Univerzita Karlova v Praze

Filozofická fakulta

Ústav Filosofie a Religionistiky

Filosofie

Mgr. Ladislav Koreň

Pravda a význam: dialektika teorie a praxe

Truth and Meaning: the Dialectics of Theory and Practice

Disertační práce

vedoucí práce - Profesor, RNDr. Jaroslav Peregrin, CSc.,

2011

Acknowledgements

I would like to thank, above all, to my supervisor, Professor Jaroslav Peregrin, who proofread several versions of the dissertation thesis and improved its text significantly**.** I am immensely grateful to him for his patience, for devoting a considerable amount of time and energy to discussing the philosophical issues I was concerned with, always being a reliable source of good advice and critical comments. I am also very grateful for the opportunity to be a member of the *Doctoral Centre for Foundations of Semantics and Representations of Knowledge*, which generously supported my research for the period of three years, especially then I would like to thank to its coordinator and *spiritus agens*, Docent Vojtěch Kolman. I want to thank to the chair of my current home department at University of Hradec Králové, Doctor Martin Paleček, for having offered me the great opportunity to teach philosophy at the department of philosophy in Hradec Králové, as well as for his support throughout. I want to thank to Professor Max Kölbel for several well-taken advices he gave to me during my stay at LOGOS Centre in Barcelona, especially, for pressing me to focus on a narrower topic than I originally planned. I am also very grateful to Professor Volker Halbach and to Professor Philippe de Rouilhan, who kindly sent me their invaluable essays, which were not available to me at the time. Last but not least, many thanks to Zuzana, for her immense support and patience. I dedicate this work to her.

Prohlašuji, že jsem dizertační práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 5.3.2011 *Podpis*

Abstract

Tarski's *semantic conception of truth* is arguably the most influential – certainly, most discussed - modern conception of truth. It has provoked many different interpretations and reactions, some thinkers celebrating it for successfully explicating the notion of truth, whereas others have argued that it is no good as a philosophical account of truth. The aim of the thesis is to offer a systematic and critical investigation of its nature and significance, based on the thorough explanation of its conceptual, technical as well as historical underpinnings. The methodological strategy adopted in the thesis reflects the author's belief that in order to evaluate the import of Tarski's conception we need to understand what logical, mathematical and philosophical aspects it has, what role they play in his project of theoretical semantics, which of them hang in together, and which should be kept separate. Chapter 2 therefore starts with a detailed exposition of the conceptual and historical background of Tarski's semantic conception of truth and his method of truth definition for formalized languages, situating it within his project of theoretical semantics, and Chapter 3 explains the formal machinery of Tarski's truth definitions for increasingly more complex languages. Chapters 4-7 form the core of the thesis, all being concerned with the problem of significance of Tarski's conception. Chapter 4 explains its logico-mathematical import, connecting it to the related works of Gödel and Carnap. Having explained the seminal ideas of the model-theoretic approach to semantics, Chapter 5 tackles the question to what extent Tarski's path-breaking article 'The Concept of Truth in Formalized Languages' (and related articles from the 1930s) anticipates this approach, and what elements might be missing from it. Chapter 6 then deals with the vexed question of its philosophical import and value as a theory of truth, reviewing a number of objections and arguments that purport to show that the method fails as an explanation (explication) of the ordinary notion of truth, and, in particular, that it is a confusion to think that Tarski's truth definitions have semantic import. Chapter 7 is devoted to the question whether Tarski's theory of truth is a robust or rather a deflationary theory of truth.

On the basis of a careful analysis, the thesis aims to substantiate the following view. [A] Tarski's theory with its associated method of truth definition was primarily designed to serve logico-mathematical purposes. [B] It can be regarded a deflationary theory of a sort, since it completely abstracts from meta-semantical issues concerning the metaphysical or epistemological basis or status of semantic properties. Indeed, [C] this can be interpreted as its laudable feature, since by separating formal (or logico-mathematical) from meta-semantical (or foundational) aspects it usefully divides the theoretical labour to be done in the area of meaning and semantic properties in general. [D] In spite of the fact that Tarski's conception of truth has this deflationary flavour, the formal structure of its method of truth-definition is quite neutral in that it can be interpreted and employed in several different ways, some of them deflationary, others more robust.

Abstrakt

Tarského sémantická koncepce pravdy je patrně nevlivnější – určitě nejdiskutovanější – moderní koncepce pravdy, která vzbudila nespočet různých interpretací a reakcí. Zatímco někteří filosofové ji oslavovali jako úspěšnou explikaci pojmu pravdy, jiní argumentovali, že nám neposkytuje adekvátní filosofický výklad tohoto pojmu. Cílem dizertace je podat systematické a kritické prozkoumání povahy a signifikance Tarského koncepce, založené na pečlivé expozici jejich konceptuálních, technických I historických předpokladů. Metodologická strategie aplikována v práci obráží autorovo přesvědčení, že nelze patřičně zhodnotit přínos Tarského koncepce bez pochopení jejich logických, matematických a filosofických aspektů, a toho jakou roli hraji v jeho širším projektu teoretické sémantiky, jak spolu souvisí (případně nesouvisí). Kapitola 2 je detailní expozicí konceptuálního i historického pozadí Tarského koncepce pravdy a metody definovaní pojmu pravdy pro formalizované jazyky, a v kapitole 3 se vysvětluje formální aparát pravdivostních definicí pro 3 typy jazyků různé komplexity. Kapitoly 4-7, které tvoří jádro celé práce, jsou věnovány ústřední otázce signifikance Tarského koncepce. V kapitole 4 se vysvětlují její logicko-matematické aspekty a přínos pro matematickou logiku, v souvislosti s výsledky Kurta Gödela a Rudolfa Carnapa. V kapitole 5 jsou pak vyloženy základní předpoklady modelové-teoretického přístupu k sémantice, a diskutuje se v ní otázka do jaké míry Tarského průkopnicky článek (1933a) „Pojem pravdy ve formalizovaných jazycích“ (a související práce z období 30 let) anticipuje tento moderní přístup. Kapitola 6 pojednává kontroverzní otázku filosofického přínosu a hodnoty Tarského koncepce, a probírá různé námitky a argumenty, který se snaží ukázat, že Tarského koncepce není filosoficky adekvátní explikací pojmu pravdy, konkrétně že sama o sobě nám neříká nic podstatného o sémantice jazyka. Kapitola 7 si klade otázku, jestli je Tarského koncepce „robustní“ teorií pravdy nebo jde spíše o „deflační“ teorii pravdy.

Na základě pečlivé analýzy se v práci pokusím motivovat a podložit následující interpretaci. [A] Tarského koncepce a na ní založena metoda definování pojmu pravdy pro daný jazyk byla primárně určena pro logicko-matematické účely. [B]] Může být považována za deflační koncepci pravdy, a sice v tom smyslu, že úplně abstrahuje od meta-sémantických otázek týkajících se metafyzické či epistemologické báze a statutu sémantických vlastností. [C] To lze ovšem vidět spíše jako její pozitivní rys, protože tím, že odděluje formální (logicko-matematické) od meta-sémantických aspektů, poukazuje na užitečnou dělbu teoretické práce v oblasti významu a sémantických vlastností obecně. Nicméně, [D] i když má Tarského koncepce pravdy tento deflační charakter, formální struktura pravdivostní definice je sama o sobě neutrální a může být interpretována a použita různými způsobami, z nichž některé jsou deflační, jiné však mohou být robustnější povahy.

[1]

Introduction

In his book-length article, *The Concept of Truth in Formalized Languages* (henceforth CTFL), the Polish logician Alfred Tarski set out to examine under what conditions and by what methods it is possible to construct a satisfactory definition of the notion of truth as predicated of sentences.[1] In the end, what he achieved was not a definition of the general notion of truth, not even of sentential truth, but, rather, a general method of constructing a definition of truth restricted to sentences of a given language belonging to a comprehensive group of formalized languages of a certain type. Tarski's method of truth definition has various logical, philosophical and mathematical aspects, owing to the fact that truth is a concept that plays a rather special role in mathematical logic, semantics, as well as in philosophy, in which areas Tarski had interest and background.[2] His work on truth has influenced all these disciplines, but its reception in them has been different.

Mathematical logicians have been concerned mainly with 'formal' aspects of Tarski's work on truth such as his analysis and solution of semantic antinomies and closely related metatheorems about definability and indefinability of truth, recursive (meta-mathematical) definitions of semantic notions and their explicit mathematical analogues within set theory, etc. It is

[1] CTFL first appeared in Polish (Tarski, 1933a), afterwards revised and augmented with the important Postscript in a German translation (Tarski, 1935), English translation of the expanded version being published in the 1st edition of *Logic, Semantics and Metamathematics* (Tarski, 1956). In this work, page references for the expanded Tarski (1935) are to the translation by J. H. Woodger published in the 2nd revised edition of *Logic, Semantics and Metamathematics* (Tarski, 1983), edited by J. Corcoran. Other relevant articles often quoted in this work are:

> — (1936a): "O pojeciu wynikania logicznego." German version (published the same year) "Über den Begriff der logischen Folgerung." English translation by J. H. Woodger published in Tarski (1983). Page references are to the translation.
>
> — (1936b): "O ungruntowaniu naukowej semantyki." German version (published the same year) "Grundlagen der Wissenschaftlichen Semantik". English translation by J. H. Woodger published in Tarski (1983). Page references are to the translation.
>
> — (1969): "Truth and Proof." Page references are to the the reprint in Hughes (1993).

[2] Tarski said of himself that he is "a mathematician (as well as a logician, perhaps a philosopher of a sort)" (Tarski 1944: 369).

well-known that his seminal results in this area of metamathematics are interestingly connected with the two incompleteness theorems of Kurt Gödel, who arrived at them around 1930-31, in which period he also obtained, quite independently of Tarski, the theorem of indefinability of arithmetical truth within arithmetic.[3] CTFL and related articles from the 1930s also contain many conceptual ingredients needed to develop a general model-theoretic take on mathematical logic that has dominated the field since the 1950s. Indeed, definitions of truth (satisfaction) of a sentence (formula) in a mathematical structure (domain) are called for to make a fully precise sense of basic metatheorems for 1st order logic, such as Löwenheim-Skolem theorem (any class of 1st order formulas that has a model, has a countable model) or Gödel's completeness theorem (all universally valid 1st-order formulas are 1st-order provable). However, such results "were proved before they were stated, so to speak", as John Burgess aptly put it.[4] Whereas Gödel, Skolem and other pioneers of metalogic seemed to be content with the informal notion of truth or satisfaction in a mathematical structure, Tarski wished to have mathematically precise definitions of such ideas, though we shall have an occasion to see that it is debatable to what extent CTFL anticipates a full-blooded model-theoretic approach to semantics. First, there was the threat of paradox; second, there were lasting philosophical worries about their metaphysical and epistemological status; last but not least, informal semantic notions were metamathematical in character, and hence beyond the realm of established mathematics.

In the 1960s, Tarski's path-breaking work on truth exerted a remarkable influence on the rapidly developing discipline of formal semantics, whose leading figure was his former disciple Richard Montague (1974), who combined a formal study of grammar with the model-theoretic approach to semantics to construct compositional semantic theories for intensional fragments of natural languages, drawing also on the seminal contributions of Carnap (1956) and the possible-worlds semantics for quantified modal logic, as worked out by Kripke (1963) and others. An alternative program in natural language semantics developed by Donald Davidson (1984) borrowed heavily from Tarski's methods, but, unlike the intensional and model-theoretic approach dominant in formal semantics, the truth-theoretic approach of Davidson was extensional, based on the method of *absolute truth*, and designed to produce ambitious philosophical implications.

Parallel to the rapid development of formal semantics there has been an intense research in semantic paradox. The natural starting point was Tarski's paradigmatic analysis of antinomies. On the one hand, he stressed the fundamental conceptual role of platitudes of the type

'Snow is white' is true iff snow is white,

which somehow capture the notion of sentential truth for a given language. On the other hand, he famously argued that elementary reasoning with the notion of truth that validates all instances of the truth-schema

'p' is true iff p

leads quickly to a contradiction, if conducted on the basis of classical bivalent

[3] Gödel (1931).
[4] Burgess (2008b: 155).

logic and with reference to a reasonably syntactically rich L that contains its own notion of truth and in which self-reference is possible. Tarski's "way out" of paradox was to say that semantic notions (their definitions or axiomatizations) for a given object-language L are to belong to a distinct metalanguage and not to L itself, where L is a regimented language of Frege-Peano type, devised for the purposes of formalizing mathematics. Classical logical reasoning utilizing the truth-schema does not lead to a contradiction, provided that we keep this principled distinction between L and meta-L. Tarski's approach to antinomies is reminiscent of the solution advanced by Bertrand Russell (1908), who suggested restricting the range of significant attributions of 'true', in accordance with his ramified theory of types. In spite of the fact that Tarski set aside natural languages on account of their expressively universal character and imprecise logico-syntactic structure, many theorists have thought it worthwhile to investigate the prospects of defining or axiomatizing truth for natural languages - or for languages approximating their expressive power - asking to what extent and by what methods it is possible to define truth for a language even within that language itself. Unfortunately, "hierarchical" solutions in the style of Tarski or Russell are not satisfactory, since it is unclear in what sense could such a language be stratified into a hierarchy of levels with different restricted truth-predicates. Kripke (1975) persuaded many that such a stratification would generate unwelcome results, it being not always possible to assign definite levels to occurrences of 'true' in sentences of natural language. In view of this, many have found it imperative or desirable to examine alternative logico-semantic frameworks that might prove to be better suited to model natural language semantics in this respect. Tarski's approach is no longer *the* dominant approach to semantic paradoxes, but it continues to be the constant source of inspiration even in the most recent debates, since it singled out two alternative paths that may be pursued here: to give up the truth-schema as a basic principle governing truth, or to weaken the underlying logic.

Unlike mathematical logicians, philosophers have focused more on the 'material' aspects of Tarski's semantic conception of truth and his method of truth definition, in particular, on his material adequacy criterion expressed in the so-called Convention T, which states, roughly speaking, that a formally unobjectionable definition of the notion of truth for L in meta-L is adequate just if it allows us to deduce from the metatheory framed in meta-L all instances of the truth-schema (or some generalized version thereof) for L. The question whether this is a philosophically satisfying theory of truth has continued to be the subject of ongoing debate, whose participants often defended positions that are hard to reconcile. Thus, it was argued – by Popper (1972b), for instance - that Tarski succeeded in rehabilitating the old good idea that *veritas* is *adequatio* or *conformitas* of language and world, thereby vindicating the realist viewpoint that does not conflate truth with epistemic ideas, which mistake is common to traditional competitors of the correspondence theory of truth such as the pragmatist, verificationist or coherence theory. After his quasi-syntacticist program reached its climax in the *Logische Syntax* (1934), Carnap came to hold the view that Tarski successfully explicated *the semantic notion of truth*, though, owing to his positivist *credo* that antithetical philosophical oppositions such as the one between *idealism* and *realism* make little sense, he was less tempted to interpret his work as a rehabilitation of the correspondence theory of truth and

vindication of realism.[5] He praised it for the intuitive plausibility of Convention T and for filtering out epistemic factors that could make only for a confusion of truth with a criterion thereof. In spite of such differences, Popper and Carnap agreed that Tarski gave a philosophically satisfying explication of the notion of truth by showing us under what conditions and by what methods we can consistently define it so as to satisfy the plausible material adequacy criterion spelled out in Convention T. Not everybody shared their highly positive evaluation of Tarski's work on truth. Otto Neurath, along with several other participants at the legendary 1935 Paris Congress for the Unity of the Science, expressed serious misgivings about it, precisely because he suspected that it attempts to resuscitate - in the modern logical guise - the idea of correspondence as a representational relation, which he held to be symptomatic of the unintelligible realistic position.[6] Still other critics have argued that Tarski's theory of truth has nothing at all to do with correspondence, since it consists of a series of platitudes compatible with virtually any metaphysical view that one may hold about the nature of truth and its relation to judgement and world.[7] Moreover, specific objections have been levelled against Tarski's method of truth definition that targeted its language-relative and "trivializing" list-like character,[8] or its allegedly counterintuitive modal or epistemic consequences;[9] moreover, a good deal of critical attention has concerned Tarski's infamous contention - made in his popular article (1936b) - to have shown that, properly relativized, truth and related semantic notions can be reduced to the notions belonging to what he called *morphology*, or, as recent theorists would say: to the notions of the object-language plus syntactical and general logical (including mathematical) notions of the metalanguage.[10] One way or another, these and related objections have questioned the widespread view, according to which Tarski's method of truth (via satisfaction) definition for L = a full-blooded semantics for L. Some have taken this to be a further confirmation of its partial or total failure as a philosophical (as opposed to logico-mathematical) theory of truth, but there have also been thinkers with *deflationary tendencies* who argued that Tarski's conception is indeed a sort of "minimal" theory of truth, but that it is to be praised for having this feature, because truth is not such a robust notion for which philosopher traditionally had it.[11]

This short overview should give the reader an initial grip on how very different interpretations are possible with respect to Tarski's work on truth. Philosophers, in particular, have showed lasting, critical obsession with it. But there is still a room for a systematic, careful and critical examination of its nature and significance, as several confusions and misunderstandings can be identified in the vast existing literature on "Tarski on truth". It is my aim to offer such an examination, based on the thorough exposition of its historical,

---

[5] See Carnap (1936), (1938) or (1942).
[6] See Carnap (1963: 61-62).
[7] Cf. Sellars (1962), Black (1948).
[8] Cf. Black (1948, Dummett (1959, Field (1972).
[9] Cf. Putnam (1985), Soames (1984) or Etchemendy (1988).
[10] Cf. Field (1972).
[11] Horwich (1982), Leeds (1978), Soames (1984). Other deflationists have complained that Tarski's framework has in a sense still "too much meat" on its bones - having in mind the compositional-style definition of satisfaction employed by Tarski to construct the definition of sentential truth for reasonably complex formalized languages. See Horwich (1990), (2005).

conceptual as well as technical underpinnings. Tarski's conception is arguably the most influential – certainly most discussed – modern theory of truth, and it is surprising that there is no book-length study covering the range of topics taken up in this work (the only monograph on Tarski's conception of truth and semantics that I know about, Fernandez Moreno's very valuable book (1992), covers many topics discussed in this work but does not discuss in depth Tarski's meta-mathematics).

The methodological strategy adopted in the thesis reflects the author's belief that in order to evaluate Tarski's conception of truth we need to understand what logical, mathematical and philosophical aspects it has, what role they play in his project of theoretical semantics, which of them hang in together, and which should be kept separate. Chapter 2 therefore starts with a detailed exposition of the conceptual background of Tarski's semantic conception of truth and his method of truth definition for formalized languages, situating it within his project of theoretical semantics, and Chapter 3 explains the formal machinery of Tarski's truth definitions for increasingly more complex languages. Chapters 4-7 form the core of the thesis, all being concerned with the problem of significance or import of Tarski's conception of truth. Chapter 4 explains its logico-mathematical import, connecting it to the related works of Gödel and Carnap. Having explained the seminal ideas of the model-theoretic approach to semantics, Chapter 5 tackles the question to what extent Tarski's path-breaking article 'The Concept of Truth in Formalized Languages' (and related articles from the 1930s) anticipates this approach, and what elements might be missing from it. Chapter 6 then deals with the vexed question of its philosophical value as a theory of truth, reviewing a number of objections and arguments that purport to show that the method fails as an explanation (explication) of the ordinary notion of truth, and, in particular, that it is a confusion to think that Tarski's truth definitions have semantic import. Chapter 7 is devoted largely to the question whether Tarski's theory of truth is to be considered a robust or rather a deflationary theory of truth.

On the basis of a careful analysis the dissertation thesis aims to substantiate the following view:

[A] Tarski's theory with its associated method of truth definition was primarily designed to serve logico-mathematical purposes.

[B] It can be regarded a deflationary theory of a sort, because it completely abstracts from the so-called meta-semantical or foundational issues concerning the metaphysical or epistemological basis or status of semantic properties. Indeed, it is its laudable feature that it separates formal (logico-mathematical) aspects from meta-semantical issues.

[C] In spite of the fact that Tarski's conception has this deflationary flavour, the formal structure of its method of truth-definition is neutral in that it can be interpreted and used in several different ways, some of them deflationary, while others being more substantive.

[2]

Semantic conception of truth

2.1 The background: between logic and philosophy

It is widely known that Tarski's aim in CTFL was to construct a satisfactory definition of sentential truth that satisfies the conditions of *formal correctness* and *material adequacy*, of which more later. He said that the problem of giving such a definition belongs "to the classical questions of philosophy",[12] or even, that the central problem of establishing the scientific foundation of the theory of truth and semantics "belongs to the theory of knowledge and forms one of the chief problem of this branch of philosophy."[13] In truth, the problem of constructing a satisfactory definition of truth was of interest to philosophy as well as mathematical logic, though not for quite the same reasons. A few remarks should be helpful to situate Tarski's work on truth within the broader context.

The *Zeitgeist* in which Tarski embarked upon his foundational investigations was one of semantic scepticism. Many of his contemporaries treated truth and related semantic ideas with a good deal of suspicion and did not believe in the possibility of a systematic and respectable theory of their properties.[14] Thus, it was long known, for instance, that the notion of truth gives rise to antinomies of the Liar-variety, when a certain sort of self-reference is present in the discourse, and the same applies to the semantic notions of denotation, definition or satisfaction, for which similar paradoxes were ingeniously constructed (such as Berry's, Richard's and Grelling-Nelson paradox respectively). Moreover, philosophical attempts at explaining truth in precise terms were not particularly successful; indeed, participants in the traditional philosophical debates tended to entangle that notion in dubious speculations about metaphysical issues concerning the nature of reality, judgment and the relation between them, and couched their conceptions in terms that they seldom tried to give precise explications of. In reaction, some influential positivist thinkers claimed that, both in its common and philosophical usage, truth is too closely associated with the mysterious idea of representing (correspondence) relations between linguistic expressions and language-independent reality that rejects explanation in scientifically respectable terms (and much the same holds for intentional notions that apply not to language but

[12] Tarski (1935: 152).
[13] Tarski (1935: 266-267).
[14] Cf. Tarski (1935: 152, 401).

to thought). But the notion of truth, so understood, has no place in science or scientific philosophy, all the more so when semantic notions are logically ill-behaved, since they give rise to paradoxes. What empirical sciences really need, according to those positivists, is some workable epistemic notion of verification or confirmation, or, in the case of deductive sciences, some purely formal-syntactic notion of provability (in a formal system). Neurath and Reichenbach held views of this sort, but while Neurath thought that in its common usage truth is absolute, hence incompatible with the holistic and dynamic conception of scientific justification and knowledge, Reichenbach concluded much the same on the different ground that truth is associated with absolute verification or certainty, which is unattainable in science. Indeed, both proposed not to use in science the common notion of truth, which was in their opinion absolute, but rather some respectable epistemic *Ersatz* notion (Neurath – *coherence*, Reichenbach – *weight or degree of confirmation*), which they tended to take as the only possibly useful notion of truth. [15]

This tendency, namely to explain the notion of truth in epistemic terms, is familiar in philosophy: to put what is a criterion or test of truth into the very definition of truth. Now, Tarski shared the view that philosophical attempts to define truth were unsuccessful, and he also had some sympathies for the positivist's critique of the marriage of truth with metaphysics. However, he deemed epistemic conceptions of truth implausible, largely on the ground that they violate the law of *excluded middle*, which intuitively holds of truth (at least by his lights). Thus

A is true or *A*'s negation is true

holds for any sentence *A*, whereas

A is confirmable (provable) or *A*'s negation is confirmable (provable)

might well fail to hold, when there is not enough evidence to decide *A* one way or another.[16]

Tarski's refusal to equate truth to some epistemic notion did not just reflect what might appear to be merely his philosophical preconception or even "realist" prejudice. It was closely connected to his specialization in metalogic (also called *the methodology of deductive science* or *metamathematics*). Metalogic studies properties of deductive disciplines axiomatized in formal-logical frameworks, as well as properties of the frameworks themselves. Consequence, validity and satisfiability are basic metalogical notions that have their intuitive semantic definitions in terms of truth (or in terms of the relative

[15] See Neurath (1983b), Reichenbach (1938). Under Neurath's influence and for very a short period of time, Carnap seemed to sympathize with such a view, but he soon became its critic and championed Tarski's theory of truth. For his critique see his (1936, 1949); a critical discussion of Reichenbach's views can be found in Soames (1999); a short exposition of Neurath and Carnap is in Candlish & Demnjanovic (2007).

[16] It should be remarked that in making such claims Tarski had in mind only sentences that are fully interpreted in that their meaning/content is sufficiently determinate to render them either true or false. In this sense, he took truth to be absolute and guided by the bivalence principle. Once we restrict attention to such sentences, various potential objections based on contextualism or relativism are beside the point. Cf. Murawski & Wolenski (2008).

notion of truth in a structure), which notion is also needed to define semantic consistency, soundness or completeness of a formalized deductive theory. Now, in the 1920s, there were strong tendencies to equate (or reduce) the notion of truth, at least for purposes of formalized deductive theories, to a syntactic notion of *provability in a formal deductive system*, where the relevant notion of provability was to be purely formal-syntactic.[17] Especially the formalists, under Hilbert's leadership, did not seem to be willing to treat *truth within a formal system* as something distinct from *provability within the system*. However, in the aftermath of Gödel's incompleteness theorems such proposals seemed doomed to failure, and Tarski was one of the first to realize that truth cannot be reduced to proof (related notions), and that semantic notions need to be conceptually distinguished from syntactic notions in metalogic (in spite of the fact that the two notions happen to coincide in extension when a formalized deductive system is complete). Gödel saw the situation in mathematical logic before his theorems as follows:

> "[...] formalists considered formal demonstrability to be an analysis of the concept of mathematical truth and, therefore were of course not in a position to distinguish the two." (Wang 1974: 9)

Indeed:

> "[...] a concept of objective mathematical truth as opposed to demonstrability was viewed with greatest suspicion and widely rejected as meaningless. (A letter to Y. Balas, in Wang 1987: 84–85)

The following passage, in which Tarski refers to Gödel's stunning results, deserves to be quoted in full:

> "Doubts continue to be expressed whether the notion of a true sentence -- as distinct from that of a provable sentence -- can have any significance for mathematical disciplines and play any part in a methodological discussion of mathematics. It seems to me, however, that just this notion of a true sentence constitutes a most valuable contribution to meta-mathematics by semantics. We already possess a sense of interesting meta-mathematical results gained with the help of the theory of truth. These results concern the mutual relations between the notion of truth and that of provability; establish new properties of the latter notion (which, as well known, is one of the basic notions of meta-mathematics); and throw some light on the fundamental problems of consistency and completeness […]" (Tarski 1944: 368).[18]

This sounds familiar to us nowadays, when we are accustomed to distinguish proof-theoretic and truth-theoretic approach to logic, so that we

[17] Carnap (1936, 1949) reports such a tendency in reaction to semantic paradoxes. Even Tarski, during the 1920s seemed to have some sympathy with the attempt to define logical consequence in terms of deducibility in a formal system (that is, in structural or syntactic terms). See for instance, his (1930).

[18] Tarski refers back to p. 354, where he mentions Gödel's results.

might forget that there was no systematic, still less mathematically precise theory of semantic notions needed in metalogic when Tarski started his work on CTFL. Even in the early 1930s, when the important metatheorems of Bernays and Post (i.e. completeness and Post-completeness of propositional logic), Löwenheim-Skolem (about the size of models of 1st-order theories) and Gödel (completeness of 1st-order logic) were already in place, there was no mathematically precise semantic theory comparable to proof theory, developed in considerable detail by Hilbert and his co-workers. The above mentioned pioneers worked commonly with the informal, metatheoretic notion of truth or satisfaction (in a mathematical structure) and related notions. That informal usage was considered to be *intuitive* and safe with respect to semantic antinomies,[19] so that conceptual analysis of truth and related semantic notions was not urgently needed to rehabilitate them in their eyes. What worried Tarski was rather the fact that truth and related ideas were used informally; no definitions or theories were available making their properties precise in terms of some respectable deductive system in which mathematics could be expressed. In his survey of model theory before 1945, Vaught remarks that since the notion of a sentence σ being true in a given structure (or domain, system)

> "[…] is highly intuitive and (perfectly clear for any definite σ), it had been possible to go even as far as the completeness theorem by treating truth (consciously or unconsciously) essentially as an undefined notion—one with many obvious [20]properties […]. But no one had made an analysis of truth […] At a time when it was quite well understood that 'all of mathematics' could be done, say, in ZF, with only the primitive notion of $\in$, this meant that the model theory (and hence much of metalogic) was indeed not part of mathematics. It seems clear that this whole state of affairs was bound to cause a lack of sure-footedness in metalogic." (Vaught 1974: 161).

The general framework that Tarski and his contemporaries commonly used in the 1920s and 1930s was either some version of (simple) type theory or some standard system of set theory (ZF = Zermelo-Fraenkel set theory). Tarski's primary aim was to show that metalogic, qua a systematic theory of truth-theoretic and proof-theoretic aspects of deductive theories, could itself be conducted in a mathematically precise way:

> "[…] meta-mathematics is itself a deductive discipline and hence, from a certain point of view, a part of mathematics […]." (Tarski 1944: 369).

Hilbert and Gödel already showed that proof-theoretic properties of formalized deductive theories can be investigated in a mathematically precise spirit. Tarski had a closely related goal of constructing mathematically precise and extensionally correct definitions of truth and related semantic notions for

---

[19] Cf. Feferman (2008a: 79). It is well-known that Russell (1908) anticipated to some extent Tarski.

[20] Cf. Frege (1893).

properly formalized languages in logically stronger frameworks. To be sure, he did not work in the vacuum and did not create the discipline of theoretical semantics *ex nihilo*. Ideas anticipating those that he elaborated in CTFL had been in the air since Plato's proto-semantic analysis of elementary predications. Brentano's neo-Aristotelian conception of intentionality and truth as well as Husserl's doctrine of semantical categories (clearly an anticipation of the modern categorial grammar) influenced strongly Lvov-Warsaw school, and reached Tarski through Twardowski, Lesniewski, Ajdukiewicz and Kotarbinski. Tarski was also thoroughly familiar with Frege's and Russell's foundational work in logic, and Frege might well be considered *the* grandfather of modern logical semantics, because he was the first to sketch a compositional-style specification of semantic values of complex expressions (of a formalized language) based on their syntax. In fact, we have seen that Tarski's freely and systematically used all fruitful ideas and methods of the flourishing discipline of mathematical logic. One can even think, as Feferman reports, that he "was only belabouring the obvious," it just took his efforts and skills to gather all essential ideas and give them a mathematically precise shape.[21] There is some truth in this thought, but it also suggests to us in what sense we can say that Tarski *co*-founded the discipline of theoretical semantics: he was the first to gather all essential ingredients and give them a mathematically satisfactory shape.[22] It is not without interest to our concerns that Tarski's logico-mathematical aims meet at this point his philosophical ambition. In the first place, he hoped to show that the semantic agenda of modern metalogic can be made formally rigorous by being expressed (interpreted) in some respectable, sufficiently powerful logico-mathematical framework, being consistent so long as the framework itself is consistent.[23] But he also hoped to show that one can put to rest various sceptical worries about the scientific respectability of semantics (at least if semantics is construed along his proposed lines). His idea here was that semantic notions are as respectable from the scientific point of view as are the notions in terms of which they are defined.[24] We shall see later that one may complain that what Tarski offered are only logico-mathematical *Ersatz* notions for full-blooded semantic notions, and that he did nothing to explain or rehabilitate the latter notions in scientifically respectable terms. There is something to this complaint and the issue will be taken up in the second part of this work.

---

[21] Feferman (2008: 90).

[22] Frege's approach to semantics was arguably systematic, but it was not formal in the sense of being *mathematically precise*. Indeed, it was not even consistent! It is a matter of controversy among commentators how seriously Frege took this informal semantic theory to be found in the *Grundgesetze* I (1893), in particular, whether he could have envisaged its use in what are genuinely metalogical investigations of his logical system, given that certain other commitments of his seem to preclude such a perspective (his logical language has a fixed universal domain and its sentences are fully interpreted so that they are either true or false – not true/false under this or that interpretation or reinterpretation of its expressions). I do not mean to enter this debate, though certain remarks of Frege show a striking connection to the metatheoretical considerations of the type that Tarski conducted in a formally precise manner, namely Frege's claim that on the basis of his semantic stipulations it can be shown that the deductive system framed on the basis of the formal language is sound and hence consistent - its axioms being true and rules of inference being truth-preserving (though there must have been something wrong with his justification as the system was inconsistent). See Heck (2010) for an interesting discussion.

[23] Cf. Vaught (1974) and Feferman (2008).

[24] Tarski (1936b: 406).

2.2. Tarski's conception of semantics

Theoretical semantics as Tarski conceived it is concerned with the totality of considerations about the notions that express relations between linguistic expressions and extra-linguistic entities.[25] Let us adopt the following terminological suggestions due to Künne to make more precise this conception of semantics:[26]

> (i) *x* is a semantic notion (predicate) in the broad sense iff *x* signifies a property that only expressions can possess, or a relation in which only expressions can stand to something, and *x* holds of an expression (of a language) in part in virtue of the expression's meaning (in the language);
>
> (ii) *x* is a semantic notion (predicate) in the narrow sense iff *x* is semantic in the broad sense such that either (a) *x* expresses a relation in which expressions stand (paradigmatically) to extra-linguistic entities, or (b) *x* is explained in terms of a broadly semantic notion that expresses such a relation.

As regards (i), Tarski tacitly assumes the important qualification to the effect that semantic notions are those that hold of expressions in part in virtue of their meanings, which is clearly needed in order to filter out (structural or syntactic) notions such as "*x* has three syllables" that hold of expressions solely in virtue of their design. (ii), on the other hand, serves to narrow down further the range of semantic notions to be considered by filtering out also notions such as "*x* means the same as *y*", which hold between two expressions (signify *word-to-word* relations) so that narrowly semantic notions will be only those that (typically) signify *world-to-world* relations, or are explainable via such notions. Not that the condition (iia), by itself, does not suffice to delimit the desired range of semantic notions. As Künne duly points out, while a predicate such as '*x* has more letters than *y* has legs' expresses a relation in which expressions stand to extra-linguistic things, hence satisfies the isolated condition (iia), it obviously does not count as a semantic predicate, since the relation it expresses obtains independently of the meaning of a word (if the word has any) that may be substituted for '*x*'.

Semantic notions that satisfy (iia) and (iib) are called 'directly' and 'indirectly' relational respectively.[27] Semantics, as Tarski conceived it, deals only with narrowly semantic notions.[28] Obviously, the notions of nominal denotation (or reference) and predicative satisfaction (or application) belong to Tarskian semantics. Still, it is fairly restricted in its scope, since broadly semantic notions such as meaning, synonymy or analyticity are neither directly relational on the face of them, nor is it clear that they can be explained via directly relational notions.[29] What about the notion of sentential truth?

---

[25] Cf. Tarski (1935: 252), (1936b: 401), (1944: 345).
[26] See Künne (2003: 179).
[27] Cf. Tarski (1969: 112).
[28] Cf. Tarski (1935: 401).
[29] See Tarski (1944: 354). He could have in mind their intensional aspects and the fact that they are not obviously 'language-to-world' relations. The matter is actually more complicated, since Tarski allowed that such notions may belong to the theoretical semantics after all, in spite of the

Grammatically, 'true' is a 1-place predicate, hence it does not directly express a relation. So, if truth is a narrowly semantic notion belonging to Tarskian semantics, it could be such only in virtue of being explicable in terms of directly relational notions. Indeed, Tarski took the notion of truth (or at least *one* intuitive notion of truth that he called *semantic*) to be indirectly relational, in part because he sensed more than a grain of truth in the intuitions to the effect that *S is true just in case things are as S says they are*,[30] and in part because he took it to have close conceptual connections to directly relational notions of denotation and satisfaction. Tarski's considered view was that the semantic character of truth consists in the fact that in order to specify the application conditions of this notion with respect to a particular declarative sentence it is necessary and sufficient to specify what the sentence says or expresses, its content or meaning, in virtue of which it "represents" the world as being a certain way. To use another platitude: a particular sentence is true iff it says that the world is a certain way and the world is indeed that way. The intuitive double-dependence of truth on content and on the way things are manifests itself in the intuitive clarity and validity of biconditionals of the type

'p' is true iff p,

or, more generally:

X is true iff p,

on the important condition that sentences whose designations replace "'p'" (or 'X') on the left-hand side have the same content as sentences that replace the dummy letter 'p' on the right-hand side.

I shall explain shortly the reasons that led Tarski to demand that an adequate definition of truth for a given language L subsume - in a sense that needs to be made precise - all such biconditionals for L as its specific cases. Suffice it to say that if the truth definition subsumes them, it captures what Tarski called *the semantic concept of truth*, with respect to L, since it captures in this way the semantic dimension of truth just sketched. Similar remarks apply to predicative satisfaction and nominal denotation, for which, Tarski points out, we have analogous paradigms of clarity:[31]

'F$x_1,...,x_n$' is satisfied by $\langle a_1,...,a_n,\rangle$ iff F$a_1,...,a_n$.

---

fact that their 'intuitive content is more involved' and their 'semantic origin is less obvious', referring the reader to his (1936a) for the definition of *logical consequence* and to Carnap's (1942) for the definition of *analyticity* (in whose terms also logical consequence can be defined. Note, however, that both these definitions rest on narrowly semantic relations and do not appeal to modal/intensional ideas. Carnap's (1942) definition of analyticity is as follows: *a sentence X is analytic in a semantical system S iff X is true in virtue of S's semantical rules alone* (synonymy being explained as equivalence under S's semantical rules). See 5.10 for my discussion of Carnap's semantic approach.

[30] For more on this see section 3 in this Chapter.

[31] These schemas are applicable to expressions of a formalized language L of 1st-order type, where L is assumed to be a part of the meta-L containing quotational names of L-expressions. When L is not a part of the meta-L but has a (non-homophonic) translation in it, or when the meta-L does not contain quotation marks but other means of forming so-called perspicuous designators (structural-descriptive designators, Gödel's numbers, and such like), things are more complicated. But I shall have to say more on these matters in due course.

> '*N*' denotes *a* iff $a = N$.[32]

Being a pioneer in the theory of definability in mathematical logic, Tarski realized that the notion of predicative satisfaction can be used to explain the notion of semantic definability of the *n*-dimensional set A (over the individual domain D of L) by an *n*-place predicate of L:[33]

> '$Fx_1,...,x_n$' defines A iff for every $a_1,...,a_n$, $\langle a_1,...,a_n,\rangle \in$ A iff $\langle a_1,...,a_n\rangle$ satisfies '$Fx_1,...,x_n$' iff $Fa_1,...,a_n$.

The same applies, *mutatis mutandis*, to the notion of denotation, since

> '*N*' denotes *a* iff *a* satisfies '$x = N$' iff $a = N$.[34]

Tarski stressed that adequate definitions (axiomatizations) of such notions for a given language L should be general formulas (or, as he also put it, logical products) subsuming all instances of such schemata (w.r.t. L). In view of this, it seems that his considered approach to semantics does not give pride of place to the idea that an adequate explanation of truth for L must render it a "correspondential" notion (*truth* =df. correspondence to facts, or designation of facts, or something of the sort). Indeed, many commentators have argued that his conception is not sufficiently robust to do duty as a full-blooded theory of truth (or semantics). But while some people take this to be its obvious shortcoming (there is more to truth than Tarski's theory reveals), others take it to be its laudable, deflationary feature (there is less to truth and semantic notions than philosophers traditionally thought). But this is to anticipate what is only to come in the second part of this work.

2.3 Sentences and language-relativity of truth definitions

Tarski's method defines the notion of truth as predicated of sentences. In CTFL Tarski did not explain this choice, but in his later, more popular papers he made some remarks to the effect that 'true' is commonly predicated of sentences, and that this is its original use in natural languages.[35] One would like to know on what evidence he based this claim, as there are reasons to disagree with it.[36] But he did not tell us, although he was aware that the claim could be disputed by those thinkers who argue that truth should be defined for beliefs or judgements or propositions, in so far as such items are taken as primarily truth-evaluable. Tarski did not want to exclude that definitions of truth can be provided for such items but declared that he, at any rate, will be concerned with the *logical* notion of truth,[37] because for purposes of logic one needs be concerned only with truth as a property of declarative sentences. At the same time, though, he pointed out

---

[32] See Tarski (1944: 345).
[33] Here, *n-dimensional set over D* is a set of ordered *n*-tuples of objects from D, for $1 \leq n$. So, a subset of D is a 1-dimensional set over D. Clearly, the definability of an object *a* (from D) by a 1-place predicate 'F*x*' of L can be explained as a special case of definability of the 1-dimensional unit set $\{a\}$. See Tarski (1948a).
[34] See Tarski (1944: 354, n. 20).
[35] Tarski (1969: 101).
[36] The *locus classicus* is Carwtright (1962). See also Soames (1999).
[37] Tarski (1969: 101).

that it is convenient to define the notion of truth for sentences, since sentences are relatively unproblematic entities (compared to propositions, say).[38] We can appreciate this strategy nowadays, as there is still no agreed upon a conception of propositions.

What is important is that Tarski thought that once the truth definition is so restricted, it has to be relativized to a particular language, since a sentence *S* that is true in language A may happen to be false in language B, depending on what it means in A and B respectively, and it may be neither true nor false in language C, in which it means nothing at all. So, the question whether *S* is true (or false) may have no definite answer, unless a particular interpretation of *S* is picked out. Such indefiniteness is to be removed by referring sentences always to a particular language: a sentence has a definite meaning, hence definite truth conditions, only in the context of a particular language.[39]

2.4 Formal correctness and material adequacy

Tarski imposed two important conditions on a satisfactory definition of truth for a language: formal correctness and material adequacy. As for *formal correctness*, he required that all the terms be listed that will be used in the definition, and that formal rules be specified in conformity to which the definition will be constructed (this, we shall see, requires the vocabulary and structure of the language to be exactly specified in which the definition will be given). The definition has to be a sentence that has the following standard form of (universally quantified) equivalence

(For every sentence *x* of L): *x* is true iff …*x*…,

where '*x*' is the only variable occurring in '…*x*…' (in the *definiens*), and what fills in the dots is an expression that contains neither 'true' nor any other term whose definition presupposes it. In a word: the definition must not be circular. Apart from that, Tarski required that the terms used in the *definiens* must not admit of any doubt or cause any methodological problems. In particular, no notion belonging to the province of semantics can occur in the *definiens*, unless it can be defined in non-semantic terms of the language (or theory) in which the definition is framed. Tarski thought this desirable, since, as we now know, many thinkers worried that semantic notions are paradoxical and/or that they cannot be explained in scientifically respectable terms (on account of the mysteriously metaphysical relation to language-independent reality). This definitional procedure ensures the consistency of the definition provided that the language (or theory) in which the definition is framed is consistent.

The motivation for the condition of *material adequacy* should not be difficult to grasp. Clearly formally correct definitions of the term 'true sentence' can be given that are intuitively inadequate as definitions of the notion of *true sentence* for a given language, e.g.:

(For every sentence *x* of English): *x* is a true sentence iff *x* has three

[38] See Tarski (1944: 342) and (1969: 101).

[39] Of course, the matter is more complicated on account of ambiguous, context-sensitive or vague sentences. See section 4.

words.

This definition is formally perfectly correct, but we see immediately that it does not get things right: it is neither a sufficient nor a necessary condition for truth of a sentence of English that it has three words. The fact that Tarski wanted a criterion of adequacy makes it clear that he thought that there is something for a truth definition to be adequate or inadequate to, that it can get things right or wrong. He did not just want to introduce a new term via a purely stipulative definition, since there is nothing for a stipulative definition to get right or wrong – it simply states how a new term is to be used. Tarski was clear on this:

> "The desired definition does not aim to specify the meaning of a familiar word used to denote a novel notion; on the contrary, it aims to catch hold of the actual meaning of an old notion […]." (Tarski 1944: 341).

If the definition aims to catch hold of the actual meaning of an old notion of truth, one would expect that Tarski took into account the way the word 'true' is actually used, since usage determines what meanings words express. And, indeed, he repeatedly claimed that he intended to give truth definitions that agree reasonably well with what he variously called the 'common', 'intuitive' or 'ordinary' usage or meaning of 'true'. At the same time, though, he admitted that the actual usage – both common and philosophical - of the word 'true' is to some extent ambiguous and imprecise and some choice has to be made if one wants to give a precise definition of the notion of truth. We have seen that epistemic definitions of truth were found wanting by Tarski on the ground that they are hard to square with the intuition that the principle of exclude middle holds of truth. Other conceptions he explicitly mentioned were pragmatic (or utilitarian), which equate truth with kind of theoretic and/or practical utility of a belief, and usually also honour epistemic virtues. However, pragmatic theories of truth do not seem to fare any better than epistemic theories, because it is counterintuitive to equate truth with utility: there seem to be true beliefs that are not useful in any reasonable sense of that word, and false beliefs that are useful in at least some reasonable sense.[40]

At any event, Tarski complained that epistemic and pragmatist theories of truth 'have little connection with the actual usage of the term 'true' and that 'none of them has been formulated so far with any degree of clarity and precision.'[41] He made it clear that his truth definitions are intended to make explicit and precise only the sound intuition about truth expressed in our familiar platitude

(P1) *S* is true iff things are as *S* says they are

or in the following statement (called 'semantical definition' of truth)

---

[40] Furthermore, both epistemic and pragmatic properties appear to be context, time or subject relative in a way that truth does not appear to be, e.g.: the sentence 'The earth is not flat' was true even in those times when all the evidence available justified rather the attitude of holding true its opposite. This was stressed by Carnap (1936) in his defense of the semantic conception of truth.

[41] Tarski (1969: 103).

> (P2) "A true sentence is one which says that the state of affairs is so and so, and the state of affairs is indeed so and so […]." (Tarski 1935: 155)

Tarski claimed that a conception or definition of truth that makes precise the intention behind them will agree 'to a very considerable extent with the common-sense usage.'[42] Such a conception of truth could rightly be called *classical*,[43] since it would make clear the same intuition that Aristotle aimed to capture in his famous definition of true (false) statement:

> (P3) "[...] to say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, or of what is not that it is not, is true [...]." (*Metaphysics*, Γ 7, 1011 b)

The intuitions expressed in such *dicta* are sometimes called *correspondence platitudes*. Tarski himself said that according to the classical conception, truth of a sentence consists in its 'corresponding with reality'[44] It has to be kept in mind, though, that one may embrace such platitudes involving the notion of truth, without endorsing a full-blooded theory that explains truth in terms of a dyadic relation (of a sort) between truth-bearers (of a sort) and truth-makers of a sort. It is, in particular, doubtful whether the platitudes (Pl), (P2) and (P3) support any full-blooded correspondential reading, and whether Tarski's conception and definition of truth that is based on such platitudes can be considered a correspondence theory.[45] I shall not address in detail this issue here. Let me just note that Tarski himself seemed to be uncertain in this respect:

> "As far as my own opinion is concerned, I do not have any doubts that our formulation does conform to the intuitive content of that of Aristotle. I am less certain regarding the later formulations of the classical conception, for they are very vague indeed […]." (Tarski 1944: 360).

He rightly worried that, qua attempts at defining the notion of truth, such slogans are not satisfactory. The definition of *truth* as *correspondence with facts* (*agreement with reality*) is only as good as our understanding of the notions used in it, in particular, of *correspondence* and *fact*. If these are not clearer than the notion of truth itself (which they do not seem to be), and are not further explained in clearer terms, we have no explanation of truth but merely the illusion of one.[46] In general, Tarski preferred Aristotle's definition, because it is not couched in the imprecise idioms of *correspondence*, *fact* or *reality*. Still, he

---

[42] Tarski (1944: 360). See also Tarski (1969: 102).
[43] Tarski (1935: 153), (1944: 343), (1969: 103).
[44] Tarski (1936b: 404).
[45] This applies both to modern theories based on the notion of correspondence to facts, as well as to more traditional theories based on the notion of correspondence to objects (derived from Aristotle). See Künne (2003) for a good discussion of these matters.
[46] Granted, some philosophers – Russell and Wittgenstein come immediately to mind here – attempted to explicate the correspondence intuition. But then their accounts faced formidable difficulties.

did not deem it sufficiently precise and general,[47] and claimed that he wanted to obtain:

> "[...] a more precise explanation of the classical conception of truth, which could supersede the Aristotelian formulation preserving its basic intentions [...]." (Tarski 1969: 103)

It was arguably not the reference to states of affairs in (P2) that Tarski took to be the sound intuition behind the platitudes, but rather this: if a sentence S says that *p*, S is true if *p*, and untrue if not *p*; in short:

S is true iff p.

If we consider the sentence 'Snow is white' and ask under what conditions it is true, what according to Tarski comes immediately to our mind is that:

(1) 'Snow is white' is true if, and only if, snow is white

(1) seems to make an obviously true claim, because the sentence used on its right side expresses the content of the sentence mentioned on its left side, the same sentence being mentioned on the left side and used on the right side. That our intuitions about the validity of (1) are not misguided can be demonstrated as follows. (1) is a material biconditional and as such it is false only if its two sides differ in truth-value, that is, if either: (a) 'Snow is white' is false and "'Snow is white' is true" is true; or (b) 'Snow is white' is true and "'Snow is white' is true" is false. But if 'Snow is white' is false, then "'Snow is white' is true" is false and not true ((a) is excluded); and if 'Snow is white' is true, then "'Snow is white' is true" is true and not false ((b) is excluded). So, the two sentences under consideration are materially equivalent and (1) holds.[48]

Nothing in principle changes when we consider the biconditional for the German sentence 'Der Schnee ist weiss':

(2) 'Der Schnee ist weiss' is true iff snow is white.

In this biconditional the sentence used on the right side expresses the content of the sentence mentioned on the left side, being its translation. For, presumably, if *S'* translates *S*, *S'* and *S* mean the same, and hence they do not differ in their truth-value.[49] So, if 'Snow is white' is false, then 'Der Schnee ist weiss' is false, and "'Der Schnee ist weiss' is true" is thus also false; and if 'Snow is white' is

---

[47] Now, generality it indeed lacks, but not for the reason Tarski mentioned. He thought that it lacks generality because it covers only non-elliptical, existential statements of the form '…is (not)' See (Tarski 1969: 102). For a persuasive exegesis of how to square Aristotle's dictum with his intentions, namely as covering categorical subject-predicate affirmations and negations, see Künne (2003). Improving on the intuition behind Aristotle's definition in light of (P1) (or something close to it) was a common practice in the Polish philosophico-logical tradition, whose members were quite agreed that the slogans in terms of correspondence or agreement with reality (facts) are problematic. For useful information about the relations of Tarski to this tradition see Murawski & Wolenski (2008).

[48] If we allow sentences that are neither true nor false, the situation changes. For if S is a sentence that is neither true nor false, it may be argued that 'S is true' is false, and the two sentences fail to be materially equivalent.

[49] Context-sensitive sentences form an important class of exceptions. See section 3.1.

true, then ‘Der Schnee ist weiss’ is true, and “‘Der Schnee ist weiss’ is true” is thus also true. Hence, the two sides of (2) do not differ in truth-value and (2) holds.

Tarski’s simple but fundamental insight was that biconditionals like (1) or (2), which refer to a particular sentence of a particular language,

> “[...] explain in a precise way, in accordance with linguistic usage, the meaning of phrases of the form ‘*x* is a true sentence’, which occur in them […].” (Tarski 1935: 187)

Thus, a particular biconditional of this kind explains wherein the truth of a particular sentence of a particular language consists, by way of specifying the condition under which the truth-predicate applies to it - its condition of truth. Such biconditionals can then be viewed as partial definitions of the notion of truth with respect to particular sentences of a given language L. This is generalized in the semantic conception of truth (henceforth SCT), apparently so-called in allusion to the semantical definition (*viz.* P2), the general intention of which it aims to make more precise and definite in the following way:

> The semantic conception of truth (SCT):
>
> (a) Every biconditional obtained from the schema *X is true iff p* by putting for ‘X’ the designator of a sentence of L and for ‘p’ either that sentence itself or its translation, holds and fixes the notion of truth for that particular sentence.
>
> (b) Taken together, such biconditionals for all sentences of L fix the notion of truth for L.

More exactly, this holds only for those sentences that do not contain ‘true’ as their significant part. The reason is that the biconditional for a sentence containing ‘true’ would contain on its right side ‘true’ and could not thus serve as a partial definition of ‘true’, because definitions are required to be non-circular. But, as he says, even though the biconditional is not a partial definition, it “is a meaningful sentence, and it is actually a true sentence from the point of view of the classical conception of truth.”[50] Furthermore, designators replacing ‘X’ should be perspicuous in that it is always possible to reconstruct from them sentences that they designate. As Tarski says:

> “Given an individual name of a sentence, we can construct an explanation of type (2)’ [namely: ‘*x* is true iff *p*’; my insertion], provided only that we are able to write down the sentence denoted by this name.’ (Tarski 1935: 156)

So, quotational names, syntactic descriptions or Gödel numbers of sentences are all perspicuous designators in this sense.

We shall follow the custom of using the label ‘T-schema’ for *X is true iff p* (or its non-English versions) and ‘T-biconditional’ for any instance of T-schema obtained in accordance with the indicated rules of substitution for its

[50] Tarski (1969: 105).

dummy letters. With SCT in place, Tarski was able to give a sharp formulation of the condition of material adequacy. If particular T-biconditionals for sentences of L fix the notion of sentential truth for particular sentences of L, then:

> "Not much more in principle is to be demanded of a general definition of phrases of the form '*x* is a true sentence' than that it should satisfy the usual conditions of methodological correctness and include all partial definitions of this type (as special cases; that it should be, so to speak, their logical product [...]." (Tarski 1935: 157)

Tarski thus proposed that a definition of the truth-predicate for L will be called or considered materially adequate (to SCT), if it entails all the T-biconditionals for L.[51] Some clarificatory comments are in order.

[A] The material adequacy criterion is supposed to capture the intention of one conception of truth only, namely SCT. Although Tarski did not mean to exclude other conceptions of truth which may suggest different criteria of material adequacy for definitions faithful to them, he nevertheless thought that SCT is a neutral ground in that it does not commit one to take any stand on issues traditionally debated by philosophers with respect to truth (e.g. idealists vs. realists, or empiricists vs. metaphysicians), which he construed as being about what, if anything, warrants or entitles assertion of a sentence of this or that kind.[52] However, he deemed it strange to uphold a conception of truth that is incompatible with SCT, since it presumably implies the denial of some T-biconditional, which denial amounts to an assertion of a sentence of the form *'p' is true iff not p*. And this is surely not a particularly attractive position.

[B] A truth definition for L that meets the material adequacy criterion is assured to be extensionally adequate in that all and only the true sentences of L fall under it. But this does not mean that it also supplies a criterion of truth for L, that is, a method or test effectively deciding whether a given sentence of L is true or not. Clearly if we are unable to decide the truth-value of a sentence, say, 'An extra-terrestrial life exists', it does not help us very much to be told that the sentence 'An extra-terrestrial life exists' is true iff an extra-terrestrial life exists. But, as Tarski pointed out, even in serious science we cannot in general expect that a definition of a notion will provide an effective method of deciding what falls under it.[53]

[C] Given Tarski's claim that his truth definition 'aims to catch hold of the actual meaning of an old notion' and repeated claims to the effect that T-biconditionals are partial definitions of 'true' for sentences of L, that together

---

[51] Tarski (1969: 106).

[52] Tarski (1944: 361).

[53] See Tarski (1944: 364), (1969: 116). The proponents of epistemic theories of truth often complained that correspondence conceptions of truth are useless as a method of discovering or recognizing or checking what is true. This seems right, but as Tarski pointed out, to object against a conception or definition of truth on the ground that it does not enable us to recognize what falls under it - to tell truth from falsehood - is a special case of asking from a definition that it supply the criterion or test enabling one to effectively decide what falls under the notion defined. Tarski rightly retorted that this is too demanding a standard of definition, which is frequently violated in scientific practice.

explain the meaning of 'true' with respect to the whole of L, it is likely that he thought that a satisfactory truth definition for L will be more than extensionally adequate. Still, he used the notion of meaning informally and loosely so that it is difficult to figure out what it was in case of 'true' (as applied to L) that he wanted to capture.[54] I shall take up the issue later, when I will discuss the problem of philosophical adequacy of Tarski's semantic definitions.

2.5 The main aims of Tarski's theory of truth

From what has been said so far the following picture emerges. A theory of truth forms the basis of semantics as conceived by Tarski, its heart being the definition that agrees reasonably well with the intuitive notion of truth used in metalogic and in science generally. No notion defined in terms of confirmation or proof will intuitively serve the needs of logic and science: whether we consider truth in logic, mathematics or any other branch of science, truth does not seem to coincide with proof, confirmation, and the like ideas. That is not to say, of course, that there is no connection between truth and epistemic notions, but the relation between them is that epistemic ideas seem to presuppose the notion of truth, because we devise our proof procedures and epistemic procedures in general to track truth, the grasp of which notion guides us in our efforts.[55] At this point we should recall the platitude:

(P1): S is true iff things are as S says they are.

(P1) may not be the most precise statement, but it is hard to deny that the sound intuition behind it is that truth of a sentence depends both on what it says (its content or meaning) and on the way things are in reality. Indeed, Tarski converted this intuition into the adequacy criterion of a definition (or theory) of truth. However, we shall see in the next chapter that he was aware that this very adequacy criterion gives rise to infamous truth-theoretic paradoxes under certain conditions, and that there were certain worries about its allegedly metaphysical character. In view of all this, Tarski set out to provide a theory of truth based on its precise definition that meets the following desiderata:

(a) it should conform to the adequacy criterion which in a way captures the above mentioned intuition (P1);

(b) it should be consistent, that is, it should not give rise to paradoxes;

(c) it should entail certain basic principles which intuitively hold of truth (e.g. the law of excluded middle and the law of non-contradiction);

(d) it should be meta-logically fruitful in that fundamental

[54] See Patterson (2008) for a detailed reconstruction of Tarski's views on meaning. Field (1972), Heck (1998) claim and Hodges (2008) argues at length that Tarski intended his material adequacy criterion to amount to no more than the extensional adequacy of truth definition, while Künne (2003) and Patterson (2008) claim that Tarski wanted, in some sense, to capture the meaning of 'true'.

[55](1969) makes this ambition of Tarski clear.

semantic notions of consequence, validity, satisfiability, etc., can be defined within it and basic meta-logical results framed in terms of such notions can be proved on its basis;

(e) it should be framed wholly in terms that are scientifically (mathematically) respectable.

Along the way, Tarski hoped to teach philosophers something important about truth-theoretic and semantic paradoxes in general: when and why they arise, what it takes to avoid them, and what consequences this has for the classical problem of truth definition – under what conditions it is possible to define truth (and semantic notions in general) both consistently and adequately, and under what conditions this is not possible.

[3]

Formal truth definitions

3.1 Exactly specified languages

Once the condition of material adequacy was sharply formulated, the problem of defining truth took the form of the question: For what languages and by what methods is it possible to construct formally correct definitions of truth that entail the T-biconditionals for all their sentences? In the sections to come we shall see that although it is easy to construct partial truth definitions for particular sentences in the form of their T-biconditionals, it is not a trivial task at all to construct a materially adequate truth definition for a reasonably rich language that contains an indefinite number of sentences. But let us first tackle the question as to for what languages it is possible to give such truth definitions.

Tarski argued that it is not possible to provide satisfactory truth definitions for natural languages. One reason he had for this negative claim was that natural languages are too loose, irregular and ill-behaved phenomena. He maintained that

> "The problem of the definition of truth obtains a precise meaning and can be solved in a rigorous way only for those languages whose structure has been exactly specified […]." (Tarski 1944: 349)

Yet, in the same breath he claimed that

> "Our everyday language is certainly not one with an exactly specified structure. We do not know precisely, which expressions are sentences, and we know even to a smaller degree which sentences are to be taken as assertible […]." (Ibid: 349)

Tarski's point is that a truth definition for a language L is materially adequate only if it entails T-biconditionals for all sentences of L. However, if it is not fixed what belongs to the set of sentences of L, it is not fixed what belongs to the set of T-biconditionals for L, and the problem of constructing a materially adequate truth definition for L becomes moot. Tarski thought that it is not settled what words belong to a natural language such as English, and that it is therefore not settled what truth-evaluable sentences belong to English, since sentences are formed from words. Moreover, he seemed to believe that the set of sentences, hence of truth-evaluable (hence assertible) sentences of a natural language,

cannot be specified in syntactical terms. All this led him to say:

> "Not every language can be described in this purely structural manner. The languages for which such a description can be given are called formalized languages. Now, since the degree of exactitude of all further investigations depends essentially on the clarity and precision of this description, it is only semantics of formalized languages which can be constructed by exact methods." (Tarski 1936b: 403)

One may well doubt if such considerations really disqualify natural languages from being given satisfactory truth definitions: for purposes of theorizing, we can conceive of a natural language as having fixed vocabulary (at a given time, say), and linguists nowadays approach natural languages by formal methods, trying to determine the category of meaningful or grammatical sentences. But Tarski had further reasons to despair of natural languages in this respect. We have seen that a sentence is to be referred to a particular language if the question as to when (or under what conditions) it is true or false is to have a definite sense. But, of course, the truth about sentential truth is more complicated. If we take sentences to be capable of being true or false, we see at once that many of them can be evaluated as true or false only if we take into account some particular context of their utterance, and that their truth-value might change across such contexts of utterance, depending on various features of contexts such as who speaks, where, when, to what addressee and with what intentions. Tarski was well aware of this, because he said that many sentences do not satisfy the condition that 'the meaning of an expression should depend exclusively on its form' and, in particular, that 'it should never happen that a sentence can be asserted in one context while a sentence of the same form can be denied in another',[56] obviously meaning that their truth-value can change across contexts of their utterance. Now, owing to the fact that the content of a context-sensitive sentence, e.g.

(3) I am hungry,

varies across contexts of its utterance (in this case, depending on who utters the sentence and when), it cannot be plausibly assigned the truth conditions in the form of the biconditional

(4) 'I am hungry' is true iff I am hungry,

because by using (4) one at best captures the truth-conditions of (3) as uttered by himself at that time, but there is no telling what the truth conditions of it are as uttered by different speakers at different times, or as uttered by himself at other times. To capture this, we need a generalization that makes explicit the dependence of the truth conditions of (3) on who utters it and when

> (5) For all *s* and *t*, 'I am hungry' is true as uttered by *s* at *t* iff *s* is hungry at *t*.

[56] Tarski (1969: 113).

This has no doubt some intuitive appeal, but we no longer have on the right side the sentence mentioned on the left side (or its translation), and hence we can no longer use the criterion of material adequacy as Tarski stated it. Tarski's strategy thus works only for eternal sentences – sentence-types all of whose tokens have the same truth conditions. In order to get a new criterion of adequacy that would apply to non-eternal sentences, we need to generalize T-schema.[57]

Other troublemakers in natural languages are ambiguous sentences that often occur within a single language, whose truth conditions might vary under their different readings or interpretations. So far as I know, Tarski did not speak of vague predicates or non-denoting terms in this respect, but it is likely that he would have considered sentences containing such expressions as equally problematic, provided that we treat them as not having a determinate truth-value and as being neither true nor false respectively. In general, he seemed to think that a precise truth definition in his style can be given only for sentences that have a sufficiently determinate sense (content) to be either true or false.[58]

3.2 Truth and paradox in natural (colloquial) language

Tarski had a second, more serious reason for claiming that formally correct and materially adequate truth definitions cannot be given for a natural language L, even if we could somehow approach L as having an exactly specified structure (and ignored context sensitive and other troubling sentences). According to him, a natural language like English is universal in that it can in principle express everything that can be expressed. In particular, then, of any English sentence we can say in English that it is true (or untrue), by appending 'is true' (or 'is untrue') to an English designator of it. Once we realize this, Tarski said, we should realize the possibility of forming a self-referential sentence of English that can (be used to) say something about itself, in particular, we can construct a sentence of English that says, of itself that it is untrue (or is equivalent to a sentence that says that). But such a sentence starts a logically plausible chain of reasoning that ends with a plain contradiction. To illustrate this, we shall use the letter '*s*' to denote the following sentence, which looks unexceptionably English:

(i) ***The only bold printed italicized sentence in this paper is not***

---

[57] The following is a tentative attempt in this direction, inspired by Garcia Carpintero (1996):

> A definition of truth for a language L is materially adequate if it entails all instances of the schema *X is true in C iff p*, where 'X' is replaced by a perspicuous designator of a sentence *x* of L, and 'p' by a sentence that says the same (expresses the same content) as *x* would have said (expressed) if uttered in C.

Context sensitivity has been intensely studied in essentially this spirit in modern truth conditional approaches to semantics heavily influenced by Tarski's own ideas - Davidson (1984), Field (1972), Montague (1974), Lewis (1983), Kaplan (1989). One may say, though, that Tarski would not have endorsed this proposal on the ground that it uses more involved ideas of 'saying the same' or 'expressing the same content'. To this it may be retorted that, in general case, he himself needed the idea of correct translation that presupposes the idea of sameness of meaning, which does not seem to be less problematic than the other two ideas.

[58] It may well be Tarski's strictures are closely connected to his absolutist view of truth. See Murawski & Wolenski (2008) for a short but instructive discussion of this idea in relation to the Polish school and Tarski.

true.

According to the material adequacy criterion stated in Convention T we are licensed to assert the following T-biconditional:

(ii) *s* is true iff the only bold printed italicized sentence in this work is not true.

But we easily confirm that the following holds:

(iii) The only bold printed italicized sentence in this paper is identical with *s*.

Given this identity, we can replace the name 'the only italicized sentence on this page' by the name '*s*' in the T-biconditional, without affecting its truth-value (according to the principle of substitutivity of identicals). What we get is this:

(iv) *s* is true iff *s* is not true

And from (iv) the contradiction

(v) '*s*' is true and *s* is not true"

follows within classic logic.

What the paradox teaches us is that the T-biconditionals entail a contradiction (within classical logic) if a certain kind of self-reference is present in the language. In particular, English has means of designating any expression or sentence that belongs to it, as well as the truth predicate 'true' ('untrue') that can be meaningfully appended to any such designator, the result being itself an English sentence. Tarski called such a language *semantically closed*. We have arrived at the contradiction assuming only

(A) the laws of classical bivalent logic hold, and

(B) the intuitive principle captured by T-schema governs the usage of 'true' to the extent that all its instances "can be asserted" (as valid, true).[59]

Unwilling to abandon (A) or (B), Tarski put blame on the factor of semantic closure and argued that no consistent classical language (i.e. one for which classical bivalent logic holds) can express its own truth-definition (truth axiomatization) that entails T-biconditionals for all its sentences. By contraposition: no classical language that can express its own truth definition (or truth theory) that entails T-biconditionals for all its sentences is consistent.

[59] Actually, an additional factor was that in English an empirically established premise 'The only bold printed italicized sentence in this paper is identical with *s*' can be formulated and accepted as true. But Tarski (1983: 168) remarked that indirect self-reference by means of empirical descriptions is not necessary, since it is possible to reconstruct a version of Grelling's paradox of *heterological predicate* in it, which does not rely on any empirical premise: let '*F*' abbreviate the predicate 'not true of itself'. Then we can apply '*F*' to itself (self-apply '*F*'), getting the sentence: "'*F*' is not true of itself". It can be shown the the sentence is false if true and true if false.

Consequently, a formally correct and materially adequate definition of truth can be constructed only for a semantically open language L, and the definition must therefore be framed in another, expressively richer language L (called 'metalanguage'), that has resources needed to define truth for L. In general, if truth for a classical L is definable (usable) in ML in a consistent and materially adequate manner, a principled distinction must exist between the object-language L and the meta-language ML: the two must not be identical or inter-translatable (though L may be a proper part of ML). If we wish to define truth for the stronger ML, we have to ascend to a yet more powerful language, since ML cannot define its own truth predicate (on pain of inconsistency).

According to Tarski's strategy,[60] we are to imagine a whole (possibly transfinite) sequence of increasingly richer languages L_0, L_1, L_2, …., where L_0 does not contain any truth-predicate, and for any n ($0 < n$), L_n contains the truth predicate 'true$_n$' that applies only to lower level sentences of L_m ($m < n$), but never to sentences of the same or higher level. The paradox is avoided, because the appropriate version of T-schema that contains an indexed truth-predicate:

(Restricted T-schema) X is true$_n$ iff p

generates meaningful (or well-formed) sentences only for sentences of a lower level than n (as substitutes for 'X'), and hence no troubling biconditionals can be asserted.

The solution of the truth-paradox along these lines is fine for formalized languages that Tarski decided to focus on, but it does not apply to natural languages, which are universal (expressively unrestricted), and because of that no principled distinction between object-language and meta-language can be made for them, since any candidate metalanguage is translatable into them. Tarski admitted that it is possible to consider fragments of a natural language L that are semantically open and provided with exactly specified structure (of a certain sort) and complete vocabularies, and then construct satisfactory truth-definitions for them on the model of formalized languages, which will be framed in richer fragments of L (possibly of another natural language). Following this suggestion we can again imagine fragments E_0 and E_1 of English, where E_0 contains only English sentences that do not contain 'true' as their significant part, and E_1 contains E_0 as well as every sentence formed by appending 'is (not) true' to the name of any sentence of E_0. If we pursue this strategy further, we might arrive at the whole hierarchy E_0, E_1, E_2,… of exactly specified and semantically open fragments of English such that for every n ($0 < n$), E_n contains the predicate true of all and only the sentences of E_{n-1}, but never of sentences of the same or higher level. Since the restricted truth-predicates in the hierarchy differ in their extensions, we had better to distinguish them as before by different indexes: E_1 contains 'true$_1$', E_2 contains 'true$_2$', and so on.

At a first glance, this looks like a perfectly rational procedure, but Tarski thought that what one is thus imagining are not so much fragments of a natural language but well-behaved products of an artificial reform of a natural

[60] Compare here Russell's type-theoretic approach: sentences, which attribute properties to sentences that attribute properties to entities of a certain type, belong to a higher level than those sentences which form their subject matter.

language:

> "It may be doubted, however, whether language of everyday life, after being 'rationalized' in this way, would still preserve its naturalness and whether it would not rather take on the characteristic features of the formalized languages […]." (Tarski 1935: 253)

Tarski's diagnosis of paradox in semantically closed natural languages such as English was that the use of 'true' in English lead to inconsistency due to it being used as an unrestricted predicate applying to all sentences of English, including sentences that contain 'true'. A decade later Tarski is more cautious, and makes the following claim:

> „But actually the case is not so simple. Our everyday language is certainly not one with an exactly specified structure. We do not know precisely, which expressions are sentences, and we know even to a smaller degree which sentences are to be taken as assertible. Thus the problem of consistency has no exact meaning with respect to this language. We may at best only risk the guess that a language whose structure has been exactly specified and which resembles our everyday language as closely as possible would be inconsistent." (Tarski 1944: 349)

This passage insinuates that the inexactness or indeterminacy characteristic of natural languages cuts both ways: if it is a serious obstacle for attempting formal truth-definitions for them, it also makes problematic unqualified claims to the effect that such languages are inconsistent. I am not sure what Tarski had in mind here. My guess is that he now admits that the assumptions (A) and (B), which are needed to draw the conclusion that truth cannot be adequately defined (used) with respect to a natural L on pain of contradiction, may be more controversial in case of natural languages than he was willing to admit in CTFL. It can even be maintained that Tarski isolated for us two alternative approaches to the problem of truth and paradox in natural languages. His preferred approach was to take semantically open fragments of such a language, formalize them, and define truth for them in logically stronger fragments. But one may seriously explore the possibility to abandon T-schema as the principle governing the notion of truth, or one may consider revisions of the classical bivalent logic, which may be taken to be inadequate for natural languages even independently of any considerations having to do with truth and paradox. It is fair to say that Tarski's views on paradox and impossibility of defining truth for a language L in L itself, dominant as they once were, are no longer universally accepted as the best possible solution, at least not so for natural languages. What he showed is not that truth for a semantically closed L cannot be consistently defined or used, *period*; rather, he showed that truth for such L cannot be defined, provided that we assume both (A) and (B). Perhaps truth can be consistently defined for L - even within L itself - once (A) or (B) or both are abandoned.[61]

[61] Kripke's (1975) *fixed point approach* is usually considered to be the decisive breakthrough. What Kripke proposed was to define truth for L within L, on the basis of some alternative logic

3.3 Formalized languages

In view of all this, Tarski proposed to focus on a comprehensive group of semantically open and well-behaved languages used by logicians to formalize deductive-mathematical theories (originally expressed in informal or semiformal fragments of natural languages). They are formalized in that in the specification of their structure only properties and relations of their sings are appealed to.[62]

For truth definitions given relative to formalized languages, Tarski was able to give the sharpest formulation of the material adequacy criterion in the form of Convention T, which reads as follows:

> Convention T:
>
> A formally correct definition of the term 'true' in the metalanguage (ML) will be called *a materially adequate definition of truth* for a given language (L) if it has among its consequences:
>
> a) all instances obtained from T-schema *X is true iff p*, by replacing 'X' by a structural-descriptive designator of any sentence of L and '*p*' by its translation into ML;
>
> b) the sentence: "For all *x*, if *x* is true, then *x* is a sentence of L."[63]

We shall have to make some clarificatory comments here. [A] Although Convention T may appear to be general, as it stands it applies only when ML is English (or better, a well-behaved fragment of it), since instances of *X is true iff p* are English sentences. For a truth definition framed in another meta-language, we have to reformulate the convention and use an appropriate version of T-schema (e.g., if ML is a fragment of German, we have to use *X ist wahr wnw p*).[64] Convention T itself is formulated in a meta-meta-language, which may or

---

(Kleene's strong three-valued schema). T-schema is not universally valid, but this is the consequence of the choice of non-bivalent background logic. The truth-predicate defined in Kripke-style is partial, in particular, paradoxical-type sentences are neither true nor false (being what he called "ungrounded"). Other important alternatives that are widely discussed are, for instance, Gupta-Belnap (1993) revision theory of truth based on the theory of circular definitions, or van McGee's theory developed in his (1991). According to many, another important breakthrough was Hintikka (1996a), who argues that for so-called IF languages (developed by Hintikka) their truth is adequately definable within them. For a good (though technical) discussion of Hintikka's conception that points out some of its limits see also de Rouilhan & Bozon (2006).

[62] But they are fully interpreted: (a) all their expression have determinate meanings so that (b) every sentence has a determinate truth-value.

[63] For the original formulation differing in certain aspects, see Tarski (1935: 187-188). Tarski notes that the condition (b) is really redundant, since is we have a definition of the set S and the set TR* that satisfies the condition (a), then we can define the set TR as the intersection of S and TR*. A structural-descriptive name is a perspicuous designator of an expression, e.g. of the following type: *the expression that consists of the word 'Snow' followed by 'is' followed by 'white'*; alternatively: *the expression that consists of three words, the 1st of which is the string of the letters Es, En, O and DoubleU, the 2nd is the string of the letters I and Es, and the third is the string of the letters DoubleU, Eitch, I, Te, and E.*

[64] As David (2008) shows, the situation is even more complicated with Tarski's original

may not contain ML. [B] If a definition meets Convention T, it is extensionally adequate – all and only the true sentences of L fall under the defined notion. For, any two definitions that satisfy Convention T are bound to be equivalent. [C] Though the 'if' in Convention T suggests that it specifies only a sufficient condition of material adequacy, it was arguably intended to state both a necessary and sufficient condition of material adequacy – *viz.* the qualifications 'it is to be demanded' and 'it should be' in the passage:

> "Not much more in principle is to be demanded of a general definition of phrases of the form '*x* is a true sentence' than that it …include all partial definitions of this type (as special cases; that it should be, so to speak, their logical product […]." (Tarski 1935: 187)

If the definition of truth for the object language L is to be formally correct, it is imperative to specify exactly the structure of L as well as of the meta-language ML in which the definition will be formulated. Now, as Tarski's analysis of semantic paradox showed that a consistent truth definition for L can be given only if L does not express its own semantics (in particular, does not contain its own truth predicate), hence cannot be given in L itself, L and ML may not be identical or inter-translatable (L may be a proper part of ML). Consequently, if truth for L is to be definable in ML, ML has to contain expressive means necessary for constructing a materially adequate definition of truth for L:

- means enabling it to talk about signs and expressions, sequences of signs and expressions, as well as of sets of expressions, operations on expressions, etc. (in short: means sufficient to express syntax of L – conceived of as an axiomatic theory).
- all non-logical and logical expressions of L or their translations (this is essential, if ML is to express T-biconditionals for sentences of L).
- variables whose order exceeds the order of any variable present in L, or quantifiers ranging over arbitrary subsets of the universe of discourse associated with L (ML has to be, in Tarski's words, essentially stronger than L).

3.4 Tarskian truth-definitions

Tarski's original strategy was to demonstrate how the method of truth definition works for a particular formalized language belonging to a certain comprehensive group (that is, fully interpreted extensional languages with quantifiers, whose syntactic structure is exactly because purely formally specified, and that are free of context-sensitive or ambiguous expressions), and then to argue that the method can be extended to other language from that group. What Tarski famously proposed was to define inductively the concept of *satisfaction* of an open sentence by an infinite sequence of objects and then to define *truth* as a

---

formulation of Convention T, which refers to the particular object-language that Tarski considers (the language of the calculus of classes) and to the particular meta-language, though it makes the appearance of generality.

limiting case of satisfaction-relation by each/some infinite sequence of objects. Tarski defined in a formally precise manner the predicate '*Tr*' (the sentential function '$x \in Tr$') that is true of all and only the true sentences of the language of the calculus of classes (henceforth LCC), which is a rather weak part of the system of simple (and finite) theory of types. This system, we have already noted, was assumed by Tarski in CTFL as a sufficiently developed and general framework of mathematical logic. A particular advantage of LCC was that it is syntactically easy to deal with - having a very poor vocabulary of simple constants and operations (syntactic constructions) by means of which complex meaningful expressions are formed.

3.4.1 Truth-definitions for code-languages

For our purposes, however, it is not necessary to explain Tarski's method of truth-definition using his preferred example (but see Appendix 1-3). On the other hand, it will be useful to demonstrate how the method works for languages L_0, L_1 and L_2 (or better, fragments of a full-blooded language) of increasing complexity, ending up with L_2 whose structure is analogous to the structure of those languages that belong to Tarski's group. Let's start with L_0, a formalized fragment of German containing only three sentences:

'Der Schnee ist weiss';

'Das Grass ist grün',

'Der Himmel ist blau',

where all of them we take to have their usual English interpretations/meanings (henceforth we assume that a well-behaved fragment of English is our meta-language). The following, list-like definition for L_0

> **(D1): *s* is a true sentence of L_1** iff *s* is a sentence of L_1 and one of the following conditions is satisfied:
>
> *s* = 'Der Schnee ist weiss' and snow is white;
>
> *s* = 'Das Grass ist grün and the grass is green;
>
> *s* = 'Der Himmel ist blau' and the sky is blue

is perfectly good by Tarski's lights, because it is explicit and uses no undefined semantic terms, and all the partial definitions of 'true' for L (T-biconditionals for sentences of L) can be deduced from it, given the following obvious syntactical facts:

'Der Schnee ist weiss' ≠ 'Das Grass ist grün';

'Der Schnee ist weiss' ≠ 'Der Himmel ist blau';

'Das Grass ist grün' ≠ 'Der Himmel ist blau'.

To check this, let us substitute "Der Schnee ist weiss" for '*x*' in the definition, thereby obtaining the following:

> 'Der Schnee ist weiss' is true iff ('Der Schnee ist weiss' = 'Der Schnee ist weiss' and snow is white) or ('Der Schnee ist weiss' = 'Das Grass ist grün' and the grass is green) or ('Der Schnee ist weiss' = 'Der Himmel ist blau' and the sky is blue)

Since we can eliminate the obviously true identity in the first disjunct:

> 'Der Schnee ist weiss' is true iff snow is white or ('Der Schnee ist weiss' = 'Das Grass ist grün' and the grass is green) or ('Der Schnee ist weiss' = 'Der Himmel ist blau' and the sky is blue)

as well as the obviously false second (since the names flanking the identity sign are obviously names of distinct objects/sentences):

> 'Der Schnee ist weiss' is true iff snow is white or ('Der Schnee ist weiss' = 'Der Himmel ist blau' and the sky is blue);

When we finally eliminate the obviously false third disjunct (since, once again, the names flanking the identity sign obviously name distinct sentences): what we are left with is the T-biconditional:

> 'Der Schnee ist weiss' is true iff snow is white

As wanted. This method, however, does not work for languages containing more than finitely many sentences, unless we are prepared to allow infinite disjunctions as acceptable definitions (as Tarski was not). But languages of theoretical interest have, as a rule, a more complex structure, which forced Tarski to look for a more general method of truth-definition.

3.4.2 Truth-definitions for propositional languages

Let us move to Tarskian truth-definitions for more complex languages. We obtain one, namely L_1, by expanding L_0, adding the constructions 'Es ist nicht der fall, dass', 'und' and 'oder' (in their usual English interpretations), by repeated application of which an infinitude of compounded sentences can be formed from the three atomic sentences of L_0. We can then define 'true' for L_1 in a well-known recursive (inductive) manner (where '*A*' and '*B*' range over arbitrary sentences of L_1 and italicized complex expressions are to be read is if they were enclosed in Quine's corner quotes):

> **(D2): *s* is a true sentence of L_1** iff *s* is a sentence of L_1 and one of the following conditions is satisfied:
>
> (a) *s* = 'Der Schnee ist weiss' and snow is white;
>
> (b) *s* = 'Das Grass ist grün' and grass is green;
>
> (c) s = 'Der Himmel ist blau' and the sky is blue;

(d) *s* = *Es ist nicht der fall dass A* and *A* is not true;

(e) *s* = *A und B* and both *A* and *B* are true;

(f) *s* = *A or B* and *A* is true or *B* is true.

The inductive definition is implicit, since the defined predicate appears also in the clauses of the *definiens* that fix its conditions of application to compound sentences. With help of an elementary set-theory (or higher order logic), however, it can be converted into an explicit definition, in which the original clauses are transformed to specify the conditions on membership in a certain set, call it 'TR':

(D3) ***s* is a true sentence of L_1** iff there is a set TR such that $s \in$ TR, and for every *x*, $x \in$ TR iff *x* is a sentence of L_1 and one of the following conditions is satisfied

(a) *x* = 'Der Schnee ist weiss' and snow is white;

(b) *x* = 'Das Grass ist grün' and grass is green;

(c) *x* = 'Der Himmel ist blau' and the sky is blue;

(d) *x* = *Es ist nicht der fall dass A* and $A \notin$ TR;

(e) *x* = *A und B* and both $A \in$ TR and $B \in$ TR;

(f) *x* = *A or B* and $A \in$ TR or $B \in$ TR

It can easily be shown that for any given sentence of L_1 its T-biconditional is derivable from this definition, the material adequacy of the definition being thereby ensured.

3.4.3 Truth-definitions for quantificational languages

We have shown how to define the truth-predicate for L_1 so that T-criterion is satisfied, by fixing the truth-conditions of any complex sentence of L in terms of the truth-conditions of its less complex component sentences, ultimately in terms of the truth-conditions of simple sentences whose truth-conditions are fixed directly in their corresponding T-biconditionals. If, now, we expand L_1 by iterative constructions of a different kind, namely quantifiers, we obtain a potentially infinite number of complex sentences that cannot be dealt with in this way, because their immediate constituents are no longer truth-evaluable sentences, but sentential functions (a generalized version of Frege-type predicates), e.g.:

x is blue,

x loves *y*,

If *x* is a man, *x* loves *y*,

For all *x*, if *x* loves *y*, *x* is a man,...

Sentential functions are obviously neither true nor false as they stand: (a) they are true or false only relative to specific assignments of appropriate values to their free variables; alternatively, (b) they are true or false only relative to specific substitutions of denoting terms for their free variables. If every object in the universe of discourse associated with a language L (e.g., a first-order language of arithmetic) is denoted by a term of L, we can well follow (b) and define the truth-conditions for quantified sentences of the form $\forall v_k A(v_k)$ in terms of the truth-conditions of sentences, in this way:

> A sentence of the form $\forall v_i\, A$ is true iff there is a term *t* of L such that $A(v_i/t)$ is true, where $A(v_i/t)$ is the result of replacing all free occurrences of v_i in *A* by *t*.

But, of course, L may not happen to contain a name of every object in its universe of discourse (L may even have no names, as was the case with the language of the calculus of classes, for which Tarski provided his truth-definition). Consequently, truth for such L cannot be defined recursively on the complexity of sentences of L. Because sentential functions can have more argument-places that are marked by different variables which can be replaced by more terms or closed by prefixing to them more quantifiers, it is better to talk, generally, about satisfaction of *n*-argument sentential functions by ordered *n*-tuples of objects. Tarski actually defined something more general: the relation *x satisfies y*, where *x* is an infinite sequence of objects (from the universe of discourse associated with L) and *y* a sentential function of L. This satisfaction relation is the converse of the relation *y is true of x*, generalized so as to cover sentential functions with arbitrary (but finite) number of argument-places.

To illustrate the method, let us consider a regimented fragment of German, L_2, which contains no simple or complex terms (in order to keep the definition simple), and its only non-logical constants are:

1-place predicates: ‘ist ein Mann’, ‘ist eine Frau’;

2-place predicates: ‘liebes’,

all of them having their usual English interpretations. Atomic sentential functions of L_2 are then formed by attaching one or two variables (possibly the same) taken from the following infinite sequence:

Individual variables: ‘x_1’, ‘x_2’,...,‘x_n’,...

to the two constants of L_2, whereas complex sentential functions and sentences of L_2 are formed from atomic sentential functions by means of (iterated applications of):

Unary operator: ‘¬’ (to be read as “it is not the case that...”)

Binary operator: ‘∧’ (to be read as “...and....”);

‘∨’ (to be read as “...or...”)

Universal quantifier: '$\forall$' (to be read as "for all/every ....").

Sentential functions of L_2 are then defined inductively as follows:

(D4) (i) *f* is a sentential function of L_2 iff one of the following conditions is satisfied:[65]

(a) $f = v_k P$, for some 1-place predicate P of L_2 and positive integer k;

(b) $f = v_k P v_l$, for some 2-place predicate P of L_2 and positive integers k, l;

(c) $f = \neg A$, for some sentential function A of L_2;

(d) $f = A \wedge B$, for some sentential functions A and B of L_2;

(e) $f = A \vee B$, for some sentential functions A and B of L_2;

(f) $f = \forall v_i A$, for some sentential function A of L_2 and positive integer

In order to define sentences (the set S of sentences of L_2,), we have to define, first, what it means for a variable to occur free in a sentential function, which can be done concisely as follows:

(ii) the variable v_i is free in the sentential function *A* iff i is a positive integer other than 0 and one of the following conditions is satisfied:

(a) A is of the form $v_i P$, for some 1-place predicate P of L_2;

(b) A is of the form $v_k P v_l$, for some 2-place predicate P of L_2 and positive integers i and k such that either $i=k$ or $i=l$;

(c) A is of the form $\neg B$, for some sentential function B of L_2, and v_i is free in B;

(d) A is of the form $B \wedge C$, for some sentential functions B and C of L_2, and v_i is free in B or v_i is free in C;

(e) A is of the form $\forall v_k B$, for some sentential function B of L_2, v_i is free in B and $k \neq i$.

Finally, sentences can be defined as follows:

(iii) *s* is a sentence of L_2 iff

s is a sentential function of L_2 that contains no free variables.

---

[65] Due the fact that L_2 contains only two 1-place predicates and one 2-place predicate, it is possible to have, in place of (a) and (b):

(a*) $f = v_k$ *ist ein man,* or $f = v_k$*ist eine Frau,* for some positive integer k; or

(b*) $f = v_k$ *liebes* v_l, for some positive integers k and l.

The definition of the satisfaction relation mimics closely the recursive definition of sentential function, proceeding by recursion on the complexity of a sentential function of L_2. That is to say, it is first specified under what conditions a given infinite sequence *p* of objects satisfies simple sentential functions, and then the conditions are specified under which *p* satisfies complex sentential functions of the forms $A \wedge B$, $A \vee B$, $\forall v_i A$, in terms of the conditions under which *p* satisfies (or does not satisfy) less complex sentential functions that are their immediate components:

(D5) the infinite sequence of objects *p* satisfies the sentential function *f* of L_2 iff one of the following conditions is satisfied:

(a) $f = v_k$ *ist ein Mann*, for some positive integer *k*, and p_k is a man;

(b) $f = v_k$ *ist eine Frau*, for some positive integer *k*, and p_k is a woman;

(c) $f = v_k$ *liebes* v_l, for some positive integers *k*, *l*, and p_k loves p_l;

(d) $f = \neg A$, for some sentential function *A* of L_2, and *p* does not satisfy *A;*

(e) $f = A \wedge B$, for some sentential functions *A* and *B* of L_2, and *p* satisfies both *A* and *B*;

(f) $f = A \vee B$, for some sentential functions *A* and *B* of L_2, and *p* satisfies *A* or *p* satisfies *B*;

(g) $f = \forall v_i A$, for some sentential function *A* of L_2 and positive integer *i*, and every infinite sequence of objects *p** satisfies *A* that differs from *p* at most at the *i*-th place.

Here, 'p_k' denotes the *k*-th member of the infinite sequence *p*. Since variables as well as objects in a sequence are ordered, every variable occurring in a sentential function gets paired with exactly one object in the sequence, via its numerical index. Such a pairing can be considered completely non-semantic so that Tarski could avoid talking about assignments of objects to variables. Now whether or not the sequence *p* satisfies the sentential function *f* depends solely on those members of *p* that are paired with the free occurrences of variables of *f*, provided it has any. But, mind you, sentences are 0-argument sentential functions with no variables free. Accordingly, whether or not a sentence is satisfied by *p* does not depend on what members of *p* are paired with its free variables. So, there are only two possibilities to consider: either a sentence is satisfied by all sequences, in which case it is true, or it is satisfied by no sequence, in which case it is not true. Truth for L_2 is thus defined directly:

Def. of truth (semantic):

***x* is a true sentence of L_2** iff *x* is a sentence of L_2 and every infinite sequence of objects satisfies *s*.

The truth definition for L_2, based on (D5), enables us to prove T-biconditionals for all sentences of L_2. But Tarski also required that no semantic

notion be used in the definition of truth unless it can be shown that it is definable in purely non-semantic terms of the language in which the truth definition is framed. Now, owing to its recursive character, (D5) fixes the application conditions of the satisfaction predicate for L_2 (fixes what sequences satisfy its sentential functions), but it does not allow us to eliminate the predicate from all contexts (we do not have a formula free of that predicate that could replace it in all contexts of the metatheory). Fortunately, (D5) can be turned to an explicit definition, provided that our metalanguage (metatheory) has a sufficiently strong set-theory, or higher order variables than L_2:

> **(D6) the infinite sequence of objects *p* satisfies the sentential function *f* of L_2** iff there is a set S such that $\langle p, f\rangle \in$ S and, for every *q* and *g*, $\langle q, g\rangle \in$ S iff *g* is a sentential function and *q* is an infinite sequence of objects and one of the following conditions is satisfied:
>
> (a) $g = v_k$ *ist ein Mann*, for some positive integer *k*, and q_k is a man;
>
> (b) $g = v_k$ *ist eine Frau*, for some positive integer *k*, and q_k is a woman;
>
> (c) $g = v_k$ *liebes* v_l, for some positive integers *k*, *l*, and q_k loves q_l;
>
> (d) $g = \neg A$, for some sentential function *A* of L_2, and $A \notin$ S;
>
> (e) $g = A \wedge B$, for some sentential functions *A* and *B* of L_2, and $\langle q, A\rangle \in$ S and $\langle q, B\rangle \in$ S;
>
> (f) $g = A \vee B$, for some sentential functions *A* and *B* of L_2, and $\langle q, A\rangle \in$ S or $\langle q, B\rangle \in$ S;
>
> (g) $g = \forall v_i A$, for some sentential function *A* of L_2 and positive integer *i*, and $\langle q^*, A\rangle \in$ S, for any infinite sequence of objects q^* that differs from *q* at most at the *i*-th place.

Once we have this explicit definition of satisfaction, truth can itself be defined in purely non-semantic terms of the metalanguage as follows:

> **Def. truth (non-semantic):**
>
> ***x* is a true sentence of L_2** iff *x* is a sentence of L_2 and, for every infinite sequence of objects *p*, $\langle p, s\rangle \in$ S, where S is as in (D6).

Since the conditions for the membership in S (in D6) have been characterized in non-semantic terms of the metalanguage, Tarski succeeded in showing how to define truth for L_2 in non-semantic terms (the definition being formally correct by his lights). With such a definition at hand, we can eliminate the semantic notion of satisfaction in favour of non-semantic terms of the meta-language - (a) the translations of expressions of L_2, (b) logical and/or set-theoretic expressions, and (c) syntactic expressions – and semantic notions are only as controversial as the apparatus in terms of which they are introduced.

In particular, a consistent metalanguage (metatheory) remains consistent if enriched by semantic notions defined explicitly. Moreover, owing to its recursive character, a Tarski-style truth definition for a quantificational L not only entails T-biconditionals for all sentences of L, but it entails also important generalizations couched in terms of truth, e.g.: that a conjunction is true just in case both its conjuncts are true; or that of a sentence and its negation exactly one is true and exactly one is false, and more of this sort. On this basis, basic metalogical theorems can be precisely stated and proved concerning consistency, soundness or completeness of deductive theories expressed in the object-languages. Given that Tarski focused on formalized languages of mathematical theories, he succeeded in showing that their metalogic, including the truth theory and semantics, can be expressed (interpreted) in their logically (set-theoretically) more powerful extensions, whose subject matter is mathematical too.

3.5 Checking material adequacy

Tarski defined truth as a limit case of satisfaction of closed sentential functions by all sequences. In fact, he pointed out that the following lemma (sometimes called *free-variable lemma*) can easily be proved by induction on the logical complexity of a sentential function (i.e. number of truth-functions and quantifiers occurring in it):

> Free variable lemma:
>
> Let f be an n-place sentential function (of L_2) and let g and g^* be arbitrary infinite sequences of objects such that $g_i = g_i^*$, for any free variable v_i ($1 \leq i \leq n$) of f. Then g satisfies f iff g^* satisfies f.

It is a consequence of this lemma (plus a few other definitions belonging to Tarski's procedure) that in case of a 0-place sentential function f - i.e. sentence - if f is satisfied by *some* sequence, it is satisfied by *any* sequence whatever. But the adequacy of the definition is tested in light of T-Convention, because T-biconditionals are glaring paradigms of clarity characterising our concept of truth, and it is by way of entailing a complete set of T-biconditionals that a formally correct truth definition is assured to be extensionally correct. The question arises at once how do we know that the proposed definition satisfies T-convention - is materially adequate and so extensionally correct? Tarski does not offer a proof of material adequacy of his truth definition for the language of calculus of classes, since it would have to be given in a meta-metatheory, being a proof about the adequacy of a definition in the metalanguage of a predicate applying to expressions of the object-language. In principle, Tarski says, it is possible to provide the proof; but it would be tedious. Mind you that for the proof to be formally rigorous, not only the meta-metalanguage, in which it should be given, but also ML would have to be formalized. Yet, up to that point Tarski conducted his metatheoretical investigations without ever bothering to formalize ML (!); though he pointed out that an eventual formalization of ML should not raise serious difficulties, he obviously thought that it would be tedious and pedagogically cumbersome to attempt it. To get a grip on the basic ideas of Tarski and get a measure of their intuitive appeal, it is better if the

definitions are framed in a (semi-formal) fragment of English. We have followed this strategy.

What Tarski offered, except of this appeal to the clarity and intuitiveness of his definitions, was not a strict (formal) proof of their material adequacy and extensional correctness, but, rather, piecemeal empirical tests to assure us that we have got things right with regard to Convention-T. Let us consider under what condition the following holds (henceforth: '$s \approx_k s^*$' shall mean that s^* is a sequence that differs from the sequence s at most at its k-th place; shortly: for every $l \neq k$: $s(l) = s^*(l)$):

S: '$\forall x_1((x_1$ ist ein Mann) $\vee$ (x_1 is eine Frau))' is true (in L_2)

By the definition of truth for sentences we have:

(1) '$\forall x_1((x_1$ ist ein Mann) $\vee$ (x_1 is eine Frau))' is true (in L_2) iff '$\forall x_1((x_1$ ist ein Mann) $\vee$ (x_1 is eine Frau))' is satisfied by every sequence.

Since, now, S is of the form $\forall x_k A$ the clause (D6f) applies:

(2) '$\forall x_1((x_1$ ist ein Mann) $\vee$ (x_1 is eine Frau))' is satisfied by every sequence iff for every sequence s, every sequence $s^* \approx_1 s$ is such that s^* satisfies '(x_1 ist ein Mann) $\vee$ (x_1 is eine Frau)'.

The sentential function '(x_1 ist ein Mann) $\vee$ (x_1 is eine Frau)' is of the form $A \vee B$, so we can apply the clause (D6f):

(3) s^* satisfies '(x_1 ist ein Mann) $\vee$ (x_1 is eine Frau)' iff s^* satisfies '(x_1 ist ein Mann)' or s^* satisfies '(x_1 is eine Frau)'

We can now apply what the clauses of (D6) tell us about satisfaction of simple sentential functions, thereby getting:

(4) s^* satisfies '(x_1 ist ein Mann)' iff s_1^* is a man

and

(5) s^* satisfies '(x_1 ist ein Mann)' iff s_1^* is a woman

Substituting right sides of (4) and (5) back into (3) (for their equivalents) we get:

(6) s^* satisfies '(x_1 ist ein Mann) $\vee$ (x_1 is eine Frau)' iff s_1^* is a man or s_1^* is a woman

By applying the rule licensing substitution of equivalents – viz. (4) and (5) – to (3), we get:

(7) s^* satisfies '(x_1 ist ein Mann) $\vee$ (x_1 is eine Frau)' iff s_1^* is a man or s_1^* is a woman,

Applying the same rule, we can now use (7) to substitute the sentence at its right side for the sentence at its left side in (2):

(8) '$\forall x_1((x_1$ ist ein Mann) $\vee$ (x_1 is eine Frau))' is satisfied by every sequence iff for every sequence s, every sequence $s^* \approx_1 s$ is such that s_1^* is a man or s_1^* is a woman.

There is a little complication with (8), since it quantifies over sequences, rather than objects. However, what (8) in effect says is that whatever infinite sequence of objects (over which we quantify in the object language) we take, and however we vary its first term, the function '(x_1 ist ein Mann) $\vee$ (x_1 is eine Frau)' will turn out satisfied. The reason should be obvious. First, *k*-alternatives to the sequence s (where $k = 1$) can be correlated in one-to-one fashion with objects (o) that are their first terms. Second, since we have to take in account all *k*-alternatives to the sequence s (including the one that agrees with s on the 1st term, and so is identical with s), this means that every object in the domain of discourse will be correlated in one-to-one fashion with exactly one *k*-alternative to s. Applying this idea to (8), what we intuitively get is this:

whatever object o happens to be s_1^* of whatever $s^* \approx_1 s$: it (o) is a man or it (o) is a woman.

In effect, quantification over all infinite sequences that are *k*-alternatives to the sequence s does the same work as would be done, more intuitively, by quantification over all objects in the domain. So that we can replace (8) by:

(9) '$\forall x_1((x_1$ ist ein Mann) $\vee$ (x_1 is eine Frau))' is satisfied by every sequence iff every object (o) is such that it (o) a man or it (o) is a woman.

Finally, substituting equivalent for equivalent in (1), we get the desired result:

(10) '$\forall x_1((x_1$ ist ein Mann) $\vee$ (x_1 is eine Frau))' is true (in L_2) iff every object (o) is such that it (o) is a man or it (o) is a woman,

which certainly looks as a *bona fide* T-biconditional for the sentence quoted on the right side.

Tarski proceeded slightly differently in "empirically" establishing that the following T-biconditional can be derived from his truth-definition for the language of calculus of classes (LCC):

$\cap_1 \cup_2 \iota_{1,2} \in Tr$ iff for every class a there is a class b such that $a \subseteq b$.

I refer the reader for all necessary details to *Appendix* (2). Suffice it to say that, at a crucial point in his informal justification of this T-biconditional, Tarski makes use of the same idea that we have seen at work in our informal proof: quantification over all *k*-variants of a sequence f does the same job as quantifying over arbitrary objects over which the quantifiers range:

> "Since $g_1 = f_1$ and the class g_2 may be quite arbitrary, only those sequences f satisfy the function $\cap_2 \iota_{1,2}$ which are such that $f_1 \subseteq b$ for any class b."

Strictly speaking, in order to deduce from the metatheory augmented with the truth-definition T-biconditionals for sentences of the form $\forall v_i A$ that do not mention infinite sequences, we need to prove in the metatheory an instance of the following schema for each given sentence in question:

$$\forall s^*[\forall k\,(k \neq i \rightarrow s(k) = s^*(i)) \rightarrow A] \text{ iff } \forall v_i A.$$

Probably because there are no special difficulties in proving such instances, Tarski did not bother to carry out the proof, being content to give informal but suggestive proofs such as the one given above.

[4]

Metamathematics of absolute truth

4.1 Basic metatheoretical results

We have seen that Tarski's main aim was to provide a set of consistent methods complementing the already well-developed proof-theoretic part of meta-mathematics, which would be based on central semantic notions of truth, satisfaction and definability.[66] This was to be achieved, preferably, by constructing *adequate definitions* of such notions for a given object-language L(T), on the basis of a metalanguage ML.[67] This demands a logically stronger ML that contains variables of higher order or stronger set-theoretical apparatus. Tarski summarized the main results as follows:

> "A. For every language of finite order a formally correct and materially adequate definition of true sentence can be constructed in the metalanguage, making use only of expressions of a general logical kind, expressions of the language itself as well as terms belonging to the morphology of language [...], i.e. names of linguistic expressions and of the structural relations existing between them.
>
> B. For formalized languages of infinite order the construction of such a definition is impossible.
>
> C. On the other hand, even with respect to language of infinite order, the consistent and correct use of the concept of truth is rendered possible by including that concept in the system of primitive concepts of the metalanguage and determining its fundamental properties by means of the axiomatic method (the question whether the theory of truth established in this way contains no contradiction remains for the present undecided)." (Tarski 1935: 265-66)

[66] That is why Tarski focused on languages of deductive theories: he wanted to cast light on the relation between formal-syntactic and semantic aspects of deductive theories

[67] Because L(T) has a built-in deductive theory T - e.g. the calculus of classes in CTFL or Peano Arithmetic, in more up-to-date accounts. To stress the fact that formalized languages in Tarski's setting are devised to formalize deductive theories in axiomatic style, we will refer to a deductive theory so formalized as 'T' and to the language itself, in which it is framed as L(T)). Thus, 'PA' refers to the deductive theory called 'Peano arithmetic', while 'L(PA)' refers to the language, in which PA is formalized.

The thesis A states the main positive result, while the thesis B states the main negative result of CTFL, if we set aside the conclusion of the argument, made at the outset of CTFL, to the effect that the notion of truth cannot be adequately defined for a semantically closed language, hence for a natural language (language of approximately its expressive power), on the ground of its expressive universality. The thesis C states a minor positive result, which compensates to some extent the major negative result. Clearly with respect to semantically closed or universal languages, it is not even possible to axiomatize consistently their semantic notions of truth on the basis of a metalanguage (embodying a metatheory), due to the fact that there is no principled distinction between object language (theory) and metalanguage (theory). Accordingly, this can be called the minor negative result. These results can be generalized so as to cover also the notions of satisfaction, definition, or denotation, telling us something about the prospects of establishing a full theoretical semantics for a given language (viz. the theses A', B' and C' immediately following in the summary the theses A, B and C).

It is to be noted that while the summary appears in the 1935 German versions of CTFL and in 1954, 1983 English versions, these, unlike the original Polish version, contain a brand new Postscript, in which the results are reformulated in a quite significant manner, the theses A, B, C being replaced by just two theses:

> "A. For every formalized language a formally correct and materially adequate definition of true sentence can be constructed in the metalanguage with the help only of general logical expressions, of expressions of the language itself, and of terms from the morphology of language – but under the condition that the metalanguage possesses a higher order than the language which is the object of investigation.
>
> B. If the order of the metalanguage is at most equal to that of the language itself, such a definition cannot be constructed." (Tarski 1935: 273)

The theses A', B' and C' are replaced in a similar manner. An obvious difference is that the thesis C (or C') drops out of the account in the Postscript. More importantly, though, the new theses extend the original results in that (i) the positive result of *adequate truth definability* is no longer restricted to languages of finite order but covers "every formalized language", including languages of infinite order, under the all important condition explicitly mentioned; (ii) the negative result undergoes a related change, spelling out the consequence of the new positive result for cases when the condition is not satisfied. There is no longer any quantification over languages of finite order.

Let me explain in more detail what was involved in this significant change of metamathematical perspective. By 1933 Tarski assumed that a system of simple theory of types, or a fragment of it that he called the general calculus of classes, was in a sense a complete and universal system of logic capable of expressing virtually any idea of logic and mathematics (this kind of completeness is of course not to confused with deductive completeness). Its language must therefore contain all the *semantical categories* of expressions

occurring in whatever deductive system. Here Tarski was strongly influenced by his teacher Lesniewski, whose conception of semantical categories was in turn indebted to Husserl, in particular, exploiting his idea that inter-changeability of expressions *salva congruitate* or *grammaticitate* (that is: preserving grammatical meaningfulness) is a criterion of their membership in the same semantical category. Tarski elaborated on this idea by giving a few illustrative examples of semantical categories: (1) sentential functions, (2) names of individuals along with individual variables, (3) names of classes of individuals and (predicates or 1-place sentence forming functors allowing only expressions of the 2nd category as arguments) along with corresponding class variables, (4) names of relations between individuals (2-place sentence forming functors allowing only expressions of the 2nd category as arguments) along with corresponding class variables, (5) names of classes of classes of individuals and (1-place sentence forming functors allowing only expressions of the 3rd category as arguments) along with corresponding class variables, etc. He then formulates two fundamental principles governing semantical categories:[68]

> i) Two expressions belong to the same semantical category iff there is a sentential function in which one of them appears as an argument, and the function remains meaningful after the replacement of it by the other.
>
> ii)Two functors belong to the same semantical category iff they have the same number of arguments in all the sentential functions in which they appear and their corresponding arguments belong to the same semantical category.

In the system of simple (finite) type-theory, viewed through the prism of Lesniewskian doctrine of semantical categories, we can recognize a hierarchy or sequence of languages, each being assigned an ordinal number n that represents its *order* in such a way that n is the supremum of the orders of all variables occurring in the language. In (1933) the guiding idea was that the order of individual constants and variables representing them is 1 (derivatively, 1 is also the order of individuals named by such constants, over which 1st order variables range). Then the order of n-argument functors (and variables representing them) that have only individual constants as arguments is 2 (derivatively, this is also the order of classes or relations that form extensions of such functors, over which 2nd-order variables range); in general, $n + 1$ (for a natural number n) is the order of n-argument functors of those sentential functions, all of whose arguments are at most of the order n and at least one argument is of the order n. It is thereby ensured that there can be no sentential function, whose functor appears in its own argument-place - just as we would expect of the type-theoretic framework.

In the end, then, every language in Tarski's hierarchy is assigned as its order either a finite ordinal, thus belonging to languages of finite order, or the first transfinite ordinal ω, so belonging to languages of infinite order. I have assumed here the terminology introduced only in the 1935-Postscript (viz. the talk about ordinals), but the account is otherwise quite faithful to Tarski's

[68] Cf. Tarski (1935: 216 - 218).

stipulations concerning the notion of *order* to be found in §4 of CTFL. Languages of finite order are there divided into three kinds:[69]

> (I) languages all of whose variables belong to one semantical category (e.g. the language LCC dealt with in §3 of CTFL – see Appendix A),
>
> (ÍI) languages whose variables belong to at least two but at most to finitely many semantical categories,
>
> (IÍI) languages to which infinitely many semantical categories of variables belong, but the orders of semantical categories are bounded above by some natural number (finite ordinal).

According to the highest order of variables occurring in them, such languages can be divided into 1st order, 2nd order, and so on, for any natural number *n*. Languages of infinite order are then characterized as those, which contain variables belonging to infinitely many semantical categories, all of whose orders are finite but not bounded above by any natural number/finite ordinal.[70] Thus, in keeping with these rules, the language of the simple type-theory or of the general calculus of classes (henceforth: LGC) is assigned an infinite order, and Tarski conceived of it as a universal logical medium adequate for the whole of logic and its ambitions, e.g. the logicist program of the *Principia Mathematica* aiming to reduce the whole of mathematics to a logical basis, approvingly mentioned by him as one of the greatest moments of modern logic. Note that the language is supposed to be complete or universal in the specific sense that every idea belonging to the realm of logic or mathematics can be expressed in it through its primitive notions or defined notions: either it is expressed by a primitive notion of it or it is shown to be reducible to such primitives by means of explicit definitions. Consequently the "universal" deductive system framed in that language contains all the semantical categories belonging to all languages of deductive sciences so that every logico-mathematical idea and proposition can be expressed in it. Indeed, every logico-mathematical theory should find an interpretation in it in such a way that all its theorems become the theorems of the universal system (again, this does not make the system "deductively universal" in the sense of proving all and only the logical or mathematical truths).

At that time, Tarski deemed it impossible for a language to contain expressions of "infinite order", because this would contradict the "finitistic character" of human languages (Tarski 1935: 253):

> "Yet neither the metalanguage which forms the basis of the present investigations, nor any other of the existing languages, contains such expressions. It is in fact not clear at all what intuitive meaning could be given to such expressions." (Tarski 1935: 244)
>
> "...the theory of semantical categories penetrates so deeply into our fundamental intuitions regarding the meaningfulness of expressions

[69] Tarski (1935: 220).
[70] Cf. Tarski (1935: 242).

> that it is scarcely possible to imagine a scientific language in which the sentences have a clear intuitive meaning but the structure of which cannot be brought into harmony with the above theory." (Tarski 1935: 244)

However, such expressions would be called for in his 1933 framework to define satisfaction relation for languages of infinite order (LGC), because such a definition needs to quantify over sequences of entities of arbitrarily high (finite) order, over sentential functions of LGC containing arguments (variables) of arbitrarily high (nut finite) order, and, finally, over relations between such sequences and sentential functions.

On this basis, Tarski argued in (1933) that truth can be defined according to his recipe for every formalized language that has a finite order, in the metalanguage, whose order is greatest at least by one. In particular, for any such language it is possible to construct such a definition on the basis of LGC or equivalent system. This is what the original thesis A summarizes. On the other hand, given the doctrine of semantical categories wedded to the simple theory of finite types, it is impossible to construct an adequate metalinguistic truth definition for languages of infinite order, because the metalanguage would have to contain expressions of infinite order and so be of higher order, which was deemed absurd by Tarski. LGC, in particular, is a language of infinite order; consequently it is not possible to construct an adequate truth definition for it on the basis of a metalanguage (metatheory), nor, of course, on its own basis, because in the later case LGC would be semantically closed, and the conditions for the semantic paradox would be satisfied. This is precisely what the original thesis B from (1933) summarizes. The remaining thesis C then states that, in spite of this limitation, the prospects of axiomatizing consistently truth for a language of infinite order (e.g. LGC) appear bright (though we shall see that Tarski made some pertinent comments on the comparatively smaller value of this procedure compared to the direct truth definition).

Using Hintikka's distinction between *logic as calculus* and *logic as universal medium* - inspired by Van Heijenoort's distinction of *logic as calculus* and *logic as language* - we can say that, at this point at least, Tarski was still an adherent of the second conception of logic, since he assumed the existence of a universal logical language that cannot - at least regarding its semantic structure - be meta-theoretically approached as it were from an "external" viewpoint of a more powerful formal-logical system.[71] On Tarski' conception of logic in 1933,

[71] Cf. Hintikka (1996b), Heijenoort (1967b). Another aspect, on the basis of which Tarski can be considered the proponent of the doctrine of logic as a universal medium is his absolutist conception of truth (in CTFL) that assumes that the notion of truth makes, strictly speaking, sense only for fully interpreted languages. It makes no sense to speak of truth relative to varying (re)-interpretations of a language, which idea is essential to the model-theoretic approach to logic. The doctrine of logic as calculus adopts a model-theoretic view, which allows various semantic re-interpretations of languages, in various structures with various domains. In the next section I discuss the question whether Tarski envisaged the model theoretic viewpoint in CTFL and related works from the period. The issue is closely connected to another question, taken up in Chapter V, of whether CTFL - in, particular, Tarski's definition of the relative notion of a correct sentence in an individual domain – anticipates the full-blooded model-theoretic definitions of semantic notions. See also an interesting article by Rodriguez-Consuegra (2005), who argues that even after CTFL Tarski continued to hold the view according to which a general set theory – 1st order or type-theoretic - is the universal framework of logic and mathematics.

there is nowhere to climb up beyond the simple type-theoretic system (or indeed, beyond its fragment, LGC, which is, in a sense, equivalent to it), since every logico-mathematical system is either a part of it or is equivalent to it. Similar sentiments were expressed by Russell, and, according to many commentators, by Frege, Wittgenstein or early Carnap (in the mid 1920s). As opposed to the conception of logic as calculus, this view is unfriendly to meta-theoretical investigations of logic, conceived of as universal and expressively or descriptively complete.[72] We can perhaps make an even bolder hypothesis to the effect that Tarski would not have deemed it possible to frame semantics for such a universal logical system even on the basis of a natural language, since human language of this type can no more contain unintelligible expressions of infinite order than formalized languages, so is not essentially stronger, in the required sense, than LGC or any other language of infinite order.[73]

This significant change in Tarski's view happened in a remarkably short passage of time. Between the publication of the 1933 and 1935 version of CTFL Lesniewski's doctrine of semantical categories was given up and Tarski explicitly allowed also for languages of transfinite orders. My own tentative explanation is that Tarski could have been influenced by Carnap's *Logische Syntax* (1934), where Carnap successfully defined analytical (logical) truth for a logico-mathematical language that was actually of ω-order (the so-called Carnap's language II); what is more, Carnap managed this in a manner remarkably close in certain important respects to Tarski's own semantic procedure.[74] In fact, Carnap independently – or almost independently, since, like Tarski, he was influenced by Gödel's results - arrived at the conclusion that analyticity for L is not definable in L but only in the meta-L that is of higher order (level – in Carnap's preferred parlance) than L. Familiarity with the following comment of Gödel – closely related to the points that Carnap made - could also play a certain role in Tarski's transition to the new conception of orders:

> "The true reason for the incompleteness inherent in all formal systems of mathematics is that the formation of ever higher types can be continued into the transfinite [...], while in any formal system at most denumerably many of them are available. For it can be shown that the undecidable propositions constructed here become decidable whenever appropriate higher types are added (for example, the type ω to the system *P*)." (Gödel 1931: 18, n. 48a).

In fact, Gödel might have been the first to get the theorem of indefinability of arithmetical truth in arithmetic in 1929/1930, as is clear from the remarks that he made in the correspondence with Bernays and Zermelo from 1930-31. Unfortunately for him, he did not mention that result in his 1931 paper, being prevented from doing that by the belief that it would not be welcome by its intended readers, since the prevalent view then was that mathematical truth must be somehow reduced to provability (albeit to relative provability "in a system") or it makes little sense. Another strong impetus could have been the fact that in the 1930s Zermelian set-theory, in its 1st order version worked out especially by

[72] Cf. Hintikka (1996b), Goldfarb (1979), Dreben & van Heijenoort (1986), Ricketts (1996).

[74] See the rewarding discussions in Coffa (1991) and Procházka (2006, 2010).

Skolem and von Neumann, was rapidly becoming the dominant setting for foundations of mathematics, removing from the pedestal the type-theoretic system of the *Principia Mathematica* (interestingly, a version of ramified theory of types was presupposed as the framework in Hilbert/Ackermann 1928 textbook, widely considered to be the first modern book on mathematical logic isolating its 1st order fragment – called "the restricted functional calculus" - and discussing rigorously its metalogical properties).[75] In the setting of set theory, Tarski notes, there are only "indefinite" variables that do not posses any definite order, but, so to speak, "run through all possible orders" (however, "orders" were to be construed ontologically, and the order of the language was to be the order of largest sets, whose existence follows from the axioms of the deductive system that is built-in the language). Tarski's improved formulation in the Postscript presupposes the existence of such indefinite variables even in languages formalized within the simple type-theory.[76]

How so? The new formulation abandons semantical categories altogether, but still retains orders, albeit defined slightly differently: individual names and variables are assigned the order 0 (and not 1, as before), names of classes of individuals and of relations between individuals (as well as corresponding class and relation variables) are assigned the order 1, but the chief difference is that *n*-argument functors are assigned their orders depending on the orders of all their arguments in all sentential functions in which they are sentence-forming functors, their order being the smallest ordinal greater than all the orders of all such arguments.[77] However, since the order of a whole language is defined as the smallest ordinal greater than the orders of all its variables, formalized languages are assigned the same orders as before. The *novum* is that the original hierarchy of languages (from 1st order up to ω-order) can be extended into the transfinite (ω+1, ..., ω^n, ..., ω^ω, ...), it being understood that as we climb up the transfinite hierarchy, the quantifiers get less and less restricted so that, in general, the range R* of quantifiers of L* whose order is greater by one than the order of L is the powerset of the range R of L's quantifiers (this holds also for languages of finite orders).

The upshot is that the principled distinction between languages of finite order and languages of infinite order loses the importance previously attached to it, because for any given L in the transfinitely extended hierarchy it is possible in

---

[75] Ten years later, in the second edition, Hilbert and Ackerman still presuppose the type-theoretic framework, this time, however, preferring a vision of simple type theory. The story of the development of axiomatized set theory in 20th century is quite interesting. Zermelo, who founded the modern axiomatized set theory, was a proponent of the 2nd order axiomatization and he was quite hostile to the idea that the proper setting for set theory is 1st order, ridiculing this position as „Skolemism", by which he meant that it is obviously absurd to think that infinite structures such as arithmetic of natural and real numbers or set theory can be adequately described by 1st-order theories admitting of non-standard models and having countable models (the results under which Löwenheim, Skolem and Gödel are signed). Whereas Gödel became soon a proponent of 1st order set theory, Skolem himself urged 1st order logic as a background for mathematics, but he seemed to take his famous paradox (which, in his view, reveals the relative character of set-theoretical notions) as speaking against set theory itself.

[76] Tarski (1983: 277).

[77] Cf. Tarski (1935: 269). The chief difference is due to the fact that one and the same functor can now be a sentential functor in different sentential functions, in which its corresponding arguments might have different orders – which was not possible before, when the doctrine of semantical categories was assumed, governed by the two principles (i) and (ii). Viz. p. 50.

principle to ascend to a higher order language L*, and construct in L* the truth definition for L that is both formally correct and materially adequate. Tarski was thus able to revise the original theses A and B as follows:

> a) Let L(T) be any formalized object-language of a classical kind and ML an appropriate metalanguage such that a) ML contains the elementary syntax of L, b) the expressions of L can be translated into ML, and c) ML is essentially stronger in its logical part than L(T) (i.e. of higher order than L(T)). Then the notion of truth can be adequately defined for L(T) on the basis of MT framed in ML.
>
> b) Let L(T) be any formalized object-language of a classical kind that contains its own syntax (or, for that matter, elementary number theory) and ML a meta-language which is not essentially stronger in its logical part than L(T) (i.e. whose order at most equals that of L(T)). Then the notion of truth cannot be adequately defined for L(T) on the basis of MT framed in ML.

So far, so good. But it is an interesting question to ask, as Hartry Field does in his recent book on truth and semantic paradox,[78] whether it makes good sense to say that for any given classical language we can define its notion of truth in a sufficiently stronger metalanguage. Maybe Tarski was closer to truth in his original 1933 diagnosis of the problem: for languages with limited expressive power we can always define their notion of truth; but we cannot define truth for expressively rich languages (say, for those containing variables of indefinite order). For many theorists, some standard 1st-order set theory (ZF) is the framework of all mathematics, so that to go beyond it means to leave the realm of mathematics.[79] Within ZF, we can construct truth definition for every language occurring in the so-called "Tarski hierarchy over arithmetic", which is the sequence starting with L(PA) not containing its own adequate truth-predicate, its 2nd term being L(PA) expanded by the adequate truth-predicate 'Tr_{PA}' restricted to sentences of L(PA), and so on, extending the hierarchy into the transfinite.[80] However, what about the language L(ZF) itself, whose

---

[78] Field (2008).

[79] Henceforth, 'ZF' will be used to refer to Zermelo-Fraenkel axioms for 1st-order set theory and 'PA' to Peano's axioms for 1st-order arithmetic (see section 4.3.1).

[80] Although we shall shortly see that Tarski showed that it is not possible to define truth for the whole L(PA) (under its standard interpretation) by a formula of L(PA), it is possible to define at least partial arithmetical truth (satisfaction) predicates within L(PA). Thus, the set of true atomic sentences of L(PA) can be adequately defined by a formula of L(PA). More generally, the set of true bounded sentences of L(PA) can be defined within L(PA) (bounded sentences, also called 'Δ_0-sentences', form the smallest set containing atomic sentences and closed under Boolean connectives and bounded quantifiers of the type '$\forall x<k$' or '$\exists x<k$'). Indeed, there is a simple decision procedure for this set via elimination of bounded quantifiers (the set of true atomic sentences being decidable). Also, the set of true Σ*-sentences* of L(PA) can be adequately defined by a formula of L(PA), where Σ–sentences are of the type $(\exists x_1),\ldots (\exists x_i)\varphi.$, where φ is a bounded formula (and the same applies to Π*-sentences* of the type $(\forall x_1),\ldots \forall x_i)\varphi$). In general, the set of true Σ_n- or Π_n*-sentences* of L(PA) that contain at most *n* logical symbols (for some *n*) can be adequately defined by a formula '$Tr_{\Sigma n}$' ('$Tr_{\Pi n}$') of L(PA), where the degree of complexity of a formula (sentence) is commonly taken to be determined by the number of alternating existential and universal quantifiers preceding its bounded core. Tarski's result about indefinability of

quantifiers range over any set whatever, so that its domain is not a set but, rather, something called "a proper class"?

As a matter of fact, already Tarski showed (in his Theorem II – of which more later) that:[81]

> for any given natural number *k*, we can define within the (type theoretic) general calculus of classes satisfaction and truth for any sub-language of LGC that contains only sentences with variables whose order is at most equal to *k*.

And something more general also holds good:

> for any definite ordinal *k*, if the quantifiers of the set-theoretical language L are restricted to range only over sets of a rank R< *k*, then satisfaction and truth for L are explicitly definable within the standard 1st order set theory (ZF).

Consequently, properly restricted set-theoretical truth-predicates can be defined within ZF.[82] However, if ZF is the universal framework of all mathematics, does not any attempt to define truth for the whole L(ZF) - its domain being a proper class - transcend mathematics? We can imagine, to be sure, a Tarski hierarchy over set-theory, starting with L(ZF) not containing its own truth-predicate, at each next level adding an appropriately restricted truth-predicate as a primitive notion ('Tr_{ZF}' to begin with). The question is whether there is a system in which each set-theoretical truth-predicate in Tarski hierarchy over set theory could be defined. Note that it won't do just to go 2nd-order, because, assuming the standard set-theoretical interpretation of 2nd order logic, everything that the second order quantifiers range over is already contained in the ranges of 1st order quantifiers of ZF.[83] If, on the other hand, we assume either the interpretation of 2nd-order quantifiers as ranging over proper classes or the non-standard interpretation of 2nd-order logic as a device of plural quantification [84], then, well, truth for 1st order set theory is explicitly definable. But the question now arises in full force with respect to 2nd order set theory so interpreted. Field observes that one may propose that there is a more powerful (1st-order) theory of „supercool entities", in which it should be possible to define truth for set theory. The problem, says Field, is that nobody seems to have a reasonably clear idea of what such entities could be like. One thing seems clear

---

arithmetical truth within arithmetic can then be put roughly like this: there is no arithmetical formula of L(PA) in the so-called arithmetical hierarchy whose extension is the set of true L(PA)-sentences. Hence, the set of L(PA)-truths is not arithmetical (is not definable by a formula of L(PA)) but it is analytical (there being a formula in the so-called analytical hierarchy whose extension is the set of L(PA)-truths). In fact, the notion of truth for L(PA) can be defined by a formula of the 2nd-order arithmetic, and the notion of truth for the language of 2nd-order arithmetic (or even of *n*th-order arithmetic, for any given *n*) can be defined within ZF. For more on this see Boolos et al (2002: 286-289), who uses a slightly different terminology. See also Smith (2007: 62-70).

[81] Tarski (1935: 255).

[82] Indeed, like partial arithmetical truth-predicates, also partial set-theoretical truth-predicates can be defined for fragments of L(ZF) that contain only sentences containing at most *n* logical symbols (of degree of complexity *n*).

[83] For more on this matter as well as on the idea of essential richness, see van McGee (1991).

[84] See Boolos (1999).

anyway: Bernays-Gödel-Von Neumann set theory is not such theory, since it is not strong enough to define set-theoretical truth. Morse-Kelley set theory could serve as a theory of such super-cool entities,[85] since we define set-theoretical truth within it. But much the same problem would still arise for this powerful theory. Are we to say, then, that there exists a yet more powerful theory of super-super-cool entities, which is related to Morse-Kelley set theory as Morse-Kelley set theory is assumed to be related to ZF, and, in general, that for any given classical theory T there is a yet more powerful classical theory T*, whose ontology is available in principle to make use of it in defining truth for L(T)?

If one is sceptical about the prospects of such a strategy, one may well suspect that a version of Tarski's 1933-thesis about indefinability of truth for infinite languages applies to a class of classical languages after all. That is to say, truth might not be definable for a language for which there is no ordinal (finite or transfinite) bound on the orders of its variables. To those thinkers who could object that, according to Tarski, it is always possible to extend the ordinals „indefinitely" so that there is simply no such thing as a language without any ordinal bound on the orders of its variables (not even transfinite), Field retorts that

> „… this, if it can be made clear at all, relies on a conception of the ordinals as 'indefinitely extensible' that Tarski does nothing to articulate. Indeed, while the Postscript is not free from ambiguity on this score, nothing is said to prohibit there being a language with variables of all possible orders." (Field 2008: 36).

Indeed, Tarski himself considered such languages, and Field reads the passages in the Postscript where he speaks of the need to introduce variables of indefinite order "running through all orders" as involving the idea of a language that has variables of all possible orders. The problem is that the incriminate passages are not very clear. It may be that Tarski wanted to say that the language needs to have variables corresponding to any expression of it of any order, and not that it needs variables corresponding to any expression whatever, of any possible order whatever. I am not sure. His position with respect to this problem seems to me to be remarkably unstable, since even after 1935 he occasionally talks about general set theory - in its 1st order or type-theoretic form - as if it was sufficient to express every idea of logic and mathematics - being universal in this sense.[86] If so, how can we ascend to a more powerful logico-mathematical language? And if there is a more powerful language, then general set theory is not expressively universal, which contradicts Tarski's claims that it is. If, on the other hand, no logico-mathematical system is expressively more powerful than the general system of set theory, then any attempt to define mathematical truth

[85] Roughly, Bernays-Gödel-Von Neumann set theory is much like ZF except that its ontology includes also proper classes (which have members but are not themselves members) and contains the class-comprehension axiom-schema '$\exists A \forall x\ (x \in A \leftrightarrow \varphi)$', where φ can contain only quantified variables that range over sets. Morse-Kelley set theory is then much like Bernays-Gödel-Von Neumann set theory, except that in it the class-comprehension axiom-schema is impredicative - φ may contain quantified variables that range over proper classes. Now, while the first theory is in fact a conservative extension of ZF, whereas the second theory is not a conservative extension of ZF - it can prove, for instance, consistency of ZF.

[86] See here Rodriguez-Consuegra (2005).

within mathematics must stop with the language of general set theory, and Tarski's Postscript thesis implicating that we can always ascend higher in the imagined transfinite Tarski hierarchy is at odds with this article of faith. While the Postscript may suggest that Tarski gave up the article of faith that allied him with the proponents of the logic-as-language view, other remarks of his from the same or later period suggests that he did not.

4.2 Gödel, Tarski and their metatheorems

Given that the reasons for the thesis (b) were persuasively explained and informally argued for, it comes as a surprise that Tarski goes on to ask whether the failure in his attempts to define truth for languages of infinite order

> "...is accidental and in some way connected with defects in the methods actually used, or whether obstacles of a fundamental kind play a part, which are connected with the nature of the concepts we wish to define, or of those with the help of which we have tried to construct the required definitions." (Tarski 1935: 246)

I take it that he meant that one could worry whether the problems do not in fact lie in his specific recipe for constructing definitions of predicative satisfaction and sentential truth (as its limiting case). In the following negative result, the question is answered in the negative: it's by no means an accidental feature of Tarski's method but a matter of principle:

> "Theorem I. *(α) In whatever way the symbol 'Tr' is defined in the metatheory, it will be possible to derive from it the negation of one of the sentences which were described in the condition (α) of the convention T;*
>
> *(β) assuming that the class of all provable sentences of the metatheory is consistent, it is impossible to construct an adequate definition of truth in the sense of convention T on the basis of the metatheory*." (Tarski 1935: 247).

Theorem I assumes that the metalanguage is of the same order as the object-language. In 1933, this was independently motivated by the fact that Tarski was then unwilling to allow for languages of transfinite order. However, the theorem retained its force even after he had abandoned the theory of semantical categories and introduced infinite (transfinite) types: it is impossible to define truth for the object-language in the metalanguage, when the later is not essentially stronger, that is, is at most of the same order as the former.

Before I go into the details of Tarski's metamathematics, it will be useful to have in place another seminal contribution, Gödel's (1931) results on incompletability of the system of *Principia Mathematica* and any related consistent and axiomatizable system of elementary arithmetic, which shattered two of the most prominent programmes in foundations of mathematics pursued in the first three decades of the 20th century. I mean of course the logicist programme of Russell and Whitehead, whose ambition was to lay the foundations of a universal logical system completely capturing all of

mathematics, and the formalist program of Hilbert, who aimed to arrive at finitary (syntactic or elementary number-theoretic) consistency proofs for the established systems of mathematics (arguably, also at the proofs of their deductive completeness, as a highly important desideratum of the axiomatic approach). Gödel's first incompleteness theorem states that:

> **(Gödel's incompleteness theorem I)**
>
> *No consistent, recursively axiomatizable theory T embedding elementary arithmetic is complete in the sense of proving, for every sentence A in its language L(T), either A or non-A.*[87]

The obvious upshot of Gödel's first incompleteness theorem seems to be that there is a true yet unprovable sentence in T, if only we are ready to assume that for any *A*: either *A* is true or *non-A* is true (*the law of excluded middle*). Gödel showed that this situation is a matter of principle, not a contingent feature of deductive systems of the sort he investigated. Suppose we have showed T to be an incomplete theory of the sort spoken about in Gödel's 1st theorem and that we add to T's axioms its true yet unprovable sentence(s), thereby obtaining a more comprehensive deductive theory T*. Still, what Gödel showed is that the nature of the case is such that we can reapply his method (to be explained shortly) to T* in order to show that there is in T* a true yet unprovable sentence. This reasoning can be repeated as many times as we want: extending T* to T** by adding to T* its true yet unprovable sentence(s) as axiom(s) does not make S** a complete system, and so on. Gödel's second incompleteness theorem then states that:

> **(Gödel's incompleteness theorem II)**
>
> *No consistent, recursively axiomatizable theory T embedding elementary arithmetic can prove its own consistency.*

The idea animating the proof of this theorem was that consistency of T can be defined as a purely syntactic (proof-theoretic) property of T (i.e.: *there is a sentence of L(T) not provable in T*) and he showed that all such properties can be indirectly expressed in T itself via his procedure of arithmetization of syntax (metamathematics), including the intuitively true claim to the effect that T is consistent (intuitively true, given that axioms capturing the structure of the

[87] T is recursively axiomatizable iff there is an algorithm (Turing machine) such that its set of axioms is effectively decidable by that algorithm (or T is equivalent to a theory T* whose set of axiom is algorithmically decidable) in the following sense: given any sentence of L(T)) the algorithm decides in a finite number of steps whether the sentence is a T-axiom or not. The very idea of formal theory or system T involves the demand that the following syntactic properties of L(T) and T are algorithmically (effectively) decidable: *term of L(T), sentence of L(T), axiom of T,* an *L(T)-sentence being a direct deductive consequence of other L(T)-sentences according to a rule of inference of T*. If so, also the property of *T-proof* is effectively decidable: given any sequence of L(T)-sentences, it is decidable in a finite number of steps whether or not the last term of the sequence is correctly derived in T from the remaining terms. If the hope of the formalist led by Hilbert was that the category of T-theorems is decidable for a reasonably rich formal T containing elementary arithmetic, then Gödel's results showed that this hope is to be dashed: any such theory is bound to be undecidable, its set of theorems not being effectively decidable.

domain of natural numbers and rules of inference are properly chosen). Then, by a simple (informal) reasoning drawing on his already proved 1st theorem, Gödel showed that neither the consistency-claim nor its negation can be proved T. The consistency-claim is thus an example of a sentence undecidable in T.

The consequences of Gödel's theorems were far-reaching, indeed. *Pace* the logicist program of Russell, mathematics cannot be completely axiomatized in one comprehensive logical system. Admittedly, Russell did not dream of a completeness or consistency proof for such a universal logical system – since he deemed it impossible to adopt as it were an external metatheoretical perspective on the system from which to produce such proofs - but he surely grasped the significance of the question whether the system is deductively complete with respect to mathematical truths, which had been in the air since at least Euclid:

> "[...] the system must embrace among its deductions all those propositions which we believe to be true and capable of deduction from logical premises alone." (Whitehead & Russell 1908-13: 12).

If the first theorem compromised logicism, the second theorem seemed to compromise even more Hilbert's project of establishing consistency of mathematics using only finitary, essentially, syntactic or elementary number-theoretic methods.[88] Partly in reaction to the intuitionist „putsch" initiated by Brouwer - joined by Hilbert's *protégé* Hermann Weyl - who levelled the constructivist challenge to „abstract mathematics" in the style of Cantorian set theory, Hilbert urged what came to be known as his *conservation programme* of reducing all abstract inferences and *ideal propositions* to finitary proof-theoretic methods (reducible to purely formal rules of manipulations of symbols) and *real propositions*.[89] The idea was to prove by concrete and finitary methods that the whole of abstract (infinitary) mathematics is conservative over real (finitary) mathematics in the sense that every proposition in the language of real mathematics proved via recourse to abstract mathematics can already be proved on the basis of real mathematics alone. Kleene summarizes this concisely, saying that finitary methods:

> „… can be characterized as methods not using any completed infinity; i.e., no objects themselves infinite are to be used, and only potentially infinite collections of them, like the natural number sequence 0, 1, 2,... considered as unbounded above but not as a completed collection." (Kleene 1986: 127).

As for real propositions, they were presumably taken to be those of the type '$\forall x(g(x) = f(x))$', where f and g are primitive recursive functions.[90] Broadly

[88] The inspiration goes back at least to Dedekind and functions of that sort were studied in closer detail by Grassmann in the 19th century. Primitive recursive functions of arithmetic and their definability therein was the problem on which Skolem systematically worked in the 1920s; he even had a sort of programme of founding mathematics on the primitive recursive part of it, a vision of which idea was propounded also by Weyl.

[89] Similar constructivist ideas were expressed earlier by Kronecker and Poincare, who could also appeal to the philosophical authority of Kant.

[90] Or, perhaps, f and g are allowed to be (*general) recursive functions*, which, given Church's thesis, coincide with computable functions (hence with Turing computable functions or any equivalent); primitive recursive functions then form a subset of such computable functions. For a

speaking, primitive recursive functions are those that can be built up from the zero function, the successor function and the projection function by applying any number of times the operation of function-composition and recursion. More precisely still, they form the set *P* such that:

(i) *P* is the intersection of all sets of functions that include
  a) $f(x) = 0$
  b) $f(x) = (x + 1)$
  c) ${}^{n}f_i\,(x_1, \ldots, x_n) = x_i$

(ii) *P* is closed under composition (every composition of the functions in *P* is also included in *P*) and recursion (every function formed by recursion from the functions in *P* is included in *P*).[91]

We may think of Hilbert's conservation programme as attempting to reduce – more in epistemological than ontological sense – all of mathematics to the (primitive) recursive arithmetic, which ambitions took the form of the requirement of proving, solely on this basis, consistency of ever more powerful fragments of mathematics, starting with arithmetic, as axiomatized by Dedekind and Peano, and finishing with Cantorian set theory, in some axiomatization of it. Attractive as this programme once was, Gödel's second incompleteness result showed it hopeless - at least in the version just sketched. As Gödel showed, it is possible to provide the proof of consistency for a deductive theory, but only on the basis of a higher order deductive theory (so by non-finitary means), which, however, is just the sort of relative consistency proof (relative, that is, to a more powerful theory) that Hilbert et al deemed unsatisfactory. For Hilbertians, the proof of relative consistency of a mathematical theory is welcome, but only if it is given on the basis of primitive recursive arithmetic, or another theory that has been reduced in this way to primitive recursive arithmetic.

Gödel himself was initially cautious in his claims concerning the prospects of Hilbert's formalist project in the aftermath of his stunning discoveries, allowing for the possibility of formal systems of a different kind than he investigated, in which, perhaps, consistency of various mathematical theories could be proved by essentially finitary means. He was initially willing to allow for the possibility that not all finitary methods must be expressible within the elementary arithmetic. Indeed, some thinkers have seen in Gentzen's consistency proofs or similar methods a way of fulfilling the intentions of Hilbert's programme. Thus, to quote Gentzen himself:

> „From Gödel's incompleteness theorems it follows that the consistency of elementary number theory, for example, cannot be established by means of part of the methods of proof used in

closer discussion of the question (still intensely debated) of what Hilbertians took to belong within the scope of finitary mathematics see Sieg (2009) or Zach (2003), (2006). Recent investigations seem to show that there was no agreement on that issue.

[91] Gödel proved that primitive recursive functions and relations are closed under (i) composition, and (ii) the logical operations of negation, disjunction, conjunction, bounded minimization, and bounded quantification. For a detailed discussion of primitive recursive and recursive functions see Boolos et al (2002), or Smith (2007).

> elementary number theory, nor indeed by all of these methods. To what extent, then, is a genuine reinterpretation still possible?
>
> It remains quite conceivable that the consistency of elementary number theory can in fact be verified by means of techniques which, in part, no longer belong to elementary number theory, but which can nevertheless be considered to be more reliable than the doubtful components of elementary number theory itself.“ (Gentzen 1936: 139).

The trouble is that Gentzen’s method is based on the method of transfinite induction up to the ordinal ε_0 and one might suspect that this amounts to abandonment of Hilbert’s original programme. The precarious situation is nicely summarized by another giant of the period, whom we could scarcely charge for not being intimately familiar with Hilbert’s programme:

> „…the hopes for a finitistic proof of consistency have become dim indeed. G. Gentzen's ingenious proof of consistency for arithmetic (1936) is not finitistic in Hilbert's sense; the price of a substantially lower standard of evidence is exacted from him, and he is forced to accept as evident a type of inductive reasoning that penetrates into Cantor's “second class of ordinal numbers.” Thus the boundary line of what is intuitively trustworthy has once more become vague. After this Pyrrhic victory nobody had the courage to carry arms into the field of analysis; yet it is here that the ultimate test for Hilbert's conception would lie.“ (Weyl 1949: 220).

Indeed, Gödel used to be quite cautious, but he grew little bolder later in his life, arguing that Gentzen’s and similar proposals cannot be considered merely cosmetic modifications of the original Hilbert’s programme. I think he would agree with Weyl that the alleged victory is Pyrrhic.

4.2.1 Gödel’s first theorem

For simplicity’s sake and certain dialectical aims of mine, I shall not adhere slavishly to Gödel’s (who focused on the system of simple type-theory based on the domain of natural numbers) and Tarski’s original proofs (for the general calculus of classes (LGC) also formalized within a simple type-theoretic framework). My object will be a standard formalized language of arithmetic, in which a deductive system is framed sufficiently powerful to embed elementary arithmetic. Both Gödel’s and Tarski’s system embed elementary arithmetic in their own way (the later contains variables of all finite orders, including variables of 3rd order ranging over classes of classes of individuals so that natural numbers can be defined following Frege-Russell’s proposal as classes of classes of individuals with the same cardinality). But nowadays type-theoretic frameworks are no longer in fashion, and it is more common to focus directly on languages of (1st-order) arithmetic with denumerably many (possibly indexed) individual variables {*x*, *y*, *z*, ...}, standard first-order logical constants {$\forall$, $\exists$, $\wedge$, $\vee$, $\rightarrow$, =}, and a finite stock of non-logical constants standing for certain designated elements of the domain of natural numbers (typically ‘0’ for zero), and certain 1-place and 2-place functions (typically ‘S’ for successor function,

'+' for addition, and '×' for multiplication) defined over the domain (N) of natural numbers.[92] Such a language (L(PA)) is called the language of Peano arithmetic (PA), and we shall have more to say about its syntactic and semantic structure in the next section devoted to the notion of relative truth. It is of interest to us because of the deductive theory framed it, which is axiomatized as follows, it being understood that quantifiers range over N:

A1 $\forall x\ (0 \neq Sx)$

A2 $\forall x\ \forall y\ (Sx = Sy \rightarrow x = y)$

A3 $\forall x\ (x \neq 0 \rightarrow \exists y(x = Sy))$

A4 $\forall x\ (x + 0 = x)$

A5 $\forall x\ \forall y\ (x + Sy = S(x + y))$

A6 $\forall x\ (x \times 0 = 0)$

A7 $\forall x\ \forall y\ (x \times Sy = (x \times y) + x)$

Induction Schema $([\phi(0) \wedge \forall x\ (\phi(x) \rightarrow \phi(Sx))] \rightarrow \forall x \phi(x))$.

A considerably weaker, yet for reasons spelled out bellow interesting axiomatization of arithmetic is Robinson Arithmetic (called **Q**) axiomatized by (A1),…(A7) only, lacking any version of the induction principle.[93] This system is elementary and deductively weak, because it proves only a few of the generalizations about natural numbers that any good theory of arithmetic should prove. However, its importance for metamathematics lies in the fact that it can be considered a minimal axiomatic system that expresses all recursive (hence all primitive recursive) functions and relations (indeed, it represents them – in the sense of 'represent' that we are yet to specify more precisely), with respect to which the fundamental metatheorems can be stated (since it contains means necessary for Gödel's method of arithmetization of its own syntax including proof-theory).[94] The point is, to anticipate, that any consistent, recursively

[92] Plus auxiliary symbols such as parentheses of various convenient types.

[93] In Peano's original formulation, 1 was used in place of 0. Moreover, Peano's axioms use only the primitive sign for successor function. The really important difference was that Peano's axiomatization (acknowledging the dept to Dedekind, who gave an equivalent axiomatization of arithmetic) had the following axiom (in place of the induction schema that generates infinitely many axioms) that we can formulate thus:

For every set X: $([0 \in X)$ and $\forall x(x \in X \rightarrow S(x) \in X)] \rightarrow \forall x(x \in X))$.

Since this axiom quantifies over the subsets of the domain of natural numbers, what Peano gave in effect was a 2nd order axiomatization of arithmetic (by today's standards), which is considerably stronger than the 1st-order axiomatization with the induction schema. Gödel's results, in tandem with his completeness theorem for 1st-order logic, shows that 1st-order PA is not categorical (it has non-standard models), whereas 2nd order PA is categorical (all its models are isomorphic). But Gödel's incompleteness results apply to both versions; they apply to any consistent, recursively axiomatizable system with a certain amount of elementary arithmetic**.** The problem with the 2nd order axiomatization lies not in the 2nd order mathematical axioms (which are categorical) but rather in the incompleteness 2nd-order logic, which is unable, so to say, to extract the content of the mathematical axioms formalized in it.

[94] **Q** is due to R. Robinson (1952) and became widely known via the influence of Tarski et al

axiomatizable theory containing **Q** is subject to Gödel's two incompleteness theorems and their consequences (and to Tarski's indefinability of truth theorem as well). **Q** represents for us "elementary arithmetic" to which Gödel's theorems refer (owing to the fact that all primitive recursive functions, properties and relations are expressible in it, **Q** is sometimes dubbed a 'primitively recursively adequate arithmetic').[95]

There is a close analogy between Gödel's demonstration of 1st incompleteness theorem that concerns any consistent recursively axiomatized theory T of elementary arithmetic and Tarski's demonstration that the notion of truth for the language L(T) of such T cannot be defined in L(T) itself or, generally, in any language of the same logical strength as L(T). Both theorems exploit the circumstance that syntax of L(T), including the proof theory of T, can be arithmetized in Gödel's celebrated style,[96] each expression of L(T), sequence of expressions of L(T), sequence of sequences of expressions of L(T), etc., being assigned a unique number as its code in such way that it is possible to determine algorithmically: (a) for any given expression (sequence of expressions, etc.) of L(T), what number encodes is, (b) for any given number, what expression of L(T) or sequence (of sequence of ...) expressions it encodes. For the purposes of my exposition, I shall assume that there is a well-defined function *gn* (of Gödel numbering) satisfying (a) and (b) - the details need not detain us.[97] Gödel's original thought was that with a suitable coding-function chosen it turns out that to syntactic properties and relations between encoded expressions (sequences of expressions, etc.) there correspond certain number-theoretic properties and relations of a rather elementary character so that the whole syntax of L(T) (including the proof-theory of T) finds an interpretation in elementary arithmetic, and hence in T that *ex hypothesi* embeds it.

Let us begin by adopting the following notational conventions:

- *n*-th numeral, abbreviated as '$\underline{n}$', is an L(T)-expression of the form 'S(S(...(0)...)), obtained by applying the successor-functor *n*-times to '0'.

- if φ is a formula of L(T), then $\langle\varphi\rangle$ is the Gödel number of φ, and $\underline{\langle\varphi\rangle}$ is the numeral denoting in L(T) that Gödel number.

Today it is usual to demonstrate Gödel first incompleteness theorem by using the so-called *diagonal lemma* (here formulated only for 1-place formulas; it can be extended to *n*-place formulas):[98]

(1953), with Robinson one of the co-authors. Tarski et al prove for **Q** that it not only expresses (semantically defines) such functions but that it also represents them. Although **Q** contains primitive recursive arithmetic, it is not to be equated with it: primitive recursive arithmetic is rather its quantifier free fragment.

[95] See Boolos et al (2002), where a different minimal system is used that is called **Q**, while **R** is used to denote Robinson Arithmetic.

[96] The idea of arithmetization of syntax or metamathematics was developed independently but in much less advanced form by Tarski.

[97] See the exposition in Boolos et al (2002) or Smith (2007). Any standard textbook of mathematical logic such as Mendelson (1997) or Enderton (2001) provides the details.

[98] Sometimes called Gödel–Tarski (self-referential) lemma (Field, 2008), or, more accurately, Gödel-Carnap lemma. The latter seems more accurate, because Carnap (1934) was the first to

DIAGONAL LEMMA:

Let $\varphi(v)$ be a formula of the language L(T) of T embedding primitive recursive arithmetic, where '*v*' is its only free variable. Then there is a sentence *A* of L(T) such that

$$T \vdash A \leftrightarrow \varphi(\underline{\langle A \rangle}).$$

A is sometimes called the fixed point of $\varphi(v)$ in T, and the lemma accordingly *the fixed point lemma*. The existence of fixed points for T-formulas depends on the following facts holding about the expressive-deductive power of T:

(1) T is a *primitively recursively adequate arithmetic* in that every primitive recursive function is *represented* in T by a formula with an appropriate number of places.

(2) Every primitive recursive property and *n*-place relation is represented in T by an *n*-place formula, as these are relations whose characteristic functions are primitive recursive.

(3) In particular, T *represents* the primitive recursive function *diag(n)* that takes $gn(\varphi)$, for any formula φ with just one free variable, and maps it to $gn(\varphi(\underline{\langle \varphi \rangle}))$, where $(\varphi(\underline{\langle \varphi \rangle}))$ is the so-called diagonalization of the formula φ obtained by replacing all occurrences of φ's free variable with the L(T)-numeral of $gn(\varphi)$.

Thus, our notational convention tell us that the value of $diag(x)$ for $gn(\varphi) = \langle \varphi \rangle$ is $\langle \varphi(\underline{\langle \varphi \rangle}) \rangle$.

In the next step we explain what it means for a 1-place function - such as $diag(x)$ - to be represented in T by a 2-place formula $\phi(x, y)$ of L(T):

REPRESENTABILITY LEMMA:

The 1-place function $f(x)$ is represented in T by a 2-place open formula $\phi(x, y)$ just if:

---

extract the lemma in its generality from Gödel proofs. In (1931) Gödel derived undecidable sentences without appealing to the lemma, providing a direct self-referential construction of them, which procedure, however, involves the diagonal trick. See sections 4.3.4. and 4.3.5.

(i) shows features of the diagonalization procedure made famous by Cantor (e.g. in his proof that a power set of any set S is greater in magnitude than S) or by Richard (who used it to formulate his paradox of definability);

(ii) its crucial step can be seen as a particular application of the lemma (Gödel sentence can be viewed as a fixed point of the formula '*Pr*(*x*)' expressing the property of *provability in T*).

In his Princeton lecture (1934), Gödel credits Carnap for recognizing the importance of the diagonal (fixed point) lemma. Tarski is a similar case. Though he did not explicitly state or prove it, he implicitly uses the diagonal lemma in his proof of the indefinability of truth theorem in (1933), inspired by Gödel's (1931). See also the sections (4.3.4) and (4.3.5).

for any m, n: if $f(m) = n$, then T $\vdash \forall y(\phi(\underline{m}, y) \leftrightarrow y = \underline{n})$.[99]

In the last preparatory step we let the formula 'DIAG(x, y)' to T-represent in this precise sense the function *diag*(x). The proof of DIAGONAL LEMMA can now be run as follows.

Let $\varphi(v)$ be an arbitrary formula or L(T) with only the variable 'v' free in it, and let ψ be the formula:

(i) $\forall u[\text{DIAG}(v, u) \rightarrow \varphi(u)]$.

Now, Gödel number of ψ is $\langle\psi\rangle$. Let γ be the diagonalization of ψ, so that γ is

(ii) $\psi(\langle\underline{\psi}\rangle)$.

We then have:

(iii) $diag(\langle\psi\rangle) = \langle\gamma\rangle$.

As, by REPRESENTABILITY LEMMA, the function *diag*(x) is represented in T by the formula DIAG(x, y), we have:

(iv) T $\vdash \forall u[\text{DIAG}(\langle\underline{\psi}\rangle, u) \leftrightarrow u = \underline{\langle\gamma\rangle}]$.

Then we notice that γ is equivalent to

(v) $\forall u[\text{DIAG}(\langle\underline{\psi}\rangle, u) \rightarrow \varphi(u)]$,

and that, in particular, T proves this equivalence, by classical logic:

(vi) T $\vdash \gamma \leftrightarrow \forall u[\text{DIAG}(\langle\underline{\psi}\rangle, u) \rightarrow \varphi(u)]$.

---

[99] Gödel (1931) himself used the label „entscheidunsdefinite" (*vide* his theorem VI). Kleene (1976, 1983), who offers a lucid short-exposition of Gödel's result, talks about „numeralwise expressibility" of a relation in a formal system (theory). Smorýnski (1976) use the term „binumerates". Smith (2007) uses *capture* and has a very useful overview of differences in usage to be found in the relevant literature. He carefully distinguishes *expressibility* from *capturability* of a function (property, n-place relation), where the first is a matter of T containing a formula (with an appropriate number of free variables) whose extension coincides with the extension of the function (property, n-place relation). It is what Tarski would call *semantic definability*. We can say generally what it means for an n-place relation to be represented in T:

The n-place relation $Rx_1, ..., x_n$ is represented in T by $\phi x_1, ..., x_n$ just if:
(i) for any $m_1, ..., m_n$: $Rm_1, ..., m_n$ iff T $\vdash \phi\underline{m_1}, ..., \underline{m_n}$
(ii) for any $m_1, ..., m_n$: $\neg(Rm_1, ..., m_n)$ iff T $\vdash \neg(\phi\underline{m_1}, ..., \underline{m_n})$.

What we have proposed above is a special case of the following general definition of what Smith calls *capturability of a function, as a function*, by T. On the assumption that T contains **Q**, this is equivalent to the following:

The 1-place function f is represented in T by $\phi(x, y)$ just if:
(i) for every m: T $\vdash \exists!y\, \phi(\underline{m}, y)$,
and for any m, n:
(i) if $f(m) = n$, then T $\vdash \phi(\underline{m}, \underline{n})$,
(ii) if $f(m) = n$, then T $\vdash \neg\phi(\underline{m}, \underline{n})$.

From (vi) we obtain the following provable equivalence, by substitution of provable equivalents on the basis of (iv):

(vii) $T \vdash \gamma \leftrightarrow \forall u[u = \underline{\langle \gamma \rangle} \rightarrow \varphi(u)]$.

And this obviously yields, within classical logic, the desired T-provable equivalence:

(viii) $T \vdash \gamma \leftrightarrow \varphi(\underline{\langle \gamma \rangle})$ QED.[100]

Having the proof of DIAGONAL LEMMA in place, and realizing that φ in

$T \vdash A \leftrightarrow \varphi(\underline{\langle A \rangle})$,

may be any formula whatever of L(T), including negative formulas, we see that there must be a sentence γ of L(T) such that

$T \vdash \gamma \leftrightarrow \neg Pr(\underline{\langle \gamma \rangle})$

where '*Pr*(*x*)' is an arithmetical formula of L(T) encoding the syntactic property of *being provable in T*:

For every *n*: *Pr*($\underline{n}$) iff *n* is the Gödel number of a formula provable in T.

Gödel showed that the syntactic property of *T-provability* can be indirectly expressed in T by the numerical formula '*Pr*(*x*)', whose extension is the set of Gödel numbers of T-theorems, owing to the fact that the property is coextensional with the syntactic property of *there being an y such that y is a T-proof of x*. Now, the important moment is that syntactic relation *y is a T-proof of x* is not just expressed but is represented in T by a numerical formula '*Prf*(*x,y*)', whose extension is the set of all ordered pairs $\langle n, m \rangle$ such that *n* encodes a sequence of formulas that form a T-proof of the formula encoded by *n*.[101] What Gödel showed is that there is a sentence γ of L(T) provably equivalent in T to the sentence saying that γ is unprovable in T. Popularly speaking: γ indirectly says of itself that it is unprovable in T. We can now finish Gödelian argument by two

---

[101] Gödel proved that the syntactic proof-relation in T, in its arithmetical encoding, is primitive recursive (because its characteristic function is primitive recursive), together with other 44 functions, properties or relations that he considered in the course of his investigations. The important exception is the 46th property of *T-provability* or *T-theorem*, expressed in T by the formula '*Pr*(*x*)' = '$\exists y$*Prf*(*y, x*)'. This property is only *weakly representable* in T in that only the positive part of the definition of T-representability of relations holds for it (for any *n*: if *n* codes a T-provable sentence, then $T \vdash Pr(\underline{n})$). The fact that the diagonal function *diag(x)* is T-representable is based on the fact that (a) T is primitively recursively adequate theory of arithmetic and that (b) *diag*(*x*) is a primitive recursive function. The proof of (a) and (b) is crucial to Gödel's proof of the incompleteness of T, and it requires quite rigorous definitions of primitive recursive functions and T-representability of functions. Although Gödel did not explicitly formulate the diagonal lemma in his (1931a), he provided precise characterizations of both primitive recursive functions and of T-representability there.

mini-proofs:

> a) Let us suppose that T proves γ. But if so, T proves falsehood, since γ is provably equivalent in T with the sentence that says that γ is not a T-theorem. Consequently, T does not prove γ or T is not sound (under the intended interpretation). QED
>
> b) Since, by (a), γ is not a T-theorem, γ is true, since it is provably equivalent in T with the sentence that says that γ is not a T-theorem. So, the negation of γ must be false. Consequently, T does not prove the negation of γ or T is not sound (under the intended interpretation). QED

From (a) and (b) it follows that T is an incomplete theory, since it is not the case that, for every sentence of L(T), T proves it or T proves its negation. For mathematical theories such as **Q**, under its intended interpretation, soundness is a rather natural assumption to make. Indeed, Gödel used it in his expository lecture (1930 ?) as well as in his informal explanations in letters to Zermelo (1931-32). But he made it clear that the assumption can be weakened to ω-consistency of T, defined as a syntactic property of T (ω-consistency implies simple consistency, but not *vice versa*).[102] Then the mini-proofs (a) and (b) can be replaced by the following proofs that can be fully formalized within T:

> a*) Suppose T $\vdash \gamma$. Then there is n such that n codes T-proof of γ. Then T $\vdash$ *Prf*($\underline{n}$, $\langle\underline{\gamma}\rangle$), in accordance with T-representability of the primitive recursive relation of *T-proof* by *Prf*(x, y). Hence T $\vdash \exists x$*Prf*(x, $\langle\underline{\gamma}\rangle$). If so, T $\vdash$ *Pr*($\langle\underline{\gamma}\rangle$). By the diagonal lemma, T $\vdash \gamma \leftrightarrow \neg$*Pr*($\langle\underline{\gamma}\rangle$), hence T $\vdash \neg\gamma$. Consequently, it is not the case that T $\vdash \gamma$, or T is inconsistent. QED
>
> b*) Suppose T $\vdash \neg\gamma$. Then, for every n, n does not code T-proof of γ. So, for every n, T $\vdash \neg$*Prf*($\underline{n}$, $\langle\underline{\gamma}\rangle$), in accordance with T-representability of *T-proof* by *Prf*(x, y). The first assumption together with the diagonal fact that T $\vdash \gamma \leftrightarrow \neg$*Pr*($\langle\underline{\gamma}\rangle$) yields: T $\vdash$ *Pr*($\langle\underline{\gamma}\rangle$), hence T $\vdash \exists x$*Prf*(x, $\langle\underline{\gamma}\rangle$). But this makes T ω-inconsistent. Consequently, it is not the case that T $\vdash \neg\gamma$, or T is ω-inconsistent.

---

[102] Here are the explanations:

> A) A theory T of arithmetic is ω-inconsistent if, for some open formula $\phi(x)$, T proves $\phi(\underline{n})$ for each n, and T also proves $\neg\forall x\phi(x)$ (equivalently: …if, for some open formula $\phi(x)$,T proves $\exists x\phi(x)$, and it proves also $\neg\phi(\underline{n})$, for each n.
>
> B) A theory T of arithmetic is ω-consistent if there is no open formula $\phi(x)$ such that when T proves $\phi(\underline{n})$ for each n, T also proves $\neg\forall x\phi(x)$.

They are closely related with the following notions:

> C) A theory T of arithmetic is ω-incomplete if, for some open formula $\phi(x)$, T proves $\phi(\underline{n})$ for each n, but T does prove $\forall x\phi(x)$.
>
> D) A theory T of arithmetic is ω-complete if there is no open formula $\phi(x)$ such that T proves $\phi(\underline{n})$ for each n, yet T does not prove $\forall x\phi(x)$.

Rosser (1936) proved that the assumption of ω-consistency of T can be weakened to the assumption of simple consistency of T, provided we choose a more complicated version of Gödel's self-referential sentences.

QED

Gödel stressed that the weakening of the assumption of soundness of T (under the intended interpretation) is called for, if the whole procedure is to satisfy also the finitists and constructivists, who would protest against any appeal to the notion of "objective" mathematical truth that is not reducible to proof-theory. We saw that such an attitude towards truth was not at all uncommon in the 1920s. As Gödel later explained, the idea of "transcendent" or "transfinite" notion of mathematical truth was the principle behind his discovery:

> "[...] it should be noted that the heuristic principle of my construction of undecidable number theoretical propositions in the formal systems of mathematics is the highly transfinite concept of "objective mathematical truth" as opposed to that of "demonstrability". (Wang 1974: 9)

Yet, for the reasons spelled out above, he did not want it to enter the demonstration in the guise of the assumption of soundness of T that he could afford to make in his informal explanations. For invoking the transfinite notion of mathematical truth – in one form or another – in order to demonstrate incompleteness of T would amount to begging the very question at issue.

> "[...] in consequence of the philosophical prejudices of our time 1. nobody was looking for a relative consistency proof because [it] was considered axiomatic that a consistency proof must be finitary in order to make sense, [and] 2. a concept of objective mathematical truth as opposed to demonstrability was viewed with greatest suspicion and widely rejected as meaningless." (a letter to Y. Balas, in Wang 1987: 85)

This might well be the reason why he did not state in 1931 something he arguably discovered along the way, and quite independently of Tarski: namely that arithmetical truth (truth in the arithmetical L(T)) is not definable within L(T) itself). This metatheoretical result is usually associated with Tarski's name, and we shall see shortly how he obtained it. But, by all available evidence, Gödel got it completely independently as the following passage from the same letter indicates:

> "...long before, I had found that the correct solution of the semantic paradoxes in the fact that truth in a language cannot be defined within itself." (Ibid: 85)

In the correspondence with A. W. Burks Gödel says there that he got his incompleteness theorem by having found out that truth for a sufficiently powerful mathematical language is indefinable within that language itself:

> "a complete epistemological description of a language A cannot be given in the same language A, because... the concept of truth of sentences of A cannot be defined in A. It is this theorem which is the true reason for the existence of undecidable propositions in the formal systems containing arithmetic. I did not, however, formulate it explicitly in my paper of 1931 but only in my Princeton lectures

> of 1934. The same theorem was proved by Tarski in his paper on the concept of truth ..." (a letter to A. W. Burks, in von Neumann 1966: 55-56)

As a matter of fact, already in his correspondence with Zermelo (in the period of 1930-31) Gödel made it quite clear that he discovered his theorem by realizing that arithmetization of syntax plus diagonalization show that truth for a sufficiently rich L(T) is indefinable within T, on pain of semantic antinomy of Liar-type.[103] By his own words, he realized that once we substitute in the diagonal construction the provability predicate for the truth predicate something close to paradox results, which, however, is not a genuine paradox, but his 1st incompleteness theorem. According to Gödel, his diagonal argument shows that the set *Tr* of (Gödel numbers of) L(T)-truths is not arithmetical (there being no arithmetical formula whose extension is *Tr*), on pain of inconsistency. But the set *Pr* of (Gödel numbers of) T-theorems is arithmetical. Now, assuming soundness of T, *Pr* is included in *Tr*, but not *vice versa*: there are L(T)-sentences that are true but unprovable in T (otherwise *Pr* and *Tr* would coincide). Since such L(T)-sentences are true, their negations are unprovable in T, if T is sound. Hence Gödel's 1st incompleteness result (in its informal version): there are true L(T)-sentences unprovable in T. All this is nicely explained in Gödel's letters to Zermelo as well as in the description of his discoveries that Gödel sent to Wang.[104] We shall see that Tarski had essentially the same idea, which he based it explicitly on Gödel's first theorem. Interestingly, in the correspondence with Bernays,[105] Gödel suggests a satisfactory definition of truth for the language of arithmetic (in a more powerful system). What Gödel says is that once '*Tr*' is defined for atomic arithmetical sentences, it can be recursively defined roughly as follows: if *A* and *B* are formulas, then

(a) $Tr(\neg A)$ iff $\neg Tr(A)$;
(b) $Tr(A \vee B)$ iff $Tr(A) \vee Tr(B)$;
(c) $Tr(\forall x\, A(x))$ iff $Tr(A(n))$, for every numerical constant *n*.

This is truly interesting, as it anticipates Tarski's truth definition as well as Carnap's definition of analyticity that we shall discuss in section (4.5). Granted, Gödel did not tell us what it takes for an atomic sentence to be true, but, I take it, it is highly likely that he had in mind clauses such as the following:

If *a* and *b* are numerical constants, then:

$Tr(a = b)$ iff $v(a) = v(b)$,

'v( )' being a function that assigns names their numerical values. If this diagnosis is on the right track, then Gödel knew, independently of Tarski, how to define arithmetical truth in the recursive manner, though he did not show how to extend such a procedure also to languages that do not contain a name for every object in their associated domain. It took Tarski's efforts to finish the task by devising his method of defining satisfaction relation.

---

[103] Gödel (2003a: 427-429).
[104] Wang (1987).
[105] Gödel (2003a: 95).

4.2.2 Gödel's second theorem

Gödel's second incompleteness theorem hinges on the fact that consistency of T is a syntactical property of T, encoded by an arithmetical formula Con_T of L(T). It can be e.g. a formula '$\neg Pr(\underline{\langle 0 = 1 \rangle})$' that says that no number encodes T-proof of T-formula encoded by '$\underline{\langle 0 = 1 \rangle}$'. The rationale behind this choice is that since T is adequate to elementary arithmetic, it surely proves something so elementary as '$\neg(0 = 1)$', and so it cannot also prove '$0 = 1$', on pain of being inconsistent. It is to be noted that the exact reasoning leading to the Gödel's 1st incompleteness theorem based on the assumptions of simple consistency or ω-consistency of T can be reconstructed in T, its conclusions being represented by the formal counterparts of:

a*) $\quad T \vdash Con_T \rightarrow \neg Pr(\underline{\langle \gamma \rangle})$

b*) $\quad T \vdash \omega\text{-}Con_T \rightarrow \neg Pr(\underline{\langle \neg \gamma \rangle})$,

where Con_T and $\omega\text{-}Con_T$ express respectively the simple consistency and ω-consistency of T. One can ask whether T can prove its own simple consistency via proving Con_T, that is, via proving '$\neg Pr(\underline{\langle 0 = 1 \rangle})$'. The positive answer to this question was expected by Hilbert and his allies in case of a theory embedding primitive recursive arithmetic. But the question has a negative answer, as Gödel informally proved (independently of him also von Neumann, inspired by Gödel's presentation of his first theorem at the mathematical congress that took place in Königsberg 1930):

> Suppose that Con_T is provable in T. Then, by (a*), $\neg Pr(\underline{\langle \gamma \rangle})$ is provable in T. Then γ is provable in T, since $\neg Pr(\underline{\langle \gamma \rangle})$ is provably equivalent in T to γ. But this contradicts the previously established result that neither γ nor its negation is provable in T. Consequently, T does not prove Con_T, or it is inconsistent.

Gödel's demonstration of his second incompleteness theorem in (1931) was this informal. Although he advertised that he will give a fully formalized proof of it on a par with the proof of the 1st theorem, he never did that, the reason being that the result became meanwhile widely accepted even among the die-hard formalists, whose programme was directly attacked by it. By an irony of fate, it was not Gödel but Hilbert and Bernays, who produced the very first formally rigorous proof of the Gödel's 2nd theorem in their monumental joint work *Grundlagen der Mathematik* II (1939).

4.2.3. Tarski's *indefinability of truth theorem*

Tarski's indefinability theorem can be demonstrated in a *reductio ad absurdum* style via DIAGONAL LEMMA. We start by assuming that we can explicitly define a predicate '*Tr*' in T so that the condition of material adequacy is satisfied. That is to say, we assume that we have

$T \vdash Tr(\underline{\langle \varphi \rangle}) \leftrightarrow \varphi$

for any sentence φ of L(T).[106] The important thing now is that DIAGONAL LEMMA applies, so that there will be a sentence γ of L(T) such that:

$$T \vdash \gamma \leftrightarrow \neg Tr(\underline{\langle \gamma \rangle})$$

But given that the definition of '*Tr*' is assumed to satisfy the condition of material adequacy, we also have:

$$T \vdash \gamma \leftrightarrow Tr(\underline{\langle \gamma \rangle})$$

If so, then by elementary logic we obtain:

$$T \vdash Tr(\underline{\langle \gamma \rangle}) \leftrightarrow \neg Tr(\underline{\langle \gamma \rangle})$$

And this yields contradiction in classic logic.

The reduction to absurdity of the assumption that T adequately defines the truth-predicate for its own language L(T) is thus completed: if T is recursively axiomatizable and embeds elementary arithmetic, T defines '*Tr*' in manner satisfying the condition of material adequacy only if T is inconsistent. By contraposition, then:

> **(Tarski's indefinability of truth theorem – syntactic version):**
>
> No consistent recursively axiomatizable theory T embedding elementary arithmetic can define the notion '*Tr*' for L(T) so that the condition of material adequacy is satisfied.

Observe that the theorem can be generalized as follows:

> Let T be a consistent and recursively axiomatizable theory embedding elementary arithmetic. Then:
>
> (a) T cannot define '*Tr*' for L(T) in manner satisfying the condition of material adequacy;
>
> (b) T cannot contain '*Tr*' for L(T) as a primitive notion in manner satisfying the condition of material adequacy.

Clearly, (b) in the generalized theorem excludes the possibility that T can provide, at the very least, an adequate axiomatization of truth for L(T) in the sense of having all T-biconditionals for L(T) among its deductive consequences.

There is a related indefinability theorem concerning truth that is also associated with Tarski's name (and sometimes attributed to Gödel), which assumes:

---

[106] The difference between this assumption and Tarski's original proof-sketch is that he assumed '*Tr*' to be definable in the metatheory MT framed in ML not essentially stronger than L(T). However, due to this circumstance, MT is translatable into (or interpretable in) T, and the argument for indefinability of truth that Tarski offered covers as a special case the assumption that truth for L(T) is definable within L(T) itself – which is how Tarski's indefinability theorem is usually understood and presented today

> '*Tr*' is a (formal) truth-predicate for L(T) iff for every sentence γ of L(T) we have: $Tr(\underline{\langle\gamma\rangle}) \leftrightarrow \gamma$.

and states that:

> **(Tarski's indefinability theorem – semantic version):**
>
> No 1-place predicate of a language L(T) embedding elementary arithmetic can express the property of *being a (Gödel number of) a true sentence of L(T)*;

or

> No 1-place predicate of a language L(T) embedding elementary arithmetic can semantically define the set of (Gödel numbers of) true sentences of L(T).

As Smith put it in his comprehensive study of Gödel's theorems,[107] the former theorem shows the limits of what can be proved in T about truth and related semantic properties of L(T), whereas the later theorem shows the limits of what can be expressed in T about such properties. The informal proof is again very simple, using only one additional assumption that T is sound and proves only truths. We are to suppose, for *reductio*, that L(T) expresses, via a sentential function '$Tr(x)$', the property of *being a (code of a) true sentence of L(T)*. Then the diagonal lemma tells us that, for some sentence γ of L(T):

$$T \vdash \neg Tr(\underline{\langle\gamma\rangle}) \leftrightarrow \gamma.$$

But, by the extra-assumption, we also have:

$$\text{If } [T \vdash \neg Tr(\underline{\langle\gamma\rangle}) \leftrightarrow \gamma], \text{ then } [\neg Tr(\underline{\langle\gamma\rangle}) \leftrightarrow \gamma]$$

and hence:

$$\neg Tr(\underline{\langle\gamma\rangle}) \leftrightarrow \gamma.$$

But this obviously contradicts the requirement that when '$Tr(x)$' is an adequate truth-predicate for L(T), then:

$$Tr(\underline{\langle\gamma\rangle}) \leftrightarrow \gamma.$$

So, by way of conclusion, it can be said that L(T) cannot even express (semantically define) its own adequate truth-predicate.

4.2.4. Tarski's original proof-sketch and the method of *diagonalization*.

Like Gödel in (1931), in CTFL Tarski did not provide the proof that we have given above, as he did explicitly mention DIAGONAL LEMMA, though he talked about the diagonalization or diagonal method. The object language (theory) he considered was that of the general calculus of classes, the metatheory

[107] Smith (2007).

being almost identical with it except that it contained some extra-syntactical baggage needed to talk about the structure of object language. The two are at any rate logically on a par, so that one can say that Tarski actually investigated what happens if we attempt to define truth for a language such as LGC within the language itself. All the more so that the metalanguage finds an interpretation in the object language, because the extra-baggage can be "arithmetized" according to Gödel-Tarski recipe, and arithmetic can in turn be developed on the basis of the higher order LGC, say, in the set-theoretical manner of Russell.

In what follows, I have simplified Tarski's informal sketch-proof to make it closer to Gödel's informal demonstration to be reviewed in the next section. Suppose, for *reductio*, that we have introduced into ML via the definition an adequate truth-predicate '*Tr*' for L(T), whose extension it the set of true sentences of L(T). We assume that all expressions of L(T) are enumerated in an infinite sequence φ without repeating terms so that every 1-place formula occurs somewhere in φ. Thus, to every sentence and formula of L(T) there corresponds a unique number *n*, according to its position in the sequence φ - the formula occupying the *i*-th position in φ being referred to as φ_i. Given that ML contains L(T) and T embeds elementary arithmetic, ML can be interpreted in T: there is an arithmetization of ML in T can be given such that to every ML-sentence there is an equivalent L(T)-sentence.[108] Let us now consider the following sentence of ML

(1) $\varphi_n(n) \notin Tr$,

which says, in effect, that the *n*-th formula of L(T) is not true for the argument *n*. This is of course a formula of ML but the method arithmetization assures us that there is a purely arithmetical formula'$\psi(n)$' of L(T) that is equivalent to it for every argument *n*, so that we have

(2) For every n [$\varphi_n(n) \notin Tr$ iff $\psi(n)$]

Since '$\psi(n)$' is a purely arithmetical must occur somewhere in the sequence φ and accordingly - say, being its *k*-th term - so that we have '$\psi(n)$' = φ_k. At this juncture, the crucial *diagonal move* comes, for Tarski invites us to instantiate (2) to *k*, thereby obtaining the tricky sentence:

(3) $\varphi_k(k) \notin Tr$ iff $\psi(k)$

[108] Tarski did not bother to spell out the details but I guess that what he had in mind is this. First, ML is assumed to contain L(T) as its part. Indeed, since, by assumption, the logical part of L(T) and ML is the same, the only expressions that ML has in addition to those that it shares with L(T) are the structural-descriptive expressions needed to study the „morphology" or syntax of L(T): expressions for purely formal operations on expressions (sequences of expressions,…), properties of expressions (sequences of expressions,…), and relations between expressions (sequences of expressions,…). So, given that we have uniquely assigned numbers to expressions of L(T) via ordering the later in the sequence φ, we have thereby assigned numbers to them, hence arithmetized the part of ML that coincides with L(T). Now, since we know - owing to Gödel and Tarski himself - that structural-descriptive (syntactic) notions can be arithmetized without residue so that the syntactic extra-part of ML (MT) can be interpreted in arithmetic as well, we have in a way interpreted the whole ML (MT) in arithmetic. But since, *ex hypothesi*, arithmetic can be developed within T, we have found an interpretation of ML (MT) in T (e.g. in the general calculus of classes that Tarski considers).

What the sentence on the left side of (3) says is that the *k*-th formula of L(T) is not true for the argument *k*. Since '$\varphi_k(k)$' designates a sentence of L(T), MT (augmented with the *adequate* truth-predicate '*Tr*' for L(T)) should prove for that sentence the material adequacy condition of the type,

$\varphi_k(k) \in Tr$ iff ____ ,

where the blank is to be filled in by its translation. By the stipulations that we have made, it should be clear that the desired T-biconditional is this:

(4) $\varphi_k(k) \in Tr$ iff $\psi(k)$.

But (3) and (4) yield a contradiction!

(3) – the output of diagonalization - is a paradoxical sentence reminiscent of Gödel's '$\gamma \leftrightarrow \neg Pr(\underline{\langle\gamma\rangle})$'. Yet, we have seen that Gödel's sentence does not really give rise to contradiction (being unprovable, hence true). Tarski's sentence does, combined with the desideratum that MT, in tandem with the adequate truth definition for L(T), proves T-biconditionals for all sentences of L(T). Reading the argument as a *reductio ad absurdum*, some assumption or step has to go: namely, Tarski says, the assumption that the symbol '*Tr*' that we have introduced into MT via definition is an adequate truth-predicate for L(T).

4.3.5 Gödel's theorems in Tarskian setting

When Gödel explained his procedure in the short informal exposition,[109] he did so in a strikingly similar style, which reveals what Tarski could have learned from him. We are to make much the same assumptions as in Tarski's informal proof, except that we do not assume that we have introduced the notion of truth for L(T) into MT, but work with the provability predicate '*Pr*'. We are now to consider the following formula of ML:

(1*) $\neg Pr(\varphi_n(n))$

It says that the formula of L(T) with the number *n* is not provable for the argument *n*. The method arithmetization of ML in T assures us that there is a purely arithmetical formula of T equivalent to this ML-formula for every *n*. We can call that formula '$\psi(n)$' and assign it a numerical index *k*, according to its place in the sequence φ, so that '$\psi(n)$' = φ_k. Consequently, we have:

(2*) For every *n* [$\neg Pr(\varphi_n(n))$ iff $\varphi_k(n)$]

The rest of the proof runs as before. Instantiating (2*) with respect to *k* – which is the diagonal move - we end up with the Gödel-type sentence:

(3*) $\neg Pr(\varphi_k(k))$ iff $\varphi_k(k)$

Having this in place, Gödel produces an informal argument for T-unprovability of both '$\varphi_k(k)$' and '$\neg\varphi_k(k)$', assuming soundness of T. Though DIAGONAL LEMMA is neither stated nor proved, it seems to be involved (in application) in

[109] Gödel (1930?).

the step from (2*) to (3*).

Familiarity with Gödel's work on incompletability of consistently axiomatized theories embedding elementary arithmetic inspired Tarski to form a more compact picture of the connections between truth and proof. He was then not only able to provide a precise proof of indefinability of truth for L(T) within L(T) (he concedes that previously he gave only some hints in this direction), but he quickly realized that Gödel's results follow from the indefinability-of-truth theorem. If Gödel proved that '*Pr*(*x*)' is an arithmetical formula of L(T) holding of all and only the Gödel-numbers of sentences of L(T) provable in T, and if Tarski showed that there is no arithmetical formula in L(T) holding of all and only the true sentences of L(T), it follows that some true sentence of L(T) is not a provable sentence of T. Tarski saw that these important results find natural explication in the setting of his theory of truth for formalized languages, all of whose main principles were already formulated (it was in preparation since 1929), except for the fundamental result about the indefinability of truth for a sufficiently powerful formalized language within that language itself. He showed that on the basis of his truth-theory for such a formalized language it is possible to prove soundness or consistency of the deductive theory framed in it (PA), but that this is possible only because the truth theory itself is framed in a higher order metalanguage. Now that squares well with Gödel's 2nd theorem and his own claim that consistency of T can be proved but we need for that the means not available in it. But there are intimate connections also to Gödel's 1st theorem:

> "Moreover Gödel has given a method for constructing sentences which—assuming the theory concerned to be consistent — cannot be decided in any direction in this theory. All sentences constructed according to Gödel's method possess the property that it can be established whether they are true or false on the basis of the metatheory of higher order having a correct definition of truth." (Tarski 1935: 274)

We can start the argument once we have constructed (3*). Assuming that we have an adequate definition of the set *Tr* of truths of L(T) in ML, we can prove in T the T-biconditional for the Gödel sentence:

(4*) $Tr(\varphi_k(k))$ iff $\varphi_k(k)$

Which, together with 3*) entails

(5*) $\neg Pr(\varphi_k(k))$ iff $Tr(\varphi_k(k))$

The truth definition will also give us (for more details see the next paragraph):

(6*) $\neg Tr(\varphi_k(k))$ or $\neg Tr(\neg\varphi_k(k))$

(7*) If $Pr(\varphi_k(k))$, then $Tr(\varphi_k(k))$

(8*) If $Pr(\neg\varphi_k(k))$, then $Tr(\neg\varphi_k(k))$

From this basis, we can easily derive in MT the following three conclusions that

together show that '$\varphi_k(k)$' is true (*vide* 6*) yet undecidable sentence in Gödel' sense (*vide* 7* and 8*):

(9*) $Tr(\varphi_k(k))$

(10*) $\neg Pr(\varphi_k(k))$

(11*) $\neg Pr(\varphi_k(k))$

Since we have just proved undecidability of Gödel sentence in T, we have thereby proved its truth as a sentence of L(T). In proving that Gödel's sentence is undecidable, hence unprovable, we have proved, hence decided Gödel's sentence after all. But is this not paradox? No, if we carefully distinguish two senses of 'prove' and 'decide' here. What we have *proved in MT*, hence *decided in MT* is that Gödel's sentence cannot be *decided in T*, hence cannot be *proved in T*. And we have thereby also *proved in MT* that Gödel's sentence is true in L(T), hence *we decided in MT* that sentence.

4.4 Definitions and axiomatizations of truth (semantics)

Let us now pay a closer attention to the import of the so far neglected thesis C, which, according to Tarski's own words "loses its importance" in light of the new theses (A) and (B) in the Postscript. Why? Well, the moral of C was that even in case when ML is not essentially stronger (higher order) than L(T), and the preferred procedure of explicit truth definition is thus not available, at least a part of the task expected from it could be attained by extending MT by a set of axioms that specify the basic properties of the notion of truth with respect to L(T), which is materially adequate, since its deductive consequences include all T-biconditionals for L(T). Here the material adequacy is achieved in a cheap way: we add '*Tr*' to MT as its primitive predicate and then to MT's axioms all the instances of T-schema for L(T).[110] At any rate, the trick consists in adding to MT the infinite set of special axioms (let us call it TRUE) that contains all and only the T-biconditionals for L(T) (for MT extended by TRUE we shall write MT$\cup$TRUE).

In (1933a) Tarski saw the value of such axiomatizations in the metalanguage of the same order in the circumstance that they provide compensation for languages of infinite order, for which, he argued, there was no possibility of constructing adequate truth definitions in higher order languages, because he did not then allow languages of transfinite order. Now, the moral of the Postscript is that there is no principal need for axiomatic truth-theories, once it was made clear that we can always ascend to a higher (transfinite) order language, and on the basis of the metatheory framed in it construct an adequate truth definition for the object-language, with all its advantages that axiomatizations cannot claim.

What advantages did he specifically have in mind? Let me approach this by asking what disadvantages pertain to truth-axiomatizations. There is a telling passage from Tarski's popular lecture (1936b), which deserves to be

[110] Perhaps, by adding to it T-schema generating them as when the induction schema is added to the axioms of **Q**, say.

quoted in full. Having explained what the materially adequate concept of truth for L(T) in MT amounts to and that it can be introduced into MT either via axiomatization or via its explicit definition, Tarski goes on to specify a couple of disadvantages of the axiomatic method. He first mentions the problem of a somewhat "accidental character" of axioms (on which he does not further elaborate), and then states what seem to be more worrisome aspects: [111]

> "Moreover, the question arises whether the axiomatically constructed semantics is consistent. The problem of consistency arises, of course, whenever the axiomatic method is applied, but here it acquires a special importance, as we see from the sad experiences we have had with the semantical concepts in colloquial language." (Tarski 1936b: 405-406)

A more serious disadvantage that Tarski mentions is that with the truth-axiomatization for L(T) in MT the question of its consistency remains in a way open, whereas with explicit truth definitions couched in non-semantic notions the question is immediately solved, the definition being conservative over the base theory MT, whose notions are used in the *definiens* (note: MT must be of higher order than L(T)). This assures that if the base theory MT is consistent, it does not cease to be so after we extend it by the explicit definition of '*Tr*' (for L(T)). Thus, the explicit definition of truth gives us immediate, if relative guarantee that the introduced notion and theory build around it is consistent. But the truth-axiomatization does not seem to give us any guarantee of consistency.

The matter, however, is more delicate than the foregoing remarks may betray. Yes, Tarski reports the problem of consistency of truth-axiomatizations as open and in his summary he explicitly states that in the thesis C (and C', generalized to cover the axiomatization of semantics in general, and not just of truth). However, the fact is that already in (1933) he states that it can be proved (*sic!*) that MT∪TRUE is consistent, provided that MT is consistent:

> "THEOREM III. *If the class of all provable sentences of the metatheory is consistent and if we add to the metatheory the symbol 'Tr' as a primitive sign, and all theorems which are described in conditions (α) and (β) of the convention T as new axioms, then the class of provable sentences in the metatheory enlarged in this way will also be consistent*." (Tarski 1935: 256),

drawing on the Theorem II:

> "THEOREM II. *For an arbitrary previously given natural number k, it is possible to construct a definition of the symbol 'Tr' on the basis of the metatheory, which has among its consequences all those sentences from the conditions (α) of the convention T in which in the place of the symbol 'p' sentence with variables of at most the k-th order occur (and moreover, the sentence adduced in the condition (β) of this convention)*" (Tarski 1935: 255)

What Theorem II states, in effect, is that for any sub-language L_k(T) of

[111] Compare a very similar passage in Tarski (1935: 255).

L(T) of finite order (for which there is a finite bound *k* upon the orders of variables of its sentences) it is possible to construct an adequate definition of truth on the basis of MT. It follows that MT augmented with the complete set of truth axioms for L_k(T) (this being a subset of TRUE) is consistent (if MT is), since the adequate truth definition for L_k(T) in MT amounts to a relative interpretation of that truth theory in MT.[112] Since in any finite subset S of TRUE only finitely many L(T)-sentences will appear, in which only finitely many variables occur, whose order is therefore not greater than some natural number *i*, any finite set of sentences of L(T) corresponding to such a subset S of TRUE forms a sub-language L_i(T) of finite order. Any such subset S of TRUE is thus consistent (if MT is consistent), because a truth definition for L_i(T) can be constructed in MT that has all T-biconditionals belonging to S among its deductive consequences, and this amounts to a relative interpretation of S in MT. So, for any finite subset S of TRUE we have: MT∪S is consistent, if MT is consistent. Having this in place, Tarski suggests the following compactness style reasoning to demonstrate Theorem III:

> If **(1)** MT∪TRUE is inconsistent, then, by compactness theorem, **(2)** there is a finite subset of MT∪TRUE that is inconsistent. But it follows from Theorem II that **(3)** every finite subset of MT∪TRUE is consistent, if MT is consistent. Consequently
>
> **(7)** MT∪TRUE is consistent, if MT is consistent. QED

As several people have recently noted,[113] Theorem III implies that MT∪TRUE is a conservative extension of MT. In view of this, one may wonder what remains of the alleged disadvantage of truth axiomatizations *vis-á-vis* the problem of consistency. In Theorem III, ML is assumed to be of the same order as L(T), and the theorem is primarily intended to deal with those cases, in which we not only *do not* but *cannot* have recourse to a higher order metatheory.[114] This, according to Tarski, happens with languages of infinite order. Of course, the theorem applies also to cases when it is possible in principle to ascend to a higher-order metatheory, but in such cases it does not seem to have much value, because we can give explicit truth definitions. But in light of Theorem III and its informal Tarski's proof sketched above, it seems to me that, despite his misleading remarks, consistency is not much of the problem for truth-axiomatizations after all - not, at least, for truth-axiomatizations on the pattern of MT∪TRUE.

That is not to say that Tarski did not have other good reasons for preferring the explicit truth definition for L(T) in MT (call such definition $D_{L\text{-}TR}$), as based (a) on the syntactic theory or morphology of L(T), including the proof-theory for T, and (b) on the recursive definition of satisfaction relation. His considered reason was that the higher-order MT expanded by $D_{L\text{-}TR}$ (shortly:

---

[112] Relative interpretation of the target theory T in the base (or background) T* is effected when all T-primitives of T are defined in terms of L*-primitives so that the axioms o T become theorems of T*. If then T were inconsistent, the contradiction would have to be derivable from the axioms of T*. So, if no contradiction is derivable in T*, then T is consistent.

[113] See also Heck (1997) and Ketland (1999).

[114] According to the conclusion of (1933a) that is retracted in the Postscript.

$MT \cup D_{L\text{-}TR}$) makes it possible not just to formulate but to prove important principles governing the notion of truth for L(T) and its relation to the notion of provability (in T), which should strengthen our conviction that the proposed Tarskian definition of truth is materially adequate. To see what is at stake, it will be useful to follow Tarski's own notational conventions: '*S*' will denote the set of sentences of L(T), '*Tr*' the set of true sentences of L(T), '*AX*' the set of axioms of T, '*C(X)*' the set of deductive consequences of an arbitrary set *X* of sentences of L(T), '*Pr*' the set of T-provable sentences, and '*Conj*(*x*, *y*)' and '*Disj*(*x*, *y*)' means 'the conjunction of *x* and *y*' and 'the disjunction of *x* and *y*' respectively. Such notions (or, if you prefer, sets) were defined in quite a rigorous manner, and, except for truth, all were defined on the basis of the morphology of L(T), which, we know, has interpretation in T via the method of arithmetization of metamathematics. Among the most basic principles that directly follow from $MT \cup D_{L\text{-}TR}$ are, first, certain recursive principles governing truth, which are generalizations of the recursive clauses in the definition of satisfaction for sentential functions of L(T):

(I) For any $x \in S$ and $y \in S$: $Conj(x, y) \in Tr$ iff $x \in Tr$ and $y \in Tr$.

(II) For any $x \in S$ and $y \in S$: $Disj(x, y) \in Tr$ iff $x \in Tr$ or $y \in Tr$.

Elementary as they are, they give us assurance that what we have defined is really the notion of truth. Other fundamental generalizations that can be derived from the whole machinery of truth-definition (as based on the morphology of L(T) and the definition of satisfaction-relation for it) in collaboration with the definitions of proof-theoretic notions (formulated in §2 of CTFL – viz. the definitions 13-20):[115]

(III) For any $x \in S$: either $x \in Tr$ or $\underline{x} \in Tr$;[116]

(IV) For any $x \in S$: either $x \notin Tr$ or $\underline{x} \notin Tr$;

(V) $AX \subseteq Tr$ (all the axioms of T are true);

(VI) If $X \subseteq Tr$, then $C(X) \subseteq Tr$; in particular: $C(Tr) \subseteq Tr$;

(VII) $Pr \subseteq Tr$ (soundness of T);

(VIII) *Tr* is complete and consistent;

(IX) *Pr* is consistent.

(VII) states the soundness of T and it follows directly from (V) and (VI), both of which, Tarski says, can be proved in MT without going into great

[115] They are, in this order, the definitions of (13) axiom, (14) substitution operation (free variables for free variable) of a sentential function, (15) the class of consequence of *n*-th degree of the class *X*, (16) the class of consequences of *X* - *C(X)*, (17) the class of provable sentences – *Pr*, (18) deductive system, (19) consistent class of sentence and (20) complete class of sentences. Some lemmas are also needed such as the free variable lemma and its direct consequence: if a sentence is satisfied by one infinite sequence it is satisfied by all infinite sequences.

[116] Henceforth: '$\underline{x}$' means 'the negation of *x*'.

pains. Similarly, (VIII) follows from (III) (the principle of excluded middle), (IV) (the principle of non-contradiction) and (VI). Having all this in place, we observe that (IX) follows directly from (IV) and (VII), which means that consistency of T is provable in MT. Indeed, the principle of non-contradiction (IV) holds obviously for any subset of *Tr*, and, by (VII), *Pr* is a subset of *Tr*. So *Pr* must be consistent. What is particularly important, as Tarski observes, is that the converse of (VII) does not generally hold: T and L(T) may well be such that some true sentences of L(T) are not provable in T, whereas all provable sentences in T are true in L(T). That is to say, T may be a sound and consistent yet incomplete theory. Gödel rigorously proved that this holds of every consistent and recursively axiomatized T embedding elementary arithmetic. And Tarski then showed, what Gödel himself anticipated, how incompleteness of such T might be demonstrated on the basis of a more powerful metatheory $MT \cup D_{L\text{-}TR}$ framed in a logically stronger ML.

From these observations it follows that while $MT \cup D_{L\text{-}TR}$ is conservative over the base theory MT (as $D_{L\text{-}TR}$ is an explicit definition), it is by no means conservative over the theory T, because it proves certain claims belonging to L(T) (in particular, the consistency of T) that T cannot prove, provided that T is consistent. Tarski's informal consistency-proof goes via the informal soundness-proof of T, as sketched above. But we can demonstrate consistency of T in a different way, using the notation that we introduced in the course of dealing with Gödel's proof of his 2nd incompleteness theorem. So we have:

a) $Con_T \quad =df. \quad \neg Pr(\underline{\langle 0 = 1 \rangle})$

b) $Sound_T \quad =df. \quad \forall x\,(Pr(x) \rightarrow Tr(x))$.

We then make two assumptions to the effect that $MT \cup D_{L\text{-}TR}$ is materially adequate and capable of proving soundness of T:

(1) $MT \cup D_{L\text{-}TR} \vdash \varphi \leftrightarrow Tr(\underline{\langle \varphi \rangle})$, for any sentence φ of L(T).

(2) $MT \cup D_{L\text{-}TR} \vdash \forall x\,(Pr(x) \rightarrow Tr(x))$.

We now unfold consequences. From (2) it follows:

(3) $MT \cup D_{L\text{-}TR} \vdash Pr(\underline{\langle 0 = 1 \rangle}) \rightarrow Tr(\underline{\langle 0 = 1 \rangle})$.

We further have

(4) $MT \cup D_{L\text{-}TR} \vdash Pr(\underline{\langle 0 = 1 \rangle}) \rightarrow 0 = 1$,

since, by (1), we can "disquote" the consequent in (3). Given that $MT \cup D_{L\text{-}TR}$ contains T, and is thus adequate to elementary arithmetic, we also have

(5) $MT \cup D_{L\text{-}TR} \vdash \neg(0 = 1)$.

Applying *modus tollens* to (4) and (5) we finally get:

(6) $MT \cup D_{L\text{-}TR} \vdash \neg Pr(\underline{\langle 0 = 1 \rangle})$ QED.

No wonder that recursive definitions of satisfaction and truth for L(T) framed in higher-order MT (recursive or explicitly set-theoretic) were preferred by Tarski to truth-axiomatizations in metatheories of the same strength as L(T), being (a) mathematically precise (interpretable in acceptable logico-mathematical system, typically set theory), and (b) metamathematically powerful ((I),...,(IX) being their consequences). This cannot be said of truth-axiomatizations in the style of MT∪TRUE for sufficiently strong languages. Tarski realized that MT∪TRUE (to forego potential confusions, I remind you, once again, that here MT is assumed to be of the same order as L(T)) does not prove the general principles (I),...,(IX). In fact MT∪TRUE *case by case* proves each instance of (I), (II), (III) or (VI), for any given sentence *x* of L(T), yet it fails to prove these general principles themselves. In case of the sentential function '$x \notin Tr \vee x \notin Tr$' the axiomatic theory can prove each its particular substitution-instance (for sentences of L(T)), but cannot prove the general law of non-contradiction '$\forall x(x \notin Tr \vee \underline{x} \notin Tr)$':[117]

> "From the intuitive standpoint the truth of all those theorems is itself already a proof of the general principle, this principle represents, so to speak, an 'infinite logical product' of those special theorems. But this does not mean at all that we can actually derive the principle of contradiction from the axioms or theorem mentioned by means of the normal modes of inference usually employed." (Tarski 1935: 257)

This phenomenon reminds us of ω-incomplete arithmetic theories that we discussed in connection with Gödel's results. In such a theory T, there is a sentential function '$P(x)$' such that, for each number *n*, '$P(\underline{n})$' is a T-theorem, yet the generalization '$\forall x(Px)$' is not a T-theorem. However, in the case of the notion of truth this is striking, as the above mentioned principle seems to be elementary.[118]

Furthermore, Tarski worried that MT∪TRUE is not categorical in the precise sense that it does not uniquely determine the extension of '*Tr*' with respect to L(T):[119]

> "...the axiom system of the theory of truth should unambiguously determine the extension of the symbol '*Tr*' which occurs in it, and in the following sense: if we introduce into the metatheory, alongside this symbol, another primitive sign, e.g. the symbol '*Tr´*'

[117] Tarski calls it „the principle of contradiction."

[118] Much the same can be said of the principles (I) and (II). And if we assume the semantic conception of truth with T-schema as governing principle of truth, (III) and (IV) should be obvious too (the universal validity of T-schema comes very close to bivalence, indeed). Of course, one may have his constructivist's worries regarding the principle of excluded middle. But, if one is a die-hard constructivist, he should have other worries about Tarski's method of truth definition, independent of this specific worry, because the method is non-finitary in character, as Tarski himself makes clear (it quantifies over infinite sequences, indeed, over sets thereof, etc.). As many commentators mentioned, Tarski - and the members of Polish logical school in general – was much more open to non-finitary methods in metamathematics than many of his contemporaries.

[119] Tarski does not target this argument directly at MT∪TRUE, but at augmented truth-axiomatizations that include also the elementary general principles that MT∪TRUE fails to prove. *Mutatis mutandis*, it applies also to MT∪TRUE.

> and set up analogous axioms for it, then the statement '*Tr = Tr´*' must be provable. But this postulate cannot be satisfied. For it is not difficult to prove that in the contrary case the concept of truth could be defined exclusively by means of the morphology of language, which would be in palpable contradiction to Th. I." (Tarski 1935: 258)

Unfortunately, it is by no means clear what "easy proof" he could have in mind when he said that it "is not difficult to prove that the postulate cannot be satisfied". But what seems reasonably clear is this: if the truth-axiomatization satisfies the postulate that he formulates, then it determines uniquely the extension of '*Tr*', hence *implicitly defines* that notion. The concept of *implicit definition* at stake here seems to be the one that goes back to Padoa (1901):

> **(Padoa's implicit definability):**
>
> A basic (primitive) notion *n* of an axiomatic theory *Th* is implicitly defined in *Th* in terms of its remaining basic notions *a, b, c*… iff there are no two interpretations of *Th* such that:
>
> (i) they make its axioms true (verify them), and
>
> (ii) they agree on what they assign to all *Th*'s basic notions *a, b, c,*… except for *n*.[120]

In modern model-theoretic *façon de parler***,** to be introduced in more detail in the next chapter, one would characterize implicit definability in a slightly different way, though retaining its semantic spirit.

> **(Model-theoretic notion of implicit definability):**
>
> Given *Th* and its language *L(Th)*, let *L(Th)** be *L(Th)*$\cup\{n\}$, where *n* is a notion not in *L(Th)*, and let *Th** be *Th*$\cup S(n)$, where $S(n)$ is a class of sentences of *L(Th)**. Then: *n* is implicitly defined in *Th** iff for every model *M* of *Th* there is exactly one way to expand *M* to the model *M** of *Th** assigning to *n* an extension in the domain of *M*.[121]

---

[120] It should be remarked that the interpretations are assumed to share the domain, but Padoa did not work with any precise notion of model.

[121] Compare the accounts in Chang & Keisler (1990), or Boolos et al (2002: 266-267). Equivalently:

> *n* is implicitly defined in *Th** iff any two models of *Th* with the same domain, and the same extensions for all the remaining basic notions of *Th*, have the same extension also for *n*.

The model-theoretic notion of implicit definability is equivalent to the following version (for *n*-place predicates):

> Let *Th*$\cup S(\varphi)$ be an expansion of *Th*, where φ is an *n*-place predicate not in *L(Th)*, and let *Th*$\cup S(\varphi^*)$ be another expansion of *Th* exactly like the former, except that φ is everywhere in $S(\varphi)$ replaced by an *n*-place predicate φ^* not in *L(Th)*. Then: φ is implicitly defined in *Th*$\cup S(\varphi)$ iff (*Th*$\cup S(\varphi)$)$\cup$(*Th*$\cup S(\varphi^*)$) $\models \forall x_1,...x_n\ (\varphi(x_1,...x_n)$

Padoa's aim was connected to the technique of demonstrating independence of an axiom (proposition) of the system of axioms (unproved propositions) on the remaining axioms via exhibiting an interpretation of the system that verifies all the remaining axioms but not the axiom to be shown independent. When this is impossible, the axiom is shown to be not independent on other axioms. The axiomatic system is then called "irreducible", if all its axioms can be shown to be mutually independent. Padoa claimed that something analogous holds also for the system of basic notions of *Th*:

> "...to prove that the system of undefined symbols is irreducible with respect to the unproved propositions it is necessary and sufficient to find, for each undefined symbol, an interpretation of the system of undefined symbols that verifies the system of unproved propositions and that continues to so if we suitably change the meaning of only the symbol considered," (Padoa 1901: 122)

where it is to be understood that the system is irreducible when *Th* does not prove any proposition equating *n* to φ, where *n* is a basic notion of *Th* and φ a formula of *Th* that contains only the remaining notions of *Th* plus logical constants. Since the derivation of such a proposition in *Th* amounts to an explicit definition of *n* in *Th*, what Padoa in effect claims (without proof) is this:

> **(Padoa's conjecture):**
>
> (A) it is not the case that an explicit definition of *n* in *Th* in terms of its remaining notions is derivable in *Th* iff it is not the case that *n* is implicitly defined in *Th*,

or, equivalently:

> (A*) an explicit definition of *n* in *Th* in terms of its remaining notions is derivable in *Th* iff *n* is implicitly definable in *Th.*

Let us now return to Tarski's problems. Being a pioneer in definability theory, he was of course thoroughly familiar with Padoa's work and it is plausible to suppose that what he had in mind when talking about the proof of non-categoricity of truth-axiomatizations was that categoricity, construed as implicit definability, entails explicit definability (at least in a range of standard logical systems). Still, this is only one part of the announced „easy proof" to the effect that the truth-axiomatization is not categorical. The crux of the matter is precisely to show that if the truth-axiomatization MT∪TRUE defines implicitly

---

$\leftrightarrow \varphi^*(x_1,...x_n)$).

And there also the syntactic version:

> Let $Th \cup S(\varphi)$ be an expansion of *Th*, where φ is an *n*-place predicate not in *L(Th)*, and let $Th \cup S(\varphi^*)$ be another expansion *of Th* exactly like the former, except that φ is everywhere in $S(\varphi)$ replaced by an *n*-place predicate φ^* not in *L(Th).* Then: φ is implicitly defined in $Th \cup S(\varphi)$ iff $(Th \cup S(\varphi)) \cup (Th \cup S(\varphi^*)) \vdash \forall x_1,...x_n\ (\varphi(x_1,...x_n) \leftrightarrow \varphi^*(x_1,...x_n))$.

'*Tr*' (in Tarski's sense), then (a) an explicit definition of '*Tr*' is already available on the basis of MT∪TRUE, hence (b) on the basis of MT (viz. *morphology*) alone. Tarski suggested that this conditional is easy to prove, but the truth is that for 1st order languages and theories this requires Beth's fundamental theorem, whose proof was published only in 1953.

The famous theorem states (for simplicity, I restrict it to 1-place predicates only, but it can be generalized to cover *n*-place predicates, *n*-place function-symbols and terms, qua 0-place function-symbols):

(Beth's definability theorem):

Let *Th*, *L(Th)*, *Th**, and *L(Th)** be as in the model-theoretic explanation of implicit definability. Then the notion *n* is implicitly defined in the 1st order *Th** (in semantic sense) if and only if *Th** explicitly defines *n* so that there is a 1-place formula φ such that

(i) φ contains only *L*(*Th*)-notions, and
(ii) $Th^* \vdash \forall x(n(x) \leftrightarrow \varphi(x))$.[122]

But, of course, neither Padoa nor Tarski proved this result. What Padoa established was at best the left-to-right direction of (A*) or the right-to-left direction of (A).[123] Tarski's claim is puzzling, given that he did nothing to outline how the "easy proof" proceeds.

However, it might be that Tarski had another proof in mind. It is to be noted that he wrote well known articles whose intent was to justify Padoa's method of establishing definitional independence of some notion on others for the framework of simple type-theory. His basic result for such a framework was stated in the (1934-5) article called 'Some Methodological Investigations on the Definability of Concepts' (I simplify Tarski's notational machinery):[124]

(Tarski's definability theorem):

Let *L*(*Th*) be a language of the simple (impredicative) type theory and *Th* a finite axiomatized theory. Then: an explicit definition of *n* in *Th* in terms of its remaining notions is derivable in *Th* iff every two interpretations of *Th* with the same domain that agree on all its basic notions except *n* agree also on *n*.

---

[122] For *n*-place predicates or function-symbols the formulation would have to be accordingly modified. In modern textbooks it is common to introduce the semantic version of Beth's theorem, in which the provability turnstile is replaced by the semantic turnstile:

> The notion *n* is implicitly defined in the 1st order *Th** (in semantic sense) if and only if *Th** explicitly defines *n* so that there is a 1-place formula φ such that
>
> (i) φ contains only *L*(*Th*)-notions, and
> (ii) $Th^* \models \forall x(n(x)$ iff $\varphi(x))$.

Beth's theorem applies to 1st-order theories or to higher-order theories, provided that the higher-order variables can be construed as 1st order variables of a different sort.

[123] Cf. Hodges (2008).

[124] Here I am indebted to Feferman (2008b).

More precisely, what he demonstrated was rather the following syntactic version:

> Let $L(Th)$ and Th be as before, and let Th^* be exactly like Th except that n is everywhere replaced by n^*. Then an explicit definition of n in Th is derivable in Th iff $Th^* \cup Th \vdash \forall x(n(x)$ iff $n^*(x))$.

Tarski sort of vindicated Padoa's method for a range of higher-order deductive systems by demonstrating that Padoa's conjectures hold for them. This, however, is not quite right, as Hodges persuasively shows in his rich paper (2008), containing a valuable discussion of the relation between Padoa's method and Tarski's justification of it. He argues, on the basis of good textual evidence, that Tarski did not in fact vindicate the right-to-left direction of (A) above (or, for that matter, the left-to-right direction of (A*)), but he showed how to translate Padoa's informal semantic method into a purely formal-syntactic method within a deductive theory. Indeed, what Tarski did in (1935) is a very careful attempt to avoid all semantic (model-theoretic) ideas from the picture:[125]

> "In short, Tarski is not claiming to make Padoa's original proposal any more plausible. He is claiming to transfer as much as possible of Padoa's method into the form of calculations within a deductive theory. The effect of Tarski's analysis of Padoa's method is to eliminate the model theory." (Hodges 2008: 109).

What is important for our purposes here is that Tarski thought that his definability theorem helps to prove non-categoricity of truth-axiomatizations. In what follows, I dare to reconstruct the "easy proof" he could have in mind (MT corresponds to *Th* and MT∪TRUE to *Th**, as these have been introduced in the formulations of Padoa's and Beth's theorem):

> Suppose for a *reductio* that **(1)** MT∪TRUE is categorical so that it implicitly defines '*Tr*' for L(T) in the following sense: if TRUE* is exactly like TRUE except that '*Tr*' is everywhere replaced by '*Tr**', then MT∪TRUE∪TRUE* $\vdash \forall x(Tr(x) \leftrightarrow Tr^*(x))$. But **(2)** if '*Tr*' is implicitly defined in MT∪TRUE, then an explicit definition of '*Tr*' is derivable in MT∪TRUE in terms of the remaining notions of MT∪TRUE (by Tarski's definability theorem). But **(3)** among those remaining notions of MT∪TRUE there are only notions belonging already to ML, and there is therefore a formula ψ containing only the notions of ML, which is provably coextensive in MT∪TRUE with '*Tr*', hence explicitly defines '*Tr*' in MT∪TRUE. **(4)** If so, we have a formula of MT - namely ψ - that defines truth for L(T) in the sense of having for its extension the set of L(T)-truths. But **(5)** since MT is sufficiently strong to satisfy DIAGONAL LEMMA, but only as logically strong as L(T), MT cannot possibly define the notion of truth for L(T) (by Tarski's indefinability of truth Theorem I). Consequently:

[125] See also Coffa's (1991) useful discussion of Tarski's views on definability.

(6) MT∪TRUE is not categorical, and hence it does not implicitly define '*Tr*'. QED

Tarski further considers possible completions of the deductively weak MT∪TRUE so that it subsume somehow the principles such as (I),...,(IX). He notes that one may first want to add the principles to MT∪TRUE as further axioms, but he rejects this alternative out of hand by pointing out, again, that such extensions of the axiomatic system are unprincipled - having "accidental character" (albeit relatively consistent - by extension of Theorem III). However, Tarski's principal worry is based on the argument we have just reconstructed for MT∪TRUE: even the truth-axiomatization extended by the general principles alone is not categorical, hence does not capture the content of truth (though it is consistent, just as MT∪TRUE - Theorem II applying to it as well).

What is arguably the most interesting strategy that Tarski considers is one that does not propose to add to the axioms of MT∪TRUE but rather to add to its inference rules. What was so unsatisfactory about MT∪TRUE? Arguably this: although it case-by-case proves all instances of certain general principles of truth, it does not prove the general principles themselves. We noted in this respect the analogy with ω-incomplete arithmetic theories. Now, along with Hilbert, Gödel and Carnap was one of the first logicians who seriously discussed the so-called ω-rule, which would allow us to infer '$\forall x P(x)$', if we proved '$P(\underline{n})$', for each number n.[126] Tarski mentions that certain elementary systems of arithmetic can be "completized", if we expand them by the ω-rule, in which case a purely structural (or syntactic) truth definition for L(T) becomes possible as *the smallest set containing elementary true sentences without variables or quantifiers and closed under the ω-rule*.[127] However, in spite of the fact that the rule seems intuitively valid (its validity with respect to T can in fact be proved via a truth definition for L(T) in a higher order MT), it has infinitary character, because its application presupposes that an infinite number of premises has been proved in T. This is worrying. Indeed, how we, humans with finite capacities, can reason with infinitely many premises? In the well-known article on logical consequence (1936) Tarski considers an interesting proposal to lay down a finitary (structural) version of the rule for (arithmetical) T:

---

[126] The rule used to be called Tarski-rule or Carnap-rule. Indeed, Carnap proves in (1934) that a system of elementary arithmetic augmented by the rule is complete. However, already in 1927 Tarski lectured on ω-incomplete and ω-inconsistent theories - although the labels are not due to him but due to Gödel - where he gave the example of both types of theories. The problem is discussed in detail in Tarski (1933b). With Hilbert, the situation is more complicated. Probably with the intent to overcome Gödel's theorems, he attempted to use a semi-finitary (*sic!*) version of the rule to show that elementary arithmetic augmented with his rule is complete. His rule stated: when '$P(x)$' is a quantifier free formula for which we can prove by finitary means that '$P(\underline{n})$' holds for each n, then we can use '$\forall x P(x)$' as a new premise in all further proofs. Unlike the ω-rule, Hilbert's rule puts restriction on what formulas can replace '$P(x)$' and on the means by which its instances are to be proved. The problem with this idea is that the rule is informal and imprecise to the extent it itself appeals to the notion o 'finitary proof' so that it is not clear if the resulting system is *bona fide* 'formal' in Hilbert's own preferred sense. On the other hand, once the rule is properly formalized - a finitary version of the ω-rule - Gödel's theorems apply to the resulting system.

[127] Tarski (1935: 261). This definition is close to one proposal of Carnap (1934), who defined in that style logical (analytical) truth - in case of arithmetical L(T) coinciding with arithmetical truth). For Carnap's approach see the next section.

> (R) If, for each *n*, '$P(\underline{n})$' is provable via the previous set of rules SR, then '$\forall x P(x)$' is to be regarded as proved.

One can add a number of such rules R*, R** ..., which are increasingly stronger, as R* presupposes SR+R, and so on. Since such rules are structural (syntactic) in character, they can expressed in T via the method arithmetization. For this very reason, though, the strategy is problematic, as, by Gödel's 1st theorem, by adding such finitary rules one cannot complete an incomplete theory T.

What is important for our discussion is that Tarski formulates *the rule of infinite induction* (RI) for metatheoretic predicates (syntactic or semantic) of expressions. In case of the predicate '$Tr(x) \vee \neg Tr(x)$' the rule licenses the inference to '$\forall x(Tr(x) \vee \neg Tr(x))$', provided that all substitution instances of that predicate (for L(T)-sentences) are provable in MT∪TRUE (as we know they are). We could thus hope to overcome the serious deductive weakness of MT∪TRUE, since, as Tarski points out, the resulting theory - MT∪TRUE plus RI – is very powerful and categorical. Unfortunately, however, now the problem of proving consistency becomes urgent:

> "Under these circumstances the question whether the theory erected on these foundations contains no inner contradiction acquires a special importance. Unfortunately this question cannot be decided at present. Th. I retains its validity: in spite of strengthening of the foundations of the metatheory the theory of truth cannot be constructed as a part of the morphology of language. On the other hand for the present we cannot prove Th.III for the enlarged metalanguage, The premise which has played the most essential part in the original proof, i.e. the reduction of the consistency of the infinite axioms system to the consistency of every finite part of this system, now completely loses its validity – as is easily seen – on account of the content of the newly adopted rule. The possibility that the question cannot be decided in any direction is not excluded ..." (Tarski 1935: 261)

So it is at this juncture, and not earlier, where the problem of consistency arises for truth-axiomatizations. According to Tarski (1933), with truth-axiomatizations we face the following dilemma: either they are assuredly consistent (indeed conservative over MT) but then they are too weak and/or non-categorical to be metatheoretically satisfying theories of truth (the case of MT∪TRUE), or they are quite powerful and categorical so that they could be plausible and metatheoretically useful theories of truth, but then their consistency remains an open problem. However, in the Postscript the question of consistency of truth-axiomatizations is no longer felt to be a problem: if, for any formalized language, we can construct its Tarskian truth-definition in a higher-order (possibly transfinite) metatheory, consistency of truth-axiomatizations of all types that we have so far considered is assured relative to this more powerful metatheory, because such truth-axiomatization become interpretable in the metatheory:

> "In view of the new formulation of thesis A the former Thesis C loses its importance. It possesses a certain value only when the

> investigations are carried out in a metalanguage which has the same order as the language studied and when, having abandoned the construction of a definition of truth, the attempt is made to build up the theory of truth by the axiomatic method. It is easy to see that the theory of truth build up in this way cannot contain an inner contradiction, provided there is freedom from contradiction in the metalanguage of higher order on the basis of which an adequate definition of truth can be set up and in which those theorems which are adopted in the theory of truth as axioms can be derived." (Tarski 1935: 273)

However, once we can construct explicit truth definitions, what then is the value of truth-axiomatizations?

4.4.1 Definitions, axiomatizations and the problem of "reduction"

Another advantage of explicit truth definitions over truth-axiomatizations hangs in closely with the foregoing aspect. Axiomatizations use the primitive notion of truth, while truth-definitions explain truth in terms of other notions, of which it can be hoped that they are unproblematic or, at the very least, less problematic. Clearly, if one's aim is in part to rehabilitate or clarify a notion that is deemed problematic in some respects, it is a dubious strategy to use it as a primitive notion and lay down its properties in axiomatic style. Tarski glossed the situation by saying that this is an objectionable procedure from the psychological perspective. However, given that we have seen that Tarski himself came to concede that consistency of truth-axiomatizations in the style of MT∪TRUE or even of MT∪TRUE augmented with the rule of infinite induction is not the problem, what residual psychological blocks could we have with respect to the primitive notion truth axiomatized in consistent and materially adequate manner? Indeed, we have remarked that there was no urgent logical need to prove consistency of informal semantic notions as they were used before Tarski, by his fellow logicians. The technical strategies were already known how to block semantic paradoxes, and they were based on essentially the same ideas that Tarski later worked out rigorously (i.e. a sort of type-restrictions, as in Russell's ramified theory of types, Gödel's independent observation that truth of arithmetical object language is not expressible in it but in its metalanguage).

My own hypothesis is that three different reasons that could motivate Tarski's sceptical attitude to truth-axiomatizations, except the one to the effect that they are either deductively too weak (the case of MT∪TRUE) or transcend the realm of well established classic logic (the case of MT∪TRUE + RI). Firstly, Tarski could have thought that even though truth-axiomatization in the style of MT∪TRUE or MT∪TRUE+RI can be shown consistent, this assurance is parasitic on the explicit truth-definitions given in higher-order metatheories. Therefore, the later should enjoy a methodological priority. Secondly, he could have thought that since axiomatizations use the primitive metatheoretical notion of truth (satisfaction, denotation, or definability), *mathematical truth* is not itself a mathematical but, strictly speaking, meta-mathematical notion. Feferman

pointed out that[128] Tarski's seemed to have a long-life feeling that the fruitful semantic methods of metamathematics (based on the method of absolute truth definition or extended to model theory) need to be formulated in the form acceptable to working mathematicians, whose attitude towards them was in his opinion one of distrust.

Here is a representative passage from the well known work, in which Tarski set out to define the notion of *definable set of real numbers*:

> "Mathematicians, in general, do not like to deal with the notion of definability; their attitude toward this notion is one of distrust and reserve. The reasons for this aversion are quite clear and understandable. To begin with, the meaning of the term 'definable' is not unambiguous: whether a given notion is definable depends on the deductive system in which it is studied ... It is thus possible to use the notion of definability only in a relative sense. This fact has often been neglected in mathematical considerations and has been the source of numerous contradictions, of which the classical example is furnished by the well-known antinomy of Richard. The distrust of mathematicians towards the notion in question is reinforced by the current opinion that this notion is outside the proper limits of mathematics altogether. The problems of making its meaning more precise, of removing the confusions and misunderstandings connected with it, and of establishing its fundamental properties belong to another branch of science—metamathematics." (Tarski 1931: 110)

Yes, the problems identified by Tarski belong by their nature to meta-mathematics rather than to mathematics, but they are not therefore entirely beyond the scope of mathematical methods, because under certain conditions metamathematical definitions can be transformed into purely mathematical: when properly relativized to formalized languages and constructed on the basis of a set-theoretical metalanguage (formalized or semi-formalized). Recall the following passage:

> "[…] meta-mathematics is itself a deductive discipline and hence, from a certain point of view, a part of mathematics; and it is well known that – due to the formal character of deductive method – the results obtained in one deductive discipline can be automatically extended to any other discipline in which the given one finds an interpretation […]." (Tarski 1944: 369).

In his (1931) article on semantic definability, the notion of *definable set of reals* (relative to the 1st-order fragment of simple type theory based on the universe of reals) is approached via the more general notion of *definable set of finite sequences of reals* (definable *n*-dimensional relations between reals). He gives its metamathematical definition in terms of *satisfaction of a sentential function by finite sequences of reals* (whose recursive metamathematical definition is only hinted there but is fully spelled out in CTFL) in roughly the

[128] Feferman (2008).

following style:

> A set *S* of *n*-termed sequences of reals is definable relative to the language L in question if and only if there is an *n*-place sentential function *f* of L that is satisfied by all and only the members of *S*.[129]

Then Tarski shows how to construct a purely mathematical definition in terms of sets of finite sequences, without appealing to satisfaction or any other semantic idea belonging to the realm of metamathematics. The mathematical definition of definable sets of finite sequences of reals closely mimics the metamathematical one based on satisfaction, since the family of definable sets of finite sequences of reals is the one that contains certain primitive sets of finite sequences of reals, each corresponding to one atomic sentential function of L, containing exactly those sequences that satisfy that function, and is closed under a couple of Boolean-type operations corresponding to the logical operations by means of which complex sentential functions of L are formed (negation, conjunction, disjunction, universal and existential quantification). For Tarski, this was paradigm that can be transposed to definitions of semantic notions of truth, satisfaction or denotation, as we have detailed them in the chapter on formal truth definition. Indeed, he suggested that from the formal perspective, truth is but a special case of satisfaction, just like definability. The basis is in both cases a metamathematical recursive definition of satisfaction of an *n*-place formula *f* of L by ordered *n*-tuples (finite or infinite sequences) of objects from the universe U of L. Now, the notion of sentential truth is defined as a limiting case of satisfaction of a 0-place sentential function of L by all/some infinite sequences of objects (alternatively, in the setting of the 1931 article, by 0-termed sequence). But all this can be turned to a purely set-theoretical definition, as we have just seen on the example of the notion of *definability*.

We have reviewed some reasons why Tarski set out to provide conceptual analyses of truth or satisfaction in mathematical terms of set theory, although he originally conceived of set theory as formalized within the simple type-theory, consistently with then prevailing practice in mathematical logic (it was only somewhat later, in the second half of the 1930s, when he came definitely to prefer Zermelian set theory, "Skolemized" into its 1st order form, as a foundational setting of mathematics). However, it is doubtful whether a full mathematization of metamathematics (including semantics) was really called for, at least for the reasons that we have attributed to Tarski. Feferman perceptively remarks that in this respect Tarski showed somewhat paranoid symptoms. By all sings those mathematicians who showed some serious interest in metamathematics including semantics (and later model theory) did not seem to be worried in the least about metamathematical definitions of its basic notions and formulations of its basic theorems in terms of them, once efficient precautions were taken against paradoxes.[130] Nowadays, it is customary to provide metamathematical recursive definitions of semantics, which, albeit expressed within the set-theoretical language, do not banish semantic notions in favour of purely mathematical notions (even if Tarski showed this to be possible

[129] Semantical definability of an *n*-dimensional relation over U by an *n*-place formula *f* of L is not to be confused with syntactical or formal definability - also examined thoroughly by Tarski - of an expression of L in terms of other expressions of L).
[130] Feferman (2008: 80).

in principle). In CTFL Tarski himself remarks that recursive definitions of sentential function, satisfaction and other notions bring out their content in a better way than do their explicit counterparts, which, by the way, he introduces only in the footnotes, being content to say that the natural recursive versions can be converted into explicit versions via Dedekind-Frege procedure. This suggests that, as regards conceptual analysis, translation of recursive definitions into explicitly set-theoretical definitions does not seem to contribute anything essential not present in the former. Still, set theory remains a powerful conceptual tool employed in formulating recursive definitions. The point is that all substance there is to conceptual analyses of semantic notions in Tarski's own framework is contained in recursive metamathematical definitions. To think that purely set-theoretical definitions in the style

> ... belongs to the smallest set (intersection of all sets) satisfying such and such closure conditions ...

offer substantial conceptual analyses is an illusion. Tarski was at least dimly aware of that, in spite of his well-known tendencies to define everything in logico- mathematical terms that can be so defined. Where, on the other hand, the method of conversion into set theory is crucial is in providing kind of mathematical assurance that metatheory enriched by semantics is consistent: being eliminable in principle in favour of set-theoretical notions, recursively defined semantic notions do not threaten to bring in any inconsistency into the metatheory. And, for this reason, they had a methodological priority over truth-axiomatizations, however powerful the later may be.

It remains to mention the last reason why Tarski tended to prefer explicit truth definitions within the general set theory. It was said that merely to axiomatize some problematic notion is not a *prima facie* attractive procedure. Now, Tarski wanted, among other things, to rehabilitate the notion of truth, which subject was the source of all sorts of confused debates in the traditional philosophy, and for this reason it was considered a metaphysically loaded *idea non grata* by philosophers critical of the traditional metaphysics.

> "Apart from the problem of consistency, a method of constructing a theory does not seem to be very natural from the psychological point of view if in this method the role of primitive concepts – thus of concepts whose meaning should appear evident – is played by the concepts which have led to various misunderstanding in the past. Finally, should this method prove to be the only possible one and not be regarded as merely a transitory stage, it would arouse certain doubts from a general philosophical point of view. It seems to me that it would then be difficult to bring this method into harmony with the postulates of the unity of science and of physicalism (since the concepts of semantics would be neither logical nor physical concepts)." (Tarski 1936b: 406)

Such complaints were often voiced by logical positivists, most prominently by Neurath who had on that matter an interesting correspondence with Tarski as well as with Carnap (after he had adopted Tarskian perspective). But it is difficult to decide what significance to attach to Tarski's claims here, if

only because nowhere in his published writings he repeats the desideratum that semantics should be in harmony with the physicalist basis of (general) logical plus physical notions. We have said at the outset that it was in the nature of Tarski's inquiry to use semantic notions, only if they can be "reduced" to acceptable non-semantic notions of the metalanguage - viz. the case of defining truth via satisfaction, which, in turn, is defined explicitly in set theoretical terms. Except of hoping to show in that manner that semantic notions so introduced do not threaten to bring in any inconsistency into an already consistent metatheory, as they are eliminable in principle (arguably the main motivation behind the procedure), and attracting the attention of mathematicians (a minor motivation), Tarski hoped to show that they are respectable notions, provided that the non-semantic primitive (or defined) notions of the metatheory are respectable, in terms of which they are introduced into metatheory. Better still, his idea seemed to be this:

> Once you accept the object-language and the extra logico-mathematical and morphological apparatus used in the metatheory, then you should not have any objection against accepting semantic notions introduced into the metatheory via explicit definitions solely in terms of the expressions of the object-language plus the extra-apparatus of the metatheory, as the notions introduced in this way are always eliminable in favour of the latter (and, in this sense, are reducible to them).

Since the metatheory was assumed to be a system sufficient to develop a general set theory (or a reasonable amount of it), it can be said that Tarski *reduced*, in the sense of *having explicitly defined*, semantic notions to set-theoretical notions *via* interpreting the semantic theory for the object-language in the system of set theory framed in the metalanguage.[131] Had he focused instead on the 1st order arithmetic (e.g. on **PA**), he could have "reduced" semantic notions to the notions of 2nd-order arithmetic. It may well be that this kind of *reduction via interpretation* is not what the semantic sceptics of Neurath's calibre would have expected one to offer in order to rehabilitate semantic (or intentional) ideas in their eyes. If so, Carnap was an important exception, presumably because Tarski's method of establishing scientific semantics on the basis of morphology in a higher order metatheory was congenial to his own quasi-syntactic approach developed in *Logische Syntax* (1934).[132] At any event, there was likely no definite consensus in the Vienna circle on what a successful reduction of a notion to a class of other notions amounts to (even today, I suspect, there is no agreement on this among the contemporary philosophers of science). It may well be that, for some period of time, the neutral minimum required for reduction was to translate the problematic (semantic, intentional) target-idiom (vocabulary, language) into a non-problematic base-idiom (vocabulary, language) that is extensional and contains only empirically-scientifically respectable notions (phenomenalistic or physicalistic) and logico-mathematical notions, where the translation is to be extensionally correct. That is, at minimum, the reduction can be achieved via an extensionally correct

[131] Viz. the passage quoted at p. 86, where Tarski says that a syntactico-semantic metatheory is a deductive theory that can find an interpretation in another deductive theory.

[132] Though Neurath had a high opinion of Carnap's approach in (1934) and seemed to approve of it, as the correspondence between them informs us.

definition of some notion in terms of empirically-scientifically respectable notions in tandem with logico-mathematical notions (relative to a privileged body of sentential contexts). In the case of narrowly semantic notions (truth, satisfaction, denotation) that are properly restricted, Tarski showed how to provide precisely such "reductions".

One may worry that the nature of semantic or intentional notions (properties) cannot be find in set theory, however useful conceptual tools it provides. In light of this, we can understand the numerous complaints to be reviewed in Chapter VII to the effect that Tarski's method of truth definition for formalized languages did not in the least put to silence the philosophical worries concerning the place of semantic properties in the natural order:

> How are semantic properties of expressions individuated?
> Can they be scientifically explained in terms of natural properties?
> Are semantic properties somehow determined by natural ones?

or the questions concerning our epistemic standing with respect to such properties:

> How do we identify semantic properties of expressions?
> How do we know what semantic properties expressions possess?

In the eyes of die-hard semantic sceptics a mere elimination of semantics in favour of set-theoretical ideology was likely an ingenious technical trick that sweeps all their foundational questions under the carpet. Imagine someone seriously pondering the question of the place of numbers in the natural order or of our epistemic standing with respect to them. If the questions make sense, I fear that telling the person who so ask in response the Frege-Russell definition of numbers as equivalence classes of equinumerous classes (or any extensional equivalent such as von Neumann's definition[133]) would not help him much. For the analogous question arises now with respect to classes: what is the place of classes (of classes) in the natural order and what is our epistemic standing with respect to them? While Carnap who was himself a proponent of a version of logicism, could be impressed by a definition of truth *etc.* in terms of logic, set theory and syntax as parallel to the successful Frege-Russell logical definition of natural numbers, this view need not have been embraced by semantic sceptics.

For this reason, Tarski could not offer a satisfying definition of truth and semantic notions to those die-hard semantic sceptics, who, like Neurath, were troubled by their ontological nature or epistemological status. There is an exegetical tradition that works with the assumption - based on the single quotation from his (1936b) article – that Tarski wanted to reduce semantic notion to the physicalist basis in roughly the way demanded by logical positivists propounding the idea of the unified science framed in something like the general language of physics. Some commentators claim that he blatantly failed,[134] others argue that he succeeded, because the original project of

[133] Von Neumann, on the other hand, defines each number *n* directly as the set of its predecessors that is, as identical with the set {0, 1, …, *n* – 1}. 0, having no predecessors, is identical with ∅; 1 = {∅}, 2 = {∅,{∅}}, 3 ={∅,{∅},{∅,{∅}}}, and so on.

[134] Field (1972), or McDowell (1980), who accepts Field's exposition of physicalism as well as his critique of Tarski's ambitions, but defends himself a different (Davidsonian) perspective upon the question of how physical (behavioural) and semantic facts are connected.

physicalism was not what the modern critics of Tarski have taken it to be.[135] Whatever we may think about the legitimacy of such questions, it seems clear that Tarski was simply unconcerned with them. He was content to show that a target notion can be defined within a preferred language of the unified science, which is allowed to include abstract mathematics in the form of logic plus set theory, in an extensionally correct manner. Once the obligatory formal requirements are satisfied by the definition and its extensional adequacy is secured, there is no residual worry about the definition, given that it brings scientific (here: meta-theoretical) fruits.

My own view of the matter is different. Admittedly, Tarski was a physicalist of a sort. The evidence for this is that he used to describe himself as a nominalist, extensionalist or even finitist (!) believing in the world of spatio-temporal particulars and properties and relations thereof; moreover, he was seriously engaged in the debates with Carnap and Quine in the late 1930s and early 1940s, which revolved around the basic question of what a genuine language of science could look like, and he suggested several nominalist proposals.[136] It was noted by several commentators that there is a tension between Tarski's "private" sympathies to nominalism and physicalism and his official metamathematical research, in which he made a heavy use of set theory and transfinite methods in general. Indeed, it was willingness to make use of any fruitful mathematical methods available, including the transfinite methods, which in Tarski's opinion characterised the approach of Lvov-Warsaw school and distinguished it from other contemporary schools such as Göttingen school pursuing Hilbert's program (aimed to justify precisely such "ideal" methods), or from various constructivist approaches. On the other hand, that Tarski was a physicalist "in private" does not mean that he was "officially" concerned to give physicalist definitions of semantic notions. His main interest was in metamathematics and the solitary remark regarding the compatibility of his method of truth definition with the idea of unified science and physicalism was likely made because he wanted to please his positivist auditorium at the Paris Congress 1930, which, as he correctly anticipated, was going to be quite sceptical with respect to his views. It is no accident, in my opinion, that Tarski made this claim in the context of his talk at the congress in Paris, after having been repeatedly encouraged by Neurath to give a talk on semantics that would be consistent with the general empirical viewpoint and unfriendly to metaphysical speculation.[137]

---

[135] Cf. Kirkham (1993).

[136] See the wealth of material contained in Frost-Arnold (2004).

[137] I can only recommend the fascinating historical material gathered by Mancosu (2008), documenting - on the basis of Neurath's correspondence with Carnap, Tarski and Kokoszýnska - Neurath's continual fears that Tarskian semantics propounded by the three thinkers contains perhaps a useful mathematics but philosophically it is dangerous, threatening to resurrect „Aristotelian metaphysics" or even „scholasticism", evidently interpreting Aristotle's truth-dictum as the mother of all correspondence theories, which in turn smell by metaphysics. Mancosu (2008) is a very good place to look at when one wants to get a better grasp of Neurath's views on truth. It has often been claimed that his conception of truth is coherentist (Schlick explicitly attributed this position to him), but Neurath denied that classification (which according to him was just as well a metaphysical position – at least in its post-Hegelian versions then widespread in England), arguing instead for kind of pragmatist-verificationist outlook with strong holistic elements, which reminds us of some of Quine's remarks on truth as a property of a scientific theory. See also Frost-Arnold (2006) for a detailed defence of the view that I have

4.4 Carnap's contributions to the semantic conception of truth

With the exception of Tarski himself, nobody else contributed more to the semantic conception of truth than Rudolf Carnap. Next to Gödel and Tarski, he was arguably the most prolific theorist concerned with the problem of the syntax-semantics interface in metamathematics and philosophy in general, and his work in this area is also related in interesting ways both to Gödel's and Tarski's research. It is not my aim to explain the details of his complex approach to metamathematics, or its development over years, starting from his universalist conception of logic influenced by Russell and Frege, continued by his quasi-syntacticist metamathematical period inspired by Hilbert, Gödel and Tarski, superseded finally around 1935 by the semantic approach along Tarskian lines (departing from Tarski's approach in certain important respects that I shall mention in due course). It is well known, at least to the scholars in the early history of analytical philosophy, that Carnap was not only the first to isolate the diagonal (fixed point) lemma in *Logische Syntax* (1934),[138] but he also anticipated there Tarski's indefinability of truth theorem – indeed, his Theorem [60c] is a version of Tarski's result for the notion of logical truth (L-truth or analyticity) with respect to a formalized language containing a sufficient amount of arithmetic (his Language II) for Gödel's method of arithmetization of its syntax and the diagonal lemma to apply to it (to be provable in its built-in deductive system embedding elementary arithmetic). In fact, Carnap explicitly states and proves many important metatheoretical observations regarding his formalized Language II, sufficiently strong to develop arithmetic (so that the DIAGONAL LEMMA is satisfied for it).[139] From our perspective, the following results are particularly interesting:

- For any consistent formal system framed in such a language that contains a sufficient amount of arithmetic there are undecidable sentences formulated within that language (Gödel's 1st incompleteness theorem). Unprovability of consistency for such a formal system within that system itself (Gödel's 2nd incompleteness theorem).

- It is impossible to define L-truth for such a language within that language, but only in a higher order language L* [As in logico-mathematical language logical truth coincides with plain truth, what Carnap established was a restricted version of Tarski's indefinability of truth theorem. As for mathematical truth, Gödel established that result independently of Tarski. Carnap arrived at it independently, as Gödel did not mention the result in print before 1934, and he was not yet familiar with CTFL]

---

urgedhere.

[138] The German version appeared already in 1934, but all page references are to the English translation of a revised and complemented version published in (1937) appeared a year later. In particular, Carnap's Gödelian argument based on DIAGONAL LEMMA was not included in the original German version but appeared in two separate papers. I owe this observation to Procházka (2006).

[139] For Carnap's statement of the lemma see (1934: 126).

- A hierarchy of languages of higher and higher order can be indefinitely extended into the transfinite: in the language L* at the level *n* it is possible (i) to define L-truth for the language L at the *n*-1th level, and (ii) to decide undecidable sentences of the formal system T framed in L, on the basis of a more powerful formal system T* framed in L*. [This was stated also by Gödel; but before 1934 only with respect to the problem of decidability; in application to the problem of L-truth definition, Carnap's approach was original, and more general than Tarski's approach in (1933), anticipating the more general strategy adopted in the 1935-Postscript]

Many instructive discussions are available in the existing literature, and I can therefore confine myself to discussing those aspects of Carnap's work that are most intimately related to our main topic – the semantic conception of truth.[140] In his "syntactic" *opus magnum* (1934) Carnap set out to work out a viable account of analytical (logico-mathematical) truth and consequence. In fact, in different stages of his philosophical development he offered different analyses and solution to the problem. Despite rejecting Wittgenstein's view to the effect that logical truth, as opposed to factual truth, is a feature that should always be discerned from the sentence's design, or the thesis that there is no standpoint from which to approach the language metatheoretically,[141] Carnap retained Wittgensteinian conception of logical truth as non-factual truth – truth by rules of language alone, and completely independent of the matters of fact:

> "...I was guided, on the one hand, by Leibniz' view that a necessary truth is one which holds in all possible worlds, and on the other hand, by Wittgenstein's view that a logical truth or tautology is characterized by holding for all possible distributions of truth-values. Therefore, the various forms of my definition of logical truth are based either on the definition of logically possible states or on the definition of sentences describing those states (state-descriptions)." (Carnap 1962: 62)

To attain the aim of characterizing logico-mathematical truth and consequence in the aftermath of Gödel's discovery of incompleteness phenomena which discredited any attempt to reconstruct logico-mathematical truth by narrowly syntactic methods (usually: reduction of logical truth within a formal system to provability within that system), in *Logical Syntax* Carnap embarked on the original project of constructing formalized languages and studying their properties in an extended formal-syntactic style on the basis of the syntactic metalanguage. Such extended formal-syntactic methods were supposed to define both logical truth and consequence in a satisfactory manner. In fact, the means that Carnap allowed in the syntactic metalanguage go far beyond the means that are available in the object-language studied within it – be it our

[140] See Coffa (1991), Procházka (2006, 2010) or several articles contained in Wagner (2009), especially de Rouilhan (2009).
[141] The view was based on the dichotomy between what can be said and what can only be shown and the ensuing thesis of ineffability of syntax and semantics.

familiar ω-rule or a higher-order apparatus of variables and quantification – as one would expect in the light of Gödel's theorems.[142] The approach proved fruitful in several respects but it was not without its problems. Eventually, Tarski's contemporary metatheoretical work persuaded Carnap around 1935 that it is more plausible and natural to bring semantic considerations into the picture, which he attempted in the *Introduction to Semantics* (1942), partly based on his earlier Encyclopaedia entry on the *Foundations of Mathematics* (1939).[143] It is fair to say, though, that neither in the syntactic nor in the semantic *opus* did Carnap succeed in giving a fully satisfactory account of logical truth.

What is particularly important to keep in mind is that the shift from the project of *logical syntax* (in the extended sense) to the project of *logical semantics* (inspired by Tarski) that Carnap pursued until the end of his career was not so dramatic as one might at first think, given the contrast between *syntax* and *semantics* that our modern ears tend to associate with this distinction. As several perceptive commentators have recently noted,[144] already in *Logical Syntax* Carnap was quite close to giving the truth-definition in Tarski-style for a formalized language under its intended interpretation. Purified and simplified to a significant extent indeed, his definition of analyticity for a formalized language embedding arithmetic runs as follows (Carnap's Language II):[145]

---

[142] For detailed accounts see Procházka (2006, 2010) or de Rouilhan (2009). The role of the ω-rule in Carnap's conception of so-called indefinite *c-concepts* (as opposed to definite or recursive *d-concepts*) is an extremely interesting chapter of modern metamathematics. We have seen that Tarski was also interest in the rule, calling it the rule of infinite induction. He claimed its intuitive validity but also that it transcends the realm of classic logic in that it requires us to consider (*per impossible* ?) an infinite number of premises. As a deductive rule, that is, it is non-standard and its applicability is questionable. Tarski claimed that MT + TRUE + the ω-rule yields a categorical system. Carnap argues in *Logical Syntax* that while PA is incomplete, PA + the ω-rule is complete. This, though, does not contradict Gödel's first incompleteness theorem concerning consistent and recursively axiomatizable extensions of PA (or Q – in its modern variant), since PA + the ω-rule is no longer recursively axiomatizable, as should be clear from the very nature of the infinitary ω-rule. Carnap's view seemed to be that *prima facie* semantic concepts of analyticity or consequence are something like quasi-syntactic concepts, for which we relaxed the requirement of *definiteness* (algorithmic decidability). Indeed, one of his definitions of analyticity and consequence is in terms of the ω-rule. Such concepts, however, can be defined in a higher order „syntax-language" with a more powerful apparatus of quantification. See Peregrin (2006) for an interesting discussion of Carnap's suggestion that the notion of consequence is a notion of quasi-inference, resting on the notion of inference, provided with relax inference rules in certain ways (e.g. allowing for infinitary inference rules).

[143] Even earlier, around 1930, he started to be influenced by Tarski's approach to metamathematics, based on the distinction between object-language and metalanguage. But, as Coffa (1991) rightly notes, there was nothing yet that he could learn from Tarski on the semantic part of metamathematics.

[144] Kleene (1939) and MacLane (1938) were the first to point out some problems with Carnap's approach in their critical reviews of Carnap's (1937). See also Coffa (1991), Procházka (2006, 2010) or de Rouilhan (2009). The general tenor of my discussion here is indebted to the account given by Awodey (2007).

[145] Language II was interpreted by Carnap as a coordinate language, in which numerals obtained from the primitive '0' by repeated application of the functor corresponding to the successor-functor denote corresponding positions in the sequence of numerals starting with '0'; predicates and functors are then interpreted accordingly over the domain of numerals. Carnap's definition of analyticity in (1934) is reductive but not, strictly speaking, recursive. (I) It reduces every sentence of Language II to an equivalent sentence that was either atomic or in the prenex-form. Then (II) rules of valuation are laid down, which assign to terms of atomic formulas (variables, individual constants, *n*-place functors and predicates) appropriate values in accordance with

(a) $a = b$ is analytic iff v(a) = v(b)
(b) Fa is analytic iff v(a) belongs to v(F)
(c) $\neg A$ is analytic iff A is not analytic
(d) $A \wedge B$ is analytic iff A is analytic and B is analytic
(e) $\forall x\, Ax$ is analytic iff An is analytic, for every numerical constant n,

where 'v(e)' represents the interpretation of a non-logical constant e in a given domain (in the case under consideration, the domain is assumed to be the set N of natural numbers so that the interpretation of an individual constant is a number and the interpretation of a 1-place predicate is a subset of N). Now, when we replace *analytic* with *true*, what we get is a recursive Tarski's truth definition for an arithmetical object language framed in the metalanguage (having a name for every object in N), as interpreted in something like the standard model of arithmetic. Carnap seemed to be dimly aware of this striking parallel when he said that when we have a sentence of the form *Pr(arg)*, the sentence "is - so to speak - true on account of" valuation V just in case the value of *arg* under V is an element of the value of *Pr* under *V*, and false otherwise.

As Coffa comments: "nowhere was Carnap closer to the semantic conception of truth than at this point".[146] However, Carnap refused to use the unqualified terms "true" or "false" – hence the hedge "so to speak" - since they connote factual truth or falsity and, as he confessed himself, before Tarski explained to him the semantic conception of truth, he did not conceive of the possibility of defining plain truth, that is, truth as applying also to factual (empirical) sentences, not just to logico-mathematical sentences (non-factual).

---

suitable type-theoretic restrictions. Finally (III) basic rules of evaluation are specified by which analytical truth or falsity of atomic sentences with respect to a given valuation is directly reduced either to the sentence '0 = 0' or else to the sentence '0 ≠ 0' (the first represents the most elementary case of analytical truth and the second represents the most elementary case of analytical falsehood). According to the remaining rules of evaluation, analytical truth (falsity) of quantified sentences is reduced to analytical truth (falsity) of their quantifier free matrixes, whose analytical truth (falsity) is in turn reduced in a step-by-step manner to analytical truth (falsity) of their atomic formulas with respect to *valuations*. The successive reduction take the form of the rules of evaluation that tell us that we can reduce

> (a) the atomic formula of the type *Pr*(*arg*) to the sentence '0 = 0' (itself reckoned analytically true by default) with respect to a given valuation V just in case the value of *arg* under V is a member of the value of *Pr* under V, otherwise to the sentence '0 ≠ 0' (itself reckoned analytically false – contradictory – by default).
>
> (a) the basic formula $arg_i = arg_k$ to the sentence '0 = 0' with respect to a given valuation V just in case the value of arg_i under *V* is a member of the value of *Pr* under V, otherwise to the sentence '0 ≠ 0'.

The remaining rules of evaluation for truth-functional quantifier free formulas should be obvious (take the recursive clauses (b),…(d) and relativize them at each appropriate point to V), and the rules of evaluation for every quantifier free formula lead in a finite number of steps either to '0 = 0' or to '0 ≠ 0'. A quantifier free formula is analytical (not just analytical with respect to a given valuation) iff its evaluation leads to '0 = 0' for every valuation V of it (contradictory iff its evaluation leads to'0 ≠ 0' for every such V). A closed quantified sentence of the form$\forall xAx$ is analytical iff its matrix *Ax* is analytical with respect to every valuation V of its free variable (and contradictory iff its matrix *Ax* is contradictory with respect to every valuation V of its free variable). Kleene (1939) was the first to transform Carnap's definition into the recursive form.

[146] Coffa (1991: 293).

Indeed, he reports that when Tarski told him that he succeeded in defining truth, he initially thought that he must have in mind some syntactic notion of derivability in a system:

> „I was surprised when he [Tarski] said that he meant truth in the customary sense, including contingent factual truth. Since I was thinking only in terms of a syntactical metalanguage, I wondered how it was possible to state the truth-condition for a simple sentence like 'this table is black'. Tarski replied: 'This is simple; the sentence "this table is black" is true if and only if this table is black.'" (Carnap 1963: 60-61).

So, in Carnap's view, the definition given above amounted to the definition of analytical truth, which, following Wittgenstein, he equated with non-factual truth. By a coincidence of circumstances, he defined also a plain truth for such a language, because in case of logico-mathematical languages truth coincides with L-truth. However, he was not aware that he could in that way define truth for descriptive languages (for a physicalist expansion of Language II), since it did not occur to him then that something like Convention T can be used as a material adequacy constraint on such a definition, though he himself used a similar criterion for analytical truth:

> ⟨*A*⟩ (of Language II) is L-true iff *A*,[147]

where '⟨*A*⟩' is a perspicuous designator of *A* in the metalanguage. Had it occurred to Carnap that much the same paradigm can be used as an adequacy-criterion for factual truth, he would have realized at once that he could define simple truth for languages with descriptive signs, as his higher-order syntactic metalanguage contained translations of all the expression of the object language.

At any event, the trouble with this procedure for defining analyticity, qua logical truth, is that it does not neatly extend also to languages that contain descriptive (non-logico-mathematical) constants, and it was Carnap's professed aim to provide a general method of defining analyticity and related notions of contradictoriness or consequence also for descriptive languages of science (e.g. for physicalistic extensions of Language II). Carnap's definition, if conceived of as the definition of simple (non-analytical) truth, can be easily extended to the case when *Fa* means, say, "Chicago is a large city":

> *Fa* is true iff v(*a*) (i.e. Chicago) belongs to v(*F*) (i.e. the set of large cities).

or

> ¬ *Fa* is true iff *Fa* is not true.

However, if we conceive of it as the definition of analyticity, then the question arises how to extend it to such cases as these? Clearly, neither "Chicago is a large city" nor its negation is analytically true by any remotely plausible conception of analyticity. Carnap's proposal was to fix this by saying that in the specific case of a descriptive sentence *A*, *A* is analytical just if *A** is analytical,

[147] Carnap (1934, § 62b: 214).

where *A** is a sentence obtained by replacing all descriptive constants of *A* uniformly by variables of appropriate types and universally closing the matrix so obtained with respect to each free variable introduced. This proposal is technically all right, so far as it goes, and it has its precursors in the logical tradition in the idea that a logically true sentence is true no matter how we reinterpret its non-logical sings.

Nevertheless, recent discussions show that the overall plausibility of Carnap's general approach to the problem of defining analyticity and logical consequence stands and falls in last instance with the possibility of drawing some principled distinction between descriptive and logical expression. Carnap was aware that the crucial step towards the satisfactory solution of his problem consists in distinguishing logical from descriptive (non-logical) constants. In this connection, he mentions the main difference between him and Tarski, who urged relativistic attitude to the distinction between logical and non-logical (descriptive) signs and consequently between logical (analytical) and non-logical truth:[148]

> "My conception of semantics starts from the basis given in Tarski's work, but differs from his conception by the sharp distinction which I draw between logical and non-logical constants, and between logical and factual truth." (Carnap 1962: 62)

Clearly the second distinction hinges on the first. Carnap spelled it out in the following way:

> "...the distinction between factual truth, dependent upon the contingency of facts, and logical truth, independent of facts and dependent merely on meaning as determined by semantical rules" (Carnap 1942: xi)

But it is by no means clear whether Carnap's way of distinguishing the two classes of expressions in (1934) or in his later writings can be deemed satisfactory, although his proposals are quite interesting in their own right.[149]

At any rate, one has the feeling that had Carnap identified the problems that we have just talked about, he would have realized that what he was in fact so close to providing in (1934) was not a general method of defining analytical truth, but a general method of defining truth for a range of formalized languages. Tarski,

---

[148] In (1936a) Tarski says that his classic definition of logical consequence in terms of models hinges on the possibility of drawing a more-or-less reasonable distinction between logical symbols (fixed) and non-logical symbols of sentences (unfixed – reinterpretable, that is, replaceable by variables of appropriate orders which may be assigned various values of appropriate types in the form of various sequences of entities satisfying resulting sentential functions), while admitting that he has no clear-cut criterion. He suggests, in a liberal manner, that different choices of sets of logical constants might yield extensionally different accounts of logical consequences, the most extreme case being when we take all expressions to be logical. In between there are various less extreme and possibly useful choices: if our language contains synonymous pairs such as *bachelor/unmarried man*, we may take them to be fixed so that a sentence of the form *A is an unmarried man* logically follows from another sentence of the form *A is a bachelor*.

[149] Carnap (1935: 177-178) defines the class of logical expressions as the largest class of terms of the language such that every sentence which contains only members of this class (and variables) is determinately true or false on the basis of the transformation rules of the language alone. Cf. Awodey (2007) or Frost-Arnold (2006).

on the other hand, was clear on the matter and carried out the agenda splendidly in CTFL. Carnap's perspectival change towards semantic methods is first manifested in a systematic way in his *Introduction to Semantics*, but the main problem remains the same: to provide a plausible account of analytical truth and consequence relation, this time, by exploiting and modifying techniques that Tarski developed in combination with his own ideas.[150] According to Carnap, the study of historical languages used by linguistic communities is an empirical, descriptive study of systems of communication (of habits, regularities or conventions prevailing in linguistic communities), whereas the study of abstract languages, qua semantical systems, is more an arm-chair investigation of abstract objects whose properties we are free to stipulate (though, possibly, with an eye to fruitful comparisons with languages-in-use, semantical systems serving as their formal models that abstract from certain features of theirs in order to make other aspects more perspicuous and easy to handle). He distinguished three fundamental aspects of language: (1) speakers, (2) expression uttered by speakers, (3) things so designated. Accordingly, there is a useful division of theoretical labour within the general semiotics: *pragmatics* deals with speaker-related aspects, *semantics* with designation related aspects (abstracting from speakers) and *syntax* with expression-related aspects (abstracting both from speakers and designata).[151] Finally, he distinguished *descriptive* from *pure semantics* (and descriptive from pure syntax): the first conducts empirical study of semantical features of existing (or historical) languages (particularly or generally), while the later is concerned with "construction and analysis of semantical systems" and "consists of definitions", being "...entirely analytic and without factual content".[152]

What exactly did Carnap have in mind? He said that a semantical system S specifies an object-language L in the metalanguage ML by means of (A) syntactic-formation rules (that specify what counts as a meaningful expression of L) and (B) semantic-interpretation rules specifying the truth-conditions for all sentences of L, following Frege and Wittgenstein (of the *Tractatus*) in claiming that knowing the truth-conditions of a declarative sentence amounts (almost) to knowing its meaning:

> "By a semantical system (or interpreted system) we understand system of rules...of such a kind that the rules determine a truth condition for every sentence of the object language...In this way the sentences are interpreted by the rules, i.e. made understandable because to understand a sentence, to know what is asserted by it, the same as to know under what conditions it would be true." (Ibid: xi).

We have to keep in mind that, at this point, Carnap did not make any particularly controversial claim, for he was careful not to apply it to natural languages (his historical languages in everyday use). As Tarski, he was initially sceptical regarding the prospects of applying ideas and techniques that proved so fruitful in studying formal or formalized languages directly to natural languages, being a leading exponent of artificial languages better suited to serve specific purposes of science than colloquial languages, which do not suffer from various "defects" of natural languages and whose properties could be easily studied, being

[150] Though there is a clear account of it already in his contribution to the Encyclopaedia of Unified Science (1939).
[151] Carnap (1942: 8).
[152] (Ibid: 11-12).

the creatures of our own. What truth conditions are supposed to model in the framework of pure semantics are *logical meanings* of sentences, the semantic value of a sub-sentential expression of a certain type encapsulating its specific contribution to truth-conditions of sentences containing it, which, in turn, is nothing less and nothing more (sic!) than what the logical constants of the object language are supposed to be sensitive to. As Carnap was still more of an extensionalist in this period (along with Tarski and Quine),[153] the logical meanings were to be extensional in character.

In the style that reminds us of Tarski, Carnap distinguishes code-languages that contain finitely many sentences and are capable of conveying only a limited number of statements from languages (*Sprachsysteme*) containing an infinite number of sentences. For a code-language it is possible to give truth-conditions for all its sentences (and so its truth definition) directly by enumeration (list). For code-languages whose sentences display (e.g. a subject-predicate) structure of constituent parts it is also possible to specify the rules of denotation for their expressions (typically: for names and predicates) and then introduce a compositional rule to the effect that a sentence consisting of a predicate followed by a name is true just in case the denotation of the name (an individual) has the denotation of the predicate (a property). The two procedures are extensionally equivalent, though the second is more illuminating of the logico-semantic structure of L (note the analogy with semantically non-illuminating Tarski's truth-definitions for finite languages and semantically more illuminating recursive definitions). On the other hand, in case of non-code languages only the second semantic method is possible - one involving various recursive rules - depending on the complexity of L - specifying truth conditions of compound sentences in terms of truth- or denotation-conditions of their immediate component parts.

Let us consider Carnap's own example (1942: 32), modified to cover quantified sentences (S is a semantical system that first describes and then interprets an object-language L in the metalanguage ML). We shall consider the object language L, whose lexicon consists of a stock of individual variables (metavariable: *v*) 'x_1','...,'x_n', two individual constants (metavariable *in*) '*a*', '*b*', two predicate constants (metavariable: *Pr*):'*P*', '*Q*', and the logical signs '¬', '∨', '∀' (all having their usual interpretations that remain fixed across various semantical interpretations of L in different semantical systems).

I Rules of formation:

A sentence of L is an expression of the form (a) *Pr(in)*;[154] or (b) ¬*A*; or (c) $A \vee B$, or (d) $\forall v_i A$.

II Rules of designation:

(a) '*a*' designates Chicago;

(b) '*b*' designates New York;

[153] The matter is actually much more complex than I indicate, because already during this period Carnap defended an intensional definition of analyticity against Tarski's and Quine's extensional proposals. See the extensive discussion in Frost-Arnold (2006).

[154] In place of a) we could simply list all the atomic sentences of L – there being only four of them.

(c) '*P*' designates the property of being large;

(d) '*Q*' designates the property of having a harbour.[155]

III **Rules of truth:** a sentence *s* of L is true (in *S*) iff one of the following conditions is satisfied

(a) *s* is of the form *Pr(in)* and the object designated by *in* has the property designated by *Pr*;

(b) *s* is of the form $\neg B$, and *B* is not true;

(c) *s* is of the form $A \vee B$, and at least *A* is true or *B* is true;

(d) *s* is of the form $\forall v_i A$ and every substitution-instance $A(n/v_i)$ of *A* is true.

In a nutshell, this is the basis of Carnap's semantic method from (1942), and essentially the same ideas inform also his *Meaning and Necessity* (1956), except that the focus shifts to modal languages that, Carnap argued, require the method of intension that would complement the method of extension developed in 1942 (roughly speaking: extensions being relativized to descriptions of possible states). One may complain that the account is not satisfactory on the ground that quantifiers are interpreted substitutionally, which method gives correct results only when applied to a language that has a name for every object in its universe. Actually, Carnap's account does not suffer from this defect (much the same can be said of his account of analytical truth given in 1934, which has been often criticised on this count), since he explicitly mentions that the range of names replacing variables is not confined to the names of the object language, but is to be extended so as to cover all individuals in the universe of discourse. The effect of this move is the same, as if one talked directly of objects satisfying formulas. Granted, for a language with a non-denumerable universe the procedure does not work (or at least not straightforwardly). But Carnap was prepared to talk directly of objects satisfying formulas.

In many respects, Carnap's semantics is arguably a predecessor of the modern model-theoretic approach, according to which an uninterpreted formal language L (typically, 1st order) can be variously interpreted and re-interpreted in admissible L-structures. Indeed, *A (of L) is true in S* can well be read as *A (of L) is true under the interpretation S* (or: *A, as interpreted in S, is true*), S being specified via the semantical rules of denotation and truth such as (II) and (III). Different such rules yield different interpretations of L - in Carnap's parlance, different semantical systems or even different languages, since he tends to individuate languages semantically, so that L, as interpreted in S, is a different language than L, as interpreted in S*, provided that S and S* differ in their rules of designation. Note that what remains invariant across different interpretations

[155] Like Tarski, Carnap also formulates explicit versions of (I-III); for designation, for instance, we have this:

> (For every name *n* and object *o*): *n* designates *o* (in S) iff (a) *n* = "*a*" and *o* = Chicago, or (b) *n* = "*b*" and *o* = New York, or (c) *n* = "*P*" and *o* = the property of being large, or (d) *n* = "*Q*" and *o* = the property of having a harbour.

(semantical systems) of L are compositional rules of truth that implicitly take care of the logical part of L (truth-functional operators and quantifiers), which feature has an obvious parallel in the model theory, since different set-theoretical interpretations of the 1st order L differ only in what they assign to non-logical primitives of L's signature, but agree in their logical part, whose interpretations remain fixed across them.

This approach influenced the model theoretic approach to semantics in one crucial respect. As opposed to Tarski's approach in CTFL, Carnap's approach makes use of the notion of an *uninterpreted non-logical constant* (individual or predicate), for which various interpretations are considered in the form of various semantical systems. We shall see that Tarski was reluctant to make a systematic use of this idea until the late 1940s, and it could have been be precisely this aspect that Carnap wanted to stress when saying that he wanted to draw a sharp line:

> „...between semantical systems as interpreted language systems and purely formal, uninterpreted calculi..." (Carnap 1942: vii),

adding, in the same breath, that

> "...for Tarski there seems to be no sharp demarcation" (Ibid: vii).

Both uninterpreted languages and calculi framed in them have the category of uninterpreted non-logical constants, as distinguished from logical constants with fixed interpretations on the one hand, and individual variables on the other hand.[156] Now, as a remark on Tarski, this sounds initially puzzling, as Tarski put a very strong emphasis on the difference between formal languages lacking any interpretation and formalized languages - meaningful, fully interpreted formalisms. What Carnap could have in mind is that Tarski did not work with uninterpreted languages, whose non-logical constants can be variously interpreted via assigning them different values in different semantical systems. Second, Tarski did not care much to distinguish languages (interpreted or interpreted) from deductive theories (calculi) framed in them. For him, formalized language is a language of a certain deductive discipline (calculus of classes, arithmetic, elementary geometry, etc.); accordingly, his specification of the object language in the metalanguage involves not just the syntax in the narrow sense (Carnap's formation rules) but a deductive system as well (axioms plus Carnap's transformation rules). Carnap, on the other hand, thought of semantical systems as ways of fixing interpretations (as specified by the semantical rules) for uninterpreted object-languages (as specified by formation rules). Via a semantical system S of L also an uninterpreted deductive calculus C framed in L can be interpreted in the following sense:

> S provides an interpretation of C iff S assigns (in a recursive manner) a criterion of truth to each sentence of C (that is: a Tarski-type biconditional satisfying Carnap's variant of Convention T).

or indeed, C can be given a true (sound) interpretation by S in the following

[156] Kemeny (1949), in one of the first standard model-theoretic accounts, explicitly appeals to Carnap's ideas, though their accounts differ in certain significant respects.

sense:

> S provides a true interpretation of C iff all C-theorems (axioms of C plus their consequences obtained from the axioms by repeated applications of transformation rules) are true sentences of S.

This corresponds to the notion of soundness of a deductive system with respect to an intended interpretation (such as soundness of **PA** under the standard interpretation).

Finally, what Carnap had to say about L-truth is that a sentence of S is L-true just in case it is true in virtue of the semantical rules of S alone – its truth-value being determined by such rules alone. What he meant by this was that a sentence of S is L-true in virtue of the recursive (compositional) rules of truth supposed to implicitly fix the meanings of logical constants, hence independently of the interpretations that the rules of denotation assign to its non-logical constants.[157] In the Encyclopaedia entry he says:

> "We call a sentence of semantical system S (logically true or) L-true if it is true in such a way that the semantical rules of S suffice for establishing its truth. If a sentence is either L-true or L-false, it is called L-determinate, otherwise (L-indeterminate or) factual. (The terms L-true, L-false, and factual correspond to the terms analytic, contradictory, and synthetic, as they are used in traditional terminology." (Carnap 1938: 155)

In the *Logical Semantics* Carnap refines this account by saying that *a sentence of S is L-true iff it is true in every state of affairs in S*, where a state of affairs in S is given by a complete assignment of S-predicates to the individuals of the universe of S (the universe being specified by a special semantical rule of values that stipulates what the variables range over in S) so that each *n*-place predicate is to be assigned a set of ordered *n*-tuples of individuals of the universe of S.[158] In slightly different words, a state of affairs is determined by any complete assignment of truth-values to atomic sentences, representing a possible world in roughly the sense that Carnap read off from Wittgenstein's *Tractatus*.[159] This, of course, is not a matter of logic or semantics, but a purely empirical matter of facts (at least when we work with descriptive languages). Carnap also modified his earlier statement to the effect that L-truth is truth in virtue of the meanings "of logical sings alone" (in virtue of the recursive semantical rules alone), saying that L-truth is truth in virtue of the meanings of

---

[157] On the basis of his previous definition of L-truth of S as a sentence true in all state descriptions in S, Carnap defined an L-true interpretation of C as follows:

S provides an L-true interpretation of C iff all the theorems of C are L-true sentences of S.

[158] Carnap is thus able to define synonymous expressions as those as those that have the same extension in all states of affairs (two sentences being synonymous iff they have the same truth value in all states of affairs).

[159] Indeed, already in this period, Carnap thought of properties assigned to predicates as essentially intensional entities: something that determined different extensions in different states of affairs. For a good discussion see Frost-Arnold (2004).

"all sings" (and of nothing else), because he wished to account also for apparent analyticity of non-logical sentences such as "Every bachelor is unmarried" and the like. Accordingly, if the language under consideration contains pairs of logically dependent predicates such as "bachelor" and "unmarried", Carnap places some extra-constraints on assignments of predicates to individuals, if they are to determine states of affairs: an assignment of the predicates to the individuals that determines a genuine state of affairs must assign to the predicate "bachelor" the class of individuals that is included in the class assigned by it to the predicate "unmarried".

Plausible as these explanations may at first appear (the compositional rules of truth are supposed to be general – being the same for various semantical systems in which L might be interpreted), there are two serious problems with it. First, despite his proclamations, Carnap failed to provide a principled criterion for distinguishing logical from non-logical constants. He was subjected to a vigorous critique from Tarski and Quine, who discussed with him the topic intensely in the early 1940s.[160] Furthermore, since his semantics anticipated the model-theoretic approach in that it allowed for various (re)interpretations of a language L in different semantical systems, one would naturally expect Carnap to define logico-analytical truth as truth in every semantical system S appropriate to L, which would eventually amount to much the same as truth in every structure (appropriate to L). Yet, Carnap's account allows only for reinterpretations (via uniform replacements) of non-logical constants of a sentence within a single semantical system S with its fixed domain (specified by the rule of values), whereas the model-theoretic account of logical truth (validity) requires truth under all interpretations where interpretations (structures) have different domains.[161] Awodey correctly says in this connection:

> "In order to determine logical truth (what we now call logical validity), it does not suffice, in general, simply to substitute different constant symbols and check the result in a single interpretation. Instead, the idea that the truth of a logically true sentence is independent of the interpretation of its non-logical symbols is captured, from a modern point of view, by considering the range of all possible interpretations of these symbols over all possible domains of quantification. It is only thus that we can show, for example, that every semantic consequence of a logical truth is itself a logical truth, thereby ensuring that logical truth is empirically empty. In the "model-theoretic" terms of modern logic, what is required is the difference between truth in a particular mode and truth in all models." (Awodey 2007: 237-38).

Carnap was close to the modern model-theoretic account of relative truth and of semantic validity, but he did not succeed in formulating it. If the analysis that I have offered is on the right track, he was surely right to emphasize the differences between him and Tarski. (1) Tarski did not conceive of language-relative semantic definitions as interpreting an hitherto uninterpreted object language; (2) Tarski's truth definitions presupposed that expressions of an

[160] An amazing material about these sessions is contained in Frost-Arnold's (2006).

[161] Carnap was aware of this inadequacy and in the footnote on the p. 85 of (1942) he makes some suggestions how to give a more plausible account that uses the notion of logical necessity. The idea is developed in more detail in (1956) as truth in all possible state-descriptions.

object-language are meaningful, their meanings being captured in their (correct) metalinguistic translations, with which the definition-giver (or evaluator) is supposed to be familiar in advance so that he can use this knowledge in constructing the materially adequate truth definition (or evaluating its material adequacy). With purely stipulative definitions á la Carnap, on the other hand, there is nothing to care about, since there is no antecedent meaning/interpretation to be preserved by semantical rules. The interpretations are simply stipulated via such semantical rules. Carnap calls such rules "definitions" or "analytical", but only in the sense of being "stipulative" or "non-factual", and not in the Kant-Frege's sense, according to which analytical definitions capture the meaning of expressions already in use. Carnap distances himself from this notion of analytical definition when he says that stipulative semantical rules cannot be confirmed or disconfirmed in light of the facts about usage – be it actual or historical.

In this section I have attempted to show, *inter alia*, that Carnap's semantic method has important points of contact with Tarski's view of semantics but also that it differs in several important respects, which we should keep in mind. Too often the two conceptions have been confused for each other in the literature. We shall see that these differences play a certain role when we consider the so-called modal objection levelled against Tarski's conception of truth by Hilary Putnam and others (and related conceptions

[5]

Relative Truth

5.1 Model theory: interpretations and uninterpreted languages

In the preceding chapter I attempted to persuade the reader that one of the chief logical aims that Tarski's semantic conception of truth was aimed to achieve was to show that meta-mathematics, as theory of truth-theoretic (semantic) and proof-theoretic (syntactic) properties of formalized deductive theories and their relations can be practised in a mathematically precise spirit. This appeared as a valuable approach, since by providing analyses (explications) of basic metamathematical notions of the semantic origin within the mathematical framework of the general set theory he hoped (1) to attract the interest of mathematicians to metamathematics, (2) to give assurance of consistency of semantic methods in metatheorizing, and (3) to effect kind of rehabilitation of semantic notions *vis-á-vis* philosophical worries that were then current (truth as a metaphysically loaded hence discredited idea incompatible with natural science, etc.). I have discussed the merits and demerits of these reasons in connection with an alternative approach represented by truth-axiomatizations, having made some critical points concerning (2) and (3). What remained intact was the meta-mathematical power of Tarski's method of truth definition for properly regimented and interpreted object-languages in logically stronger metatheories, which, due to its recursive character, entails elementary yet important generalizations involving the notion of truth (e.g. that a conjunction is true just in case its conjuncts are true; that of a sentence and its negation exactly one is true and one is not true, and the like). Only on such a basis, Tarski repeatedly stressed, basic metatheorems can be precisely stated and proved about consistency, soundness or completeness of T.

But Tarski's contribution to semantics is by no means exhausted by his theory of absolute truth. Together with his circle of students and collaborators he played a major role in the boom of model theory in the 1950-60s, which is nowadays a firmly established mathematical discipline with many interesting applications in algebra, analysis, geometry or topology. It is often called *a theory of definability*, one of its chief concerns being that of delimiting classes of mathematical systems obeying axioms of formal theories, and studying relations between axioms (qua laws of a sort) and systems obeying them (qua models or realizations of the laws). Tarski saw the nature of model theory as follows:

> "I should like to point out a new direction of meta-mathematical research —the study of the relations between models of formal systems and the syntactical properties of these systems (in other

> words the semantics of formal systems). The problems studied in this domain are of the following character: Knowing the formal structure of an axiom system, what can we say about the mathematical properties of the models of this system; conversely, given a class of models having certain mathematical properties, what can we say about the formal structure of postulate systems by means of which we can define this class of models." (Tarski 1954: 19-20)

Other theorists prefer saying that they are concerned directly with mathematical structures, approaching and studying them through statements of formal languages that hold in them. However one prefers to characterize model theory, at bottom, it is founded on the "relative" notion of *a sentence being true in a structure* (conversely: *a structure being a model of the sentence*), and, more generally, on the notion of *a formula being satisfied in a structure by a sequence of individuals from the domain of the structure*. Precise definitions of these model-theoretic notions were provided by Tarski and they are natural extensions, with appropriate modifications, of his method of absolute truth definition. Or so, at least, it is commonly believed. That view, I think, is right, as far as it goes. But the story is actually more complicated that it may insinuate. First, it took Tarski almost two decades since he had finished CTFL until he finally came to adopt a full-blooded model-theoretic account of satisfaction, truth and the related notions of model and logical consequence. Second, one should not forget that the ideas were long in the air, so that no single author can be safely identified as *the* man who defined truth and satisfaction in a structure, though Tarski would have been the best candidate, had such a person existed.

I shall consider the model-theoretic method of definition of relative truth in all formal details in the next section. Let me now explain, in a less formal manner, how the relative notion of truth in a structure differs from the absolute notion of truth. Imagine that L is a language of the sort that Tarski considered in the main body of CTFL, except that all save its logical constants are stripped off their interpretations. L is thus not fully interpreted; rather, it is uninterpreted, albeit not fully, since its logical constants have fixed interpretations. Consequently, L is to be definitely distinguished from Tarski's formalized languages:

> "[...] we are not interested here in 'formal' languages and sciences in one special sense of the word 'formal', namely sciences to the signs and expressions of which no meaning is attached. For such sciences the problem here discussed has no relevance, it is not even meaningful. We shall always ascribe quite concrete and, for us, intelligible meanings to the signs which occur in the languages we shall consider." (Tarski 1935: 166–67)

The problem referred to by Tarski is that of giving a satisfactory definition (or theory) of truth for sentences, and he maintains that any attempt to define truth is sensible only for sentences of a given language having definite meanings - hence the need to relativize the truth definition to a given formalized language so as to rule out any indefiniteness (including ambiguity, context-sensitivity or vagueness, which features are absent from a properly regimented scientific language). Unfortunately, the passage is not unambiguous. Did he want to say "all the signs and expressions" or, more qualifiedly, "all the non-logical sings and

expressions"? The difference matters, because only the later reading is compatible with the standard account of uninterpreted languages. Be that as it may, Tarski evidently thought that when it comes to define truth (*simpliciter*), uninterpreted formal languages are out of place: lacking definite meanings, their sentences cannot be evaluated with respect to their truth or falsity.

Admittedly, we cannot define what it means for an L-sentence to be true *simpliciter*, but we can at least make precise a somewhat related idea. What does a formalized language of Tarski-type talk about? Presumably, about entities belonging to a certain domain (e.g. natural numbers) its sentences expressing that such entities possess (or not) such-and-such properties (expressed by 1-place predicates) or bear (or not) such-and-such relations to one another (expressed by *n*-place predicates, for $n > 1$), eventually quantifying over the domain in order to express possession (or not) of such-and-such properties or relations by all/some elements of the domain. If so, what is to prevent us from conceiving of various interpretations of an uninterpreted L, which would make L to talk about a given domain of entities, L's sentences expressing their properties or relations via quantifiers, terms or *n*-place predicates? And if L is an uninterpreted but interpretable language, we can ask what it would take for a sentence *A* of L to be true relative to this or that conceivable interpretation of it. We know that if L is a fully interpreted formalized language, the method of absolute truth-definition can be applied, defining what it means for a sentence of L to be plain true, based on the recursive characterization of what it means for a sentential function of L to be satisfied by a sequence *s* of elements from the domain of entities that L talks about. Tarski idea was that when L is a formal language, then the method of definition of *relative truth* applies, defining what it means for a sentence *A* of L to be true in a structure *M*, the definition being based on the recursive definition of what it means for a formula *F* to be satisfied in *M* by a sequence *s* of elements from the domain of *M*.[162]

5.2 *Structure*: the idea of *interpretation* made precise

L-structure, as this notion is understood in model theory, is designed to make precise the informal idea of an admissible interpretation of sentences of the formal language L, relative to which truth of L-sentences is determined. Accordingly, an L-structure is given once a non-empty set is specified as its domain and certain labelled individuals, *n*-place functions and *n*-place relations on that set. If $\mathfrak{R}$ is a structure and T a theory framed in L, the question whether T-axioms hold in $\mathfrak{R}$ arises only if $\mathfrak{R}$ is an L-structure, that is, a structure in which L can be interpreted so that its sentences are divided into those true in $\mathfrak{R}$ and those not true in $\mathfrak{R}$. Once the primitive non-logical constants are reckoned to basic syntactic categories, each constant is interpreted by assigning it exactly one set-theoretical entity appropriate to its category and defined over the domain assigned to L-quantifiers. Moreover, the assignments are to be such that, in cooperation with the fixed interpretations of the logical constants, they suffice to fix truth-values of L-sentences. The rationale for this approach is clear. The questions that logicians want to have answered are of the sort: given a sentence of such-and-such a form - e.g. $\forall x(Fx \rightarrow Gx)$ – and the interpretations of the universal quantifier and the

[162] See Tarski and Vaught (1956).

conditional, what information is it necessary and sufficient for an interpretation to supply in order to determine its truth-value? The idea is that admissible interpretations of non-logical constants of L encapsulate so much and only so much information as is needed to determine the truth-values of L-sentences, in accordance with the interpretations of the logical constants. Thus, for instance, the quantifier is sensitive only to extensions of predicates in its domain. The idea of the *logical meaning* (or *semantic value*) of a non-logical constant C is the idea of a truth-relevant feature possessed by C of L, to which solely the logical constants of L are sensitive, which operate on non-logical expressions of C's type. An interpretation of L is then a systematic assignment of such truth-relevant features to primitive non-logical constants.

Since Aristotle logicians sought to delimit the class of valid inference schemata, whose instances never combine true premises true with false conclusion. The idea was that all particular inferences falling under such a schema preserve truth in virtue of their form alone, the form being represented by the schema. Bolzano elaborated on this idea by his account of logical truth without the detour through schemata: the sentence *A* is logically true just in case every sentence is true, which is obtained from *A* by uniformly replacing all except its logical expressions by others of appropriate types.[163] That *B* follows from sentences $A_1,...A_n$ can in turn be accounted for as a special case of logical truth of the conditional: 'If $A_1,...A_n$, then *B*'.[164] Bolzano could as well start with defining logical consequence first (without the detour through schemata), then defining logical truth as a limiting case of logical consequence from the empty set of premises. Bolzano's account is thus based on substitutions, but these are performed directly on sentences with respect to their variable expressions. The effect, though, is much the same. Now, the model-theoretic idea of an interpretation of a formal sentence in an abstract structure also aims to capture the truth relevant features of substitution-instances (of schemata or sentences). Its comparative advantage is that it does not stand and fall with the expressive richness of a language under consideration – availability of a name for each element of the domain - which problem may become serious for substitutional accounts. Moreover, even though there is a structure for every true instance, structures are more manageable than instances, encapsulating the truth-relevant features of instances that may well be shared by several instances. Thus one row in the truth-table for a truth-functionally compound formula *F* represents the dependence of its truth-value on one possible distribution of truth-values to its components, encapsulating in this way the truth-relevant features of any of an indefinite number of particular instances of *F*, whose components have so distributed truth-values.

Let me now explain the seminal ideas of model theory on the paradigmatic case of a 1st-order language L containing a stock of logical signs

---

[163] More precisely, Bolzano talked not about sentences of a natural language but about propositions - *Sätz an sich* and their variable and non-variable (fixed) component representations (ideas) – *Vorstellungen an sich*. A universally valid proposition is such that, relative to the selected set of its variable component-representations, every proposition obtained by replacing the variable representations by other appropriate representations is true. Logically analytic sentences are a special case, where the selected set of variable representations contains all its non-logical representations. Viz. Bolzano (1837).

[164] Alternatively, as logical truth of the sentence: *It is not the case that A_1 and ...and A_n and not B.*

having fixed interpretations, typically {$\forall$, $\exists$, $\wedge$, $\neg$, $\vee$, $\rightarrow$, =}, together with an infinite (but denumerable) number of individual variables {x_1 ,..., x_n ,....}.[165] In addition, L has parametric expressions having no fixed logical meanings:

a) denumerable (possibly empty) set of individual constants{c_1, c_1,..., c_n,...};

b) denumerable (possibly empty) set of *n*-place predicates {$^1P_1$, $^1P_2$,..., $^2P_1$, $^2P_2$,..., $^nP_1$, $^nP_2$,...};

c) denumerable (possibly empty) set of individual constants{c_1 ,..., c_n ,...}, *n*-place function symbols {$^1f_1$, $^1f_2$,..., $^2f_1$, $^2f_2$,..., $^nf_1$, $^nf_2$,...},

Superscripts indicate the number of places that a predicate or function symbol carries with it. Quantifiers are usually included in the logical basis, but they are parametric in character, since '$\forall x$' is to be read as: "for all elements of (the domain) U", U being not fixed once and for all, because different set-theoretic interpretations of L in different L-structures bring in different values for U. This proves quite important when it comes to define the properties of logical truth, consequence and semantic consistency (satisfiability).

In logic one is not interested so much in any particular 1st-order language, but rather in a whole range of 1st-order languages of the same structure: namely with the same logical basis and basic categories of parametric expressions, though possibly differing in what particular individual constants (if any), predicates (if any), or function symbols (if any) fall in these categories. That is to say, they possibly differ in their *signature*. This feature makes it possible to provide a fairly general account of logical properties for an arbitrary language of the type specified above. The exposition to follow is meant to adhere to this practice. I first give a general model-theoretic account of 1st-order languages of a certain standard type with the definitions of satisfaction and truth under an interpretation in $\mathfrak{R}$. Similarly as L, also $\mathfrak{R}$ can be conceived of as a variable ranging over structures appropriate to languages over which L ranges. Later on I will show how to apply this general framework to the particular case of a given formal language L(PA) and a given structure (*N*) and I shall discuss its relation to the absolute method of truth definition.

The idea of an interpretation of a 1st-order L in an L-structure $\mathfrak{R}$ is formally implemented by assigning:

a) a non-empty set U to the quantifiers of L – U being called the domain (or universe) of $\mathfrak{R}$,

and extensions of appropriate sorts – set-theoretical entities on U - to other parameters:

b) an element $c^{\mathfrak{R}} \in$ U to each individual constant *c*,

[165] The set of primitive logical expressions of L may be narrower or wider, depending on what logical basis we choose (other logical expressions may then be defined in terms of primitive ones), and whether we reckon to the logical basis the identity-symbol. If yes, then we have the 1st-order language with identity, if, no, it is a 1st-order language without identity, though the identity-symbol may be included in the set of non-logical constants).

c) for every $n \geq 1$, an n-place relation $P^{\mathfrak{R}} \subseteq U^n$ to each n-place predicate P,

d) for every $n \geq 1$, an n-place operation $f^{\mathfrak{R}}: U^n \rightarrow U$, to each n-place function symbol f.

In effect, this is a set-theoretic interpretation of L in $\mathfrak{R}$ that can be represented as an ordered pair $\langle U, I \rangle$, I being an interpretation-function that accomplishes the same job as our informal account above. The same effect can be achieved in a different way. Let us first order all non-logical constants of L in a sequence with no repeating terms (the super-scripts indicate the number of places):

$$\langle c_1, c_2,..., {}^1P_1, {}^1P_2,..., {}^nP_1, {}^nP_2,..., {}^1f_1, {}^1f_2,..., {}^nf_1, {}^nf_2,...\rangle.$$

Second, let us imagine a sequence of set-theoretical entities over some fixed and non-empty U (with no repeating terms, again)

$$\langle c_1^{\mathfrak{R}}, c_2^{\mathfrak{R}},..., {}^1P_1^{\mathfrak{R}}, {}^1P_2^{\mathfrak{R}},..., {}^nP_1^{\mathfrak{R}}, {}^nP_2^{\mathfrak{R}},..., {}^1f_1^{\mathfrak{R}}, {}^1f_2^{\mathfrak{R}},..., {}^nf_1^{\mathfrak{R}}, {}^nf_2^{\mathfrak{R}},...\rangle,$$

the sequence being such that each item occurring at a certain place in this sequence is appropriate to the logical character of the term occurring at the corresponding place in the first sequence: i.e. if it is an individual constant $\boldsymbol{c_i}$, then $c_i^{\mathfrak{R}}$ is an individual of U, if it is a n-place predicate ${}^n\boldsymbol{P_i}$, then ${}^nP_i^{\mathfrak{R}} \subseteq U^n$, and so on). The idea is put concisely in the following passage:[166]

> "We assume that all the non-logical constants of [formalized theory] T have been arranged in a (finite or infinite) sequence $<\boldsymbol{C_0},...,\boldsymbol{C_n},...>$, without repeating terms. We consider systems $\mathfrak{R}$ formed by a non-empty set U and by a sequence $<C_0,..., C_n,...>$ of certain mathematical entities, with the same number of terms as the sequence of non-logical constants. The mathematical nature of each C_n depends on the logical character of the corresponding constant $\boldsymbol{C_n}$. Thus, if $\boldsymbol{C_n}$ is a unary predicate, then C_n is a subset of U; more generally, if $\boldsymbol{C_n}$ is an m-ary predicate, then C_n is an m-ary relation the field of which is a subset of U. If $\boldsymbol{C_n}$ is an m-ary operation symbol, $\boldsymbol{C_n}$ is an m-ary operation (function of m arguments) defined over arbitrary ordered m-tuples $<x_1,...,x_n>$ of elements of U and assuming elements of U as values. If, finally, $\boldsymbol{C_n}$ is an individual constant, C_n is simply an element of U. Such a system (sequence) $\mathfrak{R} = <U, C_n,..., C_n,...>$ is called a possible realization or simply a realization of T; the set U is called the universe of $\mathfrak{R}$." (Tarski et al 1953: 8)

By appropriately ordering L's non-logical constants and set-theoretical entities over U, we obtain a unique correlation between the terms of the first sequence and the terms of the second sequence

---

[166] The only difference is that above I have opted for the notation more usual today, while Tarski *et al* differentiate linguistic symbols from set-theoretical objects by using the bold Latin for the first and normal Latin for the second, and do not use different types of letter to differentiate different types of non-logical symbols in the signature of L.

$$\langle c_1,\ c_2, \ldots,\ P_1,\ P_2, \ldots,\ f_1,\ f_2, \ldots\rangle$$
$$\langle c_1^{\Re},\ c_2^{\Re}, \ldots,\ {}^nP_1^{\Re},\ {}^nP_2^{\Re}, \ldots,\ {}^nf_1^{\Re},\ {}^nf_2^{\Re}, \ldots\rangle,$$

which effects a set-theoretical interpretation of L in the structure

$$\langle \mathrm{U}, c_1^{\Re},\ c_2^{\Re}, \ldots,\ {}^nP_1^{\Re},\ {}^nP_2^{\Re}, \ldots,\ {}^nf_1^{\Re},\ {}^nf_2^{\Re}, \ldots\rangle.$$

In this way, Tarski et al specified an interpretation for the language without ever using the very (semantic) notion of interpretation!

5.3 Satisfaction and truth in a structure.

Having the background in place, we can explain how the relative notions of truth and satisfaction are defined. For this purpose we consider a standard first-order language (L) with the following logical basis $\{\forall;\ \neg\ ;\ \wedge;\ =\}$, similar in its structure to the simple quantificational language L3, except that L is lexically and syntactically more complex, containing a denumerable (possibly empty) stock of: individual constants, *n*-place predicates and *n*-place functors (for $n \geq 1$). Instead of sentential functions of an interpreted language, we shall speak of formulas of an uninterpreted L. Addition of new lexical categories of non-logical constants requires us to formulate the syntax of L anew so as to characterize recursively the sets of L-terms and L-formulas. What follows is a quite standard recipe:

***t* is a term of L** iff *t* belongs to the smallest set such that:

> (a) every variable is a term; (b) every individual constant is a term; (c) every expression of the form $\xi(t_1,\ldots,t_n)$ is a term, if $t_1,\ldots,t_n$ are terms and ξ is an *n*–place function symbol.

***f* is an atomic formula of L** iff *f* belongs to the smallest set such that:

> (a) every expression of the form $t_i=t_k$ is an atomic formula, if t_i and t_k are
> terms; (b) every expression of the form if $P(t_1,\ldots,t_n)$ is an atomic formula,
> if $t_1,\ldots,t_n$ are terms and *P* is an *n*–place predicate.

***f* is a formula of L** iff *f* belongs to the smallest set such that:

> (a) every atomic formula is a formula; (b) every expression of the form $\neg A$ is a formula, if *A* is a formula; (c) every expression of the form $A \wedge B$ is a formula, if *A* and *B* are both formulas; (d) every expression of the form $\forall v_i\, A$ is a formula, if *A* is a formula and *i* is a positive integer.

The variable v_i is free in the formula *A* iff one of the following conditions is satisfied:

> (a) *A* is an atomic formula and v_i occurs in *A* as a term; (b) *A* is of the form $\neg B$, for a formula *B*, and v_i is free in *B*; (c) *A* is of the form $B \wedge C$, for some formulas *B* and *C*, and v_i is free in *B* or v_i is free in *C*;

(d) A is of the form $\forall v_k B$, for a formula B and positive integer k, and $i \neq k$, and v_i us free in B.

***S* is a sentence of L** iff

S is a formula of L that contains no free variables.

Having the syntax of L in place, we can define the model-theoretic semantics for L. Our previous truth-definition for quantificational language shall guide us, but we should now define not just satisfaction-conditions of (simple and complex) sentential functions w.r.t. an infinite sequence *s* of objects, but, rather, satisfaction-conditions of formulas of L, as interpreted in the model-structure $\mathfrak{R}$, w.r.t. infinite sequences of individuals from $\mathfrak{R}$ (belonging to U). The two kinds of definitions are quite similar in that both proceed by recursion on the logical complexity of formulas and sentential functions respectively. We need to define recursively also the notion of denotation. And, for the sake of uniformity, this definition will be relativized to sequences as well.

The denotation (value) of a term *t* in $\mathfrak{R}$ with respect to the infinite sequence *s* – shortly: $\mathbf{D}(t)_{\mathfrak{R}s}$ - is defined recursively as follows:[167]

a) $D(t)_{\mathfrak{R}s} = s_i$ if t is the i-th variable

b) $D(t)_{\mathfrak{R}s} = c_i^{\mathfrak{R}}$ if $t = c_i$

c) $D(t)_{\mathfrak{R}s} = \zeta^{\mathfrak{R}}(D(t_1)_{\mathfrak{R}s},...,D(t_n)_{\mathfrak{R}s})$ if t is of the form $\zeta(t_1,...,t_n)$, where $t_1,..,\ t_n$ are terms.

Satisfaction of a formula *f* in $\mathfrak{R}$ w.r.t. *s* – shortly: $\mathfrak{R} \models_s f$ - is defined recursively as follows:

a) If f is of the form $t_i=t_k$, where $t_i,\ t_k$ are terms:

$\mathfrak{R} \models_s t_i=t_k$ iff $D(t_i)_{\mathfrak{R}s} = D(t_k)_{\mathfrak{R}s}$

b) If f is of the form $P(t_1,...,t_n)$, where $t_1,\ t_n$ are terms:

$\mathfrak{R} \models_s P(t_1,...,t_n)$ iff $\langle D(t_1)_{\mathfrak{R}s},..., D(t_n)_{\mathfrak{R}s}\rangle \in P^{\mathfrak{R}}$

c) If f is of the form $\neg A$, where A is a formula:

$\mathfrak{R} \models_s \neg A$ iff it is not the case that $\mathfrak{R} \models_s A$

d) If f is of the form $A \wedge B$, where A and B are formulas:

$\mathfrak{R} \models_s A \wedge B$ iff $\mathfrak{R} \models_s A$ and $\mathfrak{R} \models_s B$

---

[167] Instead of *denotation* for terms of L an assignment function s^* is sometimes recursively defined, the idea being that while variables and individual constants of L are assigned values by s and $\mathfrak{R}$ respectively, s^* extends s and $\mathfrak{R}$ also to complex terms of L (cf. Mendelson (1997)). The effect is the same; in fact, $D(x)_{\mathfrak{R}s}$ is the same function as s^*.

e) If f is of the form $\forall v_k A$, where A is a formula and k a positive integer:

$$\mathfrak{R} \models_s \forall v_k A \text{ iff } \mathfrak{R} \models_{s^*} A, \text{ for every } k\text{-variant } s^* \text{ of } s.$$

Truth of a sentence A in $\mathfrak{R}$ - shortly: $\mathfrak{R} \models A$ - is defined directly as follows:

A is true in $\mathfrak{R}$ iff A is a sentence of L and $\mathfrak{R} \models_s A$, for every s in $\mathfrak{R}$.[168]

Once we have this model-theoretic relative definition of truth in place, we can define various model-theoretic notions, including the classic notion of *logical consequence*:

Satisfiability and unsatisfiability in $\mathfrak{R}$:

- A formula A (of L) is satisfiable in $\mathfrak{R}$ iff there is a sequence s in $\mathfrak{R}$ such that $\mathfrak{R} \models_s A$; otherwise A is unsatisfiable in $\mathfrak{R}$.

- A theory T of formulas (of L) is satisfiable in $\mathfrak{R}$ iff there is a sequence s in $\mathfrak{R}$ such that $\mathfrak{R} \models_s A$, for each formula A belonging to T; otherwise T is unsatisfiable in $\mathfrak{R}$.

Model

- $\mathfrak{R}$ is a model of a sentence A (of L) iff $\mathfrak{R} \models A$.

- $\mathfrak{R}$ is a model of a theory T of sentences (of L) iff $\mathfrak{R} \models A$, for each sentence A belonging to T.

Satisfiability (**semantic consistency**):

- A formula A (of L) is satisfiable iff there is a structure $\mathfrak{R}$ (of L) in which A is satisfiable.

- A sentence A (of L) is satisfiable iff there is a structure $\mathfrak{R}$ (of L) that is a model of A.

- Theory T of formulas (of L) is satisfiable iff there is a structure $\mathfrak{R}$ (of L) in which T is satisfiable.

- Theory T of sentences (of L) is satisfiable iff there is a structure $\mathfrak{R}$ (of L) that is a model of T.

Validity (semantic)

- Formula A (of L) is valid iff for every structure $\mathfrak{R}$ (of L) and every sequence s in $\mathfrak{R}$ it holds that $\mathfrak{R} \models_s A$.

- Sentence A (of L) is valid iff every structure $\mathfrak{R}$ (of L) is a model of A.

---

[168] In case of the absolute definition of truth, we can equally say that:
A is true in $\mathfrak{R}$ ($\mathfrak{R} \models A$) iff A is a sentence of L and $\mathfrak{R} \models_s A$, for <u>some</u> s in $\mathfrak{R}$.

Consequence (semantic):

- Formula *A* (of L) is a logical consequence of a set of formulas *T* iff for every structure $\mathfrak{R}$ and every sequence *s* in $\mathfrak{R}$ it holds: if $\mathfrak{R} \models_s B$, for every formula *B* in T, then $\mathfrak{R} \models_s A$

- Sentence *A* (of L) is a logical consequence of a set of sentences *T* iff every structure $\mathfrak{R}$ (of L) that is a model of *T* is a model of *A*.

The definition of $\mathfrak{R} \models \boldsymbol{A}$ in terms of $\mathbf{D}(t)_{\mathfrak{R}s}$ and $\mathfrak{R} \models_s \boldsymbol{f}$ is very much in the spirit of Tarski's absolute truth-definitions, because it uses recursion on the complexity of formulas, making use of infinite sequences (alternatively: assignment functions), and because it *mathematizes* semantics. Both satisfaction and denotation conditions of formulas and terms respectively reduce in last instance to interpretations of L-primitives in $\mathfrak{R}$ plus assignments of values to free variables. Both interpretations and assignments are construed or modelled as sets of order pairs: i.e. functions whose domains are certain sets of expressions (variables, individual constants, 1-place predicates,....,*n*-place predicates) and whose co-domains are certain sets of appropriate set-theoretical objects defined over U. In this sense, then, both interpretations and assignments (valuations) are rendered purely mathematical, being set-theoretical in character. Recall that in the truth-definition for an interpreted language there is a non-semantic pairing between variables ordered in the sequence $\langle x_1,..., x_n,...\rangle$ and terms of the sequence $\langle s_1,..., s_n,...\rangle$ achieved simply via correlating their numerical indexes,

$$\begin{array}{l}\langle x_1, x_2,..., x_n,...\rangle \\ \quad\blacktriangledown\quad\blacktriangledown\quad\blacktriangledown \\ \langle s_1,\ s_2,..., s_n...\rangle\end{array}$$

Now, an interpretation of an uninterpreted L in a model-structure $\mathfrak{R}$

$$\begin{array}{l}\langle \boldsymbol{C_0},...,\boldsymbol{C_n},...\rangle \\ \quad\blacktriangledown\quad\blacktriangledown \\ \langle C_0,...,C_n,...\rangle\end{array}$$

seems to be a similar business: pairing of non-logical constants of an uninterpreted L and items of the structure $\mathfrak{R}$ (namely set-theoretical entities defined over the domain U or $\mathfrak{R}$). This is the vital point behind the claims to the effect that Tarski showed how to interpret semantics in set-theory, as Gödel showed how to interpret syntax (broadly construed) in arithmetic. It does not follow from this, in my opinion, that semantic and syntactic phenomena are mathematical in nature. It only follows that we can approach them – to some extent - by using rigorous mathematical methods.

5.4 The framework applied: truth in the standard model of L(PA)

The definition offered above is intentionally schematic so as to cover a wide range of 1st-order languages with the same structure, and their interpretations in suitable structures. Truth has been defined, in a way, for a variable $\mathfrak{R}$. One consequence of this is that we can no longer apply the adequacy criterion spelled out in Convention T. But, once particular choices are made for L and $\mathfrak{R}$ it is

possible to define the notion of true sentences of the specific language relative to the specific structure in such a way that we can derive from the definition analogues of the material-adequacy conditions for absolute truth of the form:

$\mathfrak{R} \models A$ (in L) iff p

where *A* is a sentence of L and what replaces 'p' at the right side is a sentence of ML that says the same as *A* under its interpretation in $\mathfrak{R}$. Thus let L be specified as the first-order language of Peano arithmetic - L(PA) – that has the same set of logical constants as before (that is: with identity), and its non-logical constants (signature) are specified by the following sequence:

$\langle \mathbf{0}, \mathbf{S}, +, \times \rangle$,[169]

where the first term is individual constant, the second is a 1-place functional symbol, and the third and fourth terms are two-place function-symbols. We then specify $\mathfrak{R}$ as *N* (called the standard model of arithmetic), the sequence consisting of the domain N of natural numbers together with certain distinguished elements, and *n*-place functions on the domain N:[170]

$N = \langle \mathrm{N}, \mathbf{0}^N, \mathbf{S}^N, +^N, \times^N \rangle$,

More simply, we can represent that structure using the same letters as in the signature, but printed on normal style: ⟨N, 0, S, +, ×⟩. We thereby obtain - in essentially Tarski's pairing fashion - the following set-theoretical interpretation of $\mathrm{L_{PA}}$ in *N*:

Interpretation of $\mathbf{L_{PA}}$

The universal quantifier ∀ of L is assigned as its domain the set N = {0,1, 2,…}

$\mathbf{0}^N = 0$

$\mathbf{S}^N$ { S: N → N, S(*x*) = the successor of *x*;

{⟨*x*, *y*⟩: *x*, *y*∈ N and *y* = *x*+1}

$+^N$ { +: N^2 → N, +(*x*, *y*) = the sum of *x* and *y*;

{⟨*x*, *y*, *z*⟩: *x*, *y*, *z* ∈ N and *z* = *x* + *y*}

[169] If, instead, we treated '=' as a non-logical constant, we could include it into the signature and we would have to assign it an interpretation in a structure, say in the structure $\mathfrak{R} = \langle N, \mathbf{0}^N, \mathbf{S}^N, +^N, \times^N, =^N \rangle$: $=^{\mathfrak{R}}$ would then be the equality relation on N: {⟨*x*, *x*⟩: *x*∈ N}.

[170] Note that the signature of L(PA) contains no relation symbol. Such signatures are sometimes called *algebraic*, whereas signatures that do not contain individual constants or function-symbols, but only relation symbols, are called relations signatures, and structures appropriate to such signatures are called relational structures (Hodges, 1997: 5). Note, however, that model-structures of first-order languages (signatures) used to be called, without difference, relational systems or structures (cf. Henkin's (1967) lucid introductory entry on model theory).

$$\times^{N} \begin{cases} \times: \mathrm{N}^2 \rightarrow \mathrm{N},\ \times(x, y) = \text{the product of } x \text{ and } y; \\ \{\langle x, y, z\rangle: x, y, z \in \mathrm{N} \text{ and } z = x \times y\} \end{cases}$$

So interpreted, L(PA) can deal with natural numbers and certain elementary properties of them and is accordingly called the language of elementary arithmetic. Since in L(PA) each element of N is designated by a numeral ***n*** of the type **S**(**S**(...(**0**)...), we can simplify the definition of truth for this language, doing entirely without the notion of satisfaction w.r.t. to an infinite sequence (of numbers from N). We first define syntax of L(PA) in the following way (given the signature and infinite set of variables of L(PA):

***x* is a term of L(PA)** iff

> *x* belongs to the smallest set that contains all individual constants and variables and is closed under the syntactic operations corresponding respectively to **S**, + and ×.[171]

***x* is a closed term of L(PA)** iff

> *x* belongs to the subset of the set of terms that contains all individual constants and is closed under the syntactic operations corresponding respectively to **S**, + and ×.[172]

***x* is an atomic formula of L(PA)** iff

> *x* belongs to the set of all and only those expressions that have the form $t_i = t_k$, where t_i, t_k are terms;

***x* is an atomic sentence of L(PA)** iff

> *x* belongs to the subset of the set of atomic formulas to which all and only those expressions belong that have the form $t_i = t_k$, where t_i, t_k are both closed terms;

***x* is a formula of L(PA)** iff

> *x* belongs to the smallest set that contains all atomic formulas and is closed under syntactic operations corresponding respectively to ¬, ∧ and ∀.[173]

The variable v_i is free in a formula *A* iff

> (a) *A* is an atomic formula and v_i occurs in *A* as a term, or (b) *A* is of the form ¬*B*, for a formula *B*, and v_i is free in *B*, or (c) *A* is of the form $B \wedge C$, for some formulas *B* and *C*, and v_i is free in *B* or v_i is free in *C*, or (d) *A* is of the form $\forall v_k B$, for a formula *B* and positive integer *k*, and $i \neq k$ and v_i occurs free in *B*

---

[171] Strictly speaking, we should define for each *n*-place function symbol *f*, an *n*-place term-building operation F_f on expressions as follows:

$F_f(t_1,..., t_n) = f\, t_1,..., t_n$.

and then define the set of formulas as the set of expressions that can be built up from the individual constants and variables by applying (zero or more times) the F_f operations (viz. e.g. Enderton (2001)).

[172] The same point applies as in the previous note.

[173] The same point applies, once again.

***x* is a sentence of L(PA)** iff

> *x* belongs to the subset of the set of formulas to which such and only such formulas belong in which there is no free variable.

On this basis we can define truth directly for atomic sentences of L(PA) in terms of denotation of closed terms of L(PA), truth of truth-functional compounds in terms of truth of its component sentences, and truth of universally quantified sentences in terms of truth of all numerical instances of formulas being quantified:

(I) $\mathbf{D}(t)_N$ is defined recursively as follows:

(a) $D(t)_N = \mathbf{0}^N$, if $t = \mathbf{0}$;

(b) $D(t)_N = \xi^N(D(t_1)_N,...,D(t_n)_N)$, if t is of the form $\xi(t_1,..., t_n)$, where ξ is an n-place function-symbol and $t_1,..., t_n$ are terms.

(II) $N \models \varphi$ is defined recursively as follows:

(a) if φ is atomic sentence of the form $(t_i = t_k)$, where t_i and t_k are closed terms:

$N \models (t_i = t_k)$ iff $D(t_i)_N = D(t_k)_N$,

(b) If φ is a sentence of the form $\neg A$, where A is a formula:

$N \models \neg A$ iff it is not the case that $N \models A$;

(c) If φ is a sentence of the form $A \wedge B$, where A and B are formulas:

$N \models A \wedge B$ iff $N \models A$ and $N \models B$;

(d) If φ is a sentence of the form $\forall v_i A$, where A is a formula and k a positive integer:

$N \models \forall v_i A$ iff $N \models A(\underline{\mathbf{n}})$, for every n.[174]

Having these basic definitions in place, we might want to check whether the definition gives indeed correct predictions for sentences of L(PA). For instance, let us ask under what conditions the following statement holds:

i) $N \models +(\mathbf{0}, \mathbf{S}(\mathbf{0})) = \mathbf{S}(\mathbf{0})$

We see that the sentence is of the form $t_i = t_k$, hence the rule (IIa) applies:

---

[174] It is a well know technical result that although the recursive definition of denotation can be turned to a purely arithmetical one within PA itself, this cannot be done with the recursive definition of truth, because of the last clause to the effect that the truth-value of a universally quantified sentence depends on the truth-values of infinitely many sentences. The whole definition, however, can be converted to a fully explicit one within 1st order set theory or 2nd-order arithmetic.

ii) $N \models +(\mathbf{0}, \mathbf{S}(\mathbf{0})) = \mathbf{S}(\mathbf{0})$ iff $D(+(\mathbf{0}, \mathbf{S}(\mathbf{0})))_N = D(\mathbf{S}(\mathbf{0}))_N$

Since the first term on the right side is of the form $\xi(t_1,..., t_n)$, for $n = 2$, we can apply the rule (Ib):

iii) $N \models +(\mathbf{0}, \mathbf{S}(\mathbf{0})) = \mathbf{S}(\mathbf{0})$ iff $+^N(D(\mathbf{0})_N, D(\mathbf{S}(\mathbf{0}))_N) = D(\mathbf{S}(\mathbf{0}))_N$

We further apply two times the rule (Ib), since D_N appears twice on the right side attached to '$\mathbf{S}(\mathbf{0})$', which is of the form $\xi(t_1,..., t_n)$, for $n = 1$:

iv) $N \models +(\mathbf{0}, \mathbf{S}(\mathbf{0})) = \mathbf{S}(\mathbf{0})$ iff $+^N(D(\mathbf{0})_N, \mathbf{S}^N(D(\mathbf{0})_N)) = \mathbf{S}^N(D(\mathbf{0})_N)$

At this point, D_N and N are attached only to primitive non-logical constants of L(PA). So, we can first eliminate certain occurrences of D_N in favour of N, in accordance with the rule (Ia):

v) $N \models +(\mathbf{0}, \mathbf{S}(\mathbf{0})) = \mathbf{S}(\mathbf{0})$ iff $+^N(\mathbf{0}^N, \mathbf{S}^N(\mathbf{0}^N)) = \mathbf{S}^N(\mathbf{0}^N)$

Next we eliminate N in accordance with the way it was recursively specified. By employing first our knowledge about $\mathbf{0}^N$ and then about $\mathbf{S}^N$, we obtain (bold-printed occurrences of zero-sign and successor-sign being replaced by normal-printed):

vi) $N \models +(\mathbf{0}, \mathbf{S}(\mathbf{0})) = \mathbf{S}(\mathbf{0})$ iff $+^N(0, \mathbf{S}^N(0)) = \mathbf{S}^N(0)$

and

vii) $N \models +(\mathbf{0}, \mathbf{S}(\mathbf{0})) = \mathbf{S}(\mathbf{0})$ iff $+^N[0, S(0)] = S(0)$.

Finally, we exploit our knowledge of what $+^N$ amounts to in order to eliminate the very last occurrence of N on the right side so that we get

viii) $N \models +(\mathbf{0}, \mathbf{S}(\mathbf{0})) = \mathbf{S}(\mathbf{0})$ iff $(0 + S(0)) = S(0)$,

which is a desired result in keeping with our intended truth-conditions for the sentence in question.

By means of such transformations we could derive for each sentence A of L(PA) an analogous equivalence stating its condition of truth relative to its interpretation in the structure N. Such equivalences are analogues of T-biconditionals for fully interpreted sentences of formalized languages. In view of this, it may appear natural to think of formalized languages as languages with built-in interpretations in intended structures, and that, for this reason, they admit of absolute truth-definitions. Indeed, some thinkers are tempted to interpret CTFL as follows: since Tarski had in mind languages talking about and describing certain intended structures, there was no need for him to relativize truth-definitions explicitly to an additional structure-parameter.

I agree that there is not much formal difference between a relative definition of truth w.r.t. N for L(PA), qua uninterpreted language, and an absolute truth definition for L(PA), qua interpreted language of elementary arithmetic. For any given sentence A of the interpreted L(PA) we can derive the T-biconditional:

A is true iff *p*

the right side translating *A*. By analogy, for any given sentence *A* of uninterpreted L(PA) we can derive

$N \models A$ iff *p*

the right side giving the intuitive reading of *A* as interpreted in *N*. From the perspective common in contemporary mathematical logic, specifying a particular interpretation for a given language L does the same job that was done in CTFL by translation of L into meta-L:

'**0**' of L(PA) reads 'zero' in our ML

'**S**(*x*)' of L(PA) reads 'the successor of (*x*)' in our ML

etc.

I take it that this was Tarski's mature take on mathematical logic from 1950s on. It rhymes well with his life-long program of mathematizing metamathematics and semantics in particular by means of function- and set-theoretic procedures such as the method of pairing variables with the terms of sequences via numerical indexes, or the method of converting inductive definitions into explicitly set-theoretical ones. It is to be noted that model-theoretic definitions of relative semantic notions can be turned into purely set-theoretical ones, because the domain of an L-structure is bound to be a set, and satisfaction can therefore be defined in a more powerful metatheory, whose quantifiers range over all sequences on and subsets of the domain, hence, in the standard set theory, whose intended domain comprises every set. Of course, we have seen that the problem arises as to what, if anything is the standard model of set theory, given that its domain is not a set but a proper class. One alternative is to consider set theory as the ultimate framework of mathematics whose semantics cannot be model-theoretically defined in a yet more powerful language. In keeping with what was said about Tarski hierarchies, one can define truth for any sub-language of set theory all of whose variables are bounded above by a definite order (this is a version of Tarski's Theorem II, which was formulated for the language of the general calculus of classes). Eventually, one can axiomatize the implicit semantics of the language of set-theory – taking appropriate measures to avoid paradoxes - for which procedure one does not need an essentially stronger metatheory. Another alternative, which we also discussed in connection with Tarski's indefinability results, is to extend the notion of model so that domains can be proper classes, and not just sets, and then define set theoretic semantics - including truth in set theory - in a more powerful theory of super-cool entities such as classes of Morse-Kelley set theory.

This, however, does not mean that this was Tarski's – or anybody's - considered view in the 1930s. I shall now turn my attention to the question to what extent was the model-theoretic approach anticipated in CTFL.

5.5 Model-theoretic definitions and CTFL

The view was once fairly widely held, and there are still many traces of it to be found in the literature, that in CTFL and related papers – especially, the 1936 article defining logical consequence in terms of models – Tarski took a giant step towards generalizing his method of truth definition to uninterpreted formal languages. In fact, there is a grain of truth in this belief, if properly understood. It is that the definition of relative truth appears to be an easy technical extension of the definition of absolute truth: while a sentence of a formal L does not have any definite meaning and so is not true (false) outright, it comes out true or false depending on this or that admissible interpretation of its non-logical expressions, their logical meanings being sensitive to the fixed interpretations of L's logical constants. That dependence is then directly reflected in the definition of satisfaction by introducing at appropriate points of it – i.e. in its base and recursion clauses - a parameter for a structure, in which L's expressions are supposed to be interpreted. What we add is an additional parameter relative to which satisfaction of sentential functions - this time dubbed 'formulas' - is determined. It seems accurate to say, with Burgess, that the step from an absolute truth-definition for a set-theoretically interpreted L to a relative, model-theoretic definition for an uninterpreted language is an extremely short one:

> "An interpreted language is naturally thought of as an ordered pair consisting of an *un*interpreted language and an interpretation. And an interpretation is simply a set, the domain, and an assignment of a relation or operation of the right number of places on it to each non-logical primitive. But that is essentially what a mathematical structure is: a set, the domain, and certain distinguished relations and/or operations on it, distinguished from each other by certain symbols associated with them....And formally, the step from a two-place relation between a sentence and an ordered pair consisting of an uninterpreted language plus an interpretation or structure to a three-place relation among a sentence, an uninterpreted language, and a structure or interpretation is a very short one." (Burgess 2008b: 155).

This is how many theorists are inclined to view the matter today, when the model theoretic approach dominates mathematical logic and to a considerable extent formal semantics. It was suggested by the end of the previous section that we can treat an interpreted L as if it had a built-in intended interpretation so that we can represent it by the ordered pair $\langle L, I_{\mathfrak{R}}\rangle$: L as interpreted in the structure $\mathfrak{R}$. Once we have isolated the intended interpretation $I_{\mathfrak{R}}$ of L, the next step is to treat the second term of the ordered pair as a variable that can receive different values. We thus arrive at the idea of reinterpreting L in various admissible ways - talking about and describing different structures than it actually does under its intended or standard interpretation $I_{\mathfrak{R}}$. This forces us to think of L as an uninterpreted language to be interpreted in different ways, in different set-theoretical structures of its type. In order to define truth for such L, we then need to relativize the definition to a structure $\mathfrak{R}$, defining what it means for formal L-sentences to be true as interpreted in $\mathfrak{R}$. When we see the matter in this light, the absolute truth definition for $\langle L, I_{\mathfrak{R}}\rangle$ can be deemed a special case of the general definition of truth in a structure, for the intended interpretation of L

in its standard structure.

One then easily gets the reassuring feeling that the model-theoretic method of truth definition is a natural extension or modification of the method of absolute truth definition, and that it therefore must be present somewhere in CTFL or related articles, and, along with it, the definitions of the related notions of *validity*, *satisfiability* and *logical consequence*, which we are used to contrast with their proof-theoretical counterparts: *theorem*, *consistency*, *derivability*. In his penetrating analysis, Hodges reports that he once had that reassuring feeling himself. Yet, to his surprise, it proved definitely misleading:

> "... a few years ago I had a disconcerting experience. I read Tarski's famous monograph 'The concept of truth in formalized languages' (1935) to see what he says himself about the notion of truth in a structure. The notion was simply not there. This seemed curious, so I looked in other papers of Tarski. As far as I could discover, the notion first appears in Tarski's address (1952) to the 1950 International Congress of Mathematicians, and his chapter 'Contributions to the theory of models I'(1954). But even in those papers he doesn't define it. In the first chapter he mentions the notion only in order to explain that he won't be needing it for the purpose in hand. In the second chapter he simply says "We assume it to be clear under what conditions a sentence ... is satisfied in a system..."" (Hodges 1985/86: 137)

Let it be said that Hodges does not deny that CTFL contains quite a few seminal ideas that shaped the standard model-theoretic approach as we know it today. Indeed, the recursive definition of satisfaction is the basis of that approach, and it was first given in a fully rigorous manner in CTLF. But he points out that no standard definitions of truth and satisfaction in a structure are to be found in CTFL, these being provided only in the joint work of Tarski and Vaught.[175] In fact, the definitions from the mid 1950s had their predecessors in Tarski' own earlier writings from the early 1950s, and in the important articles of Kemeny and Henkin,[176] influenced also by Carnap's pioneering work in semantics that I have already reviewed.[177] But Hodges seems to be right that prior to the 1940s no precise model-theoretic definitions of truth and satisfaction were available, even though Tarski had almost all essential ingredients at his disposal needed to formulate them: the notion of sequence and the recursive method of definition of satisfaction of a formula by a sequence of individuals from the domain associated with quantifiers. Yet until the early 1950s something prevented him from providing the standard model-theoretic definitions and Hodges inquiries what was responsible for that remarkable delay, especially when we bear in mind that the idea of structure was in the air since the 19th century debates about non-Euclidean geometries.

What Hodges reminds us of is that getting an accurate picture of the connections between the semantics of CTFL and the model-theoretic semantics current today may be a surprisingly delicate matter. It seems that the decisive shift

---

[175] Tarski & Vaught (1956).
[176] Kemeny (1949) and Henkin (1949).
[177] Carnap (1938), (1942), (1947).

of Tarski's perspective towards full-blooded model theory came only in the 1950s. This appears remarkably late, given that notion was implicit in the informal 'model-theoretic' work in geometry and algebra since the second half of the 19th century. Observing that the common word for mathematical structure in the first three decades of 20th century was "system" (in German: 'System von Dingen'), he traces the historical trajectory of the idea of *truth in a structure*, or rather, its predecessors:

> "Tarski's 1933 paper brought into focus a number of ideas that were in circulation earlier. The notion of an assignment satisfying a formula is implicit in George Peacock ([151], 1834) and explicit in George Boole ([25] p. 3, 1847), though without a precise notion of 'formula' in either case. The word 'satisfy' in this context may be due to Edward V. Huntington (for example in [97] 1902). Geometers had spoken of gypsum or paper 'models' of geometrical axioms since the 17th century; abstract 'models' appeared during the 1920s in writings of the Hilbert school (von Neumann [147]1925, Fraenkel [59] p. 342, 1928)." (Hodges 20??: 2).

The talk about *Systeme von Dingen* and interpretations was part and parcel of theoretical mathematics already in the 19th century. As a matter of fact, geometry was thought of by its practitioners as being concerned with space (or with a class of spaces), which is a kind of system or structure, by today's lights. Today, it is a well known fact that new axiomatizations of non-Euclidean systems were motivated by the discovery of independence of Euclid's parallel postulate on the remaining Euclidean postulates, which was demonstrated via exhibiting interpretations in which all but the parallel postulate are true or satisfied. In a related vein, consistency of axiomatic systems of non-Euclidean geometries was established via exhibiting their verifying interpretations in Euclidean geometry or some other system; today one would say: via exhibiting their models. Indeed, *systems* or *interpretations* were common currency in German axiomatic tradition represented by Dedekind (viz. his characterization of the structure of natural numbers as determined "up to isomorphism" by his 2nd-order axioms), by Moritz Pasch (his work in geometry), and, above all, by David Hilbert, who conducted model-theoretic inquiries into geometry and analysis. What can be more telling than the statement made in Hilbert's address to the 1900 mathematical congress, which he says that consistency of an axiomatic system of geometry can be demonstrated by:

> "....constructing an appropriate domain of number such that to the geometrical axioms correspond analogous relations among the objects of this domain." (Hilbert 1900: 1104) ?

In his path-breaking work on the axiomatic foundations of geometry[178] Hilbert not only showed the independence of each of his axioms on the remaining axioms by exhibiting an interpretation verifying all the axioms except the one to be showed independent, but he showed its relative consistency by exhibiting an interpretation of the axiom system within the arithmetic of real numbers, by choosing a suitable domain of algebraic numbers along with certain relations on it

[178] Hilbert (1899).

corresponding to the relational-notions of the system. In fact, his followers talked about abstract 'models', under the label of 'Systeme' even earlier than Hodges indicates. As early as 1910, Weyl talks about one and the same statement being true with respect to different systems, and he does so in the course of discussing the notion of isomorphic systems.[179] Much the same can then be said of American postulate theorists, Veblen and Huntington,[180] who analysed the notions of categoricity of an axiom system and of isomorphic systems, and of Peano's axiomatic school, where Padoa and Pieri came with seminal contributions. Besides Hilbert, the idea of an uninterpreted non-logical constant is most visible in their work. Recall Padoa's conception of implicit definability (dependence) of a notion in terms of others and the method of showing irreducibility (independence) of the system of basic non-logical notions of an axiomatic system, based on the idea of considering different (re)-interpretations of such notions.[181]

Being an expert on axiomatizations in topology, geometry, analysis or algebra, Tarski could not have been unaware of these developments. The question then arises as to why he did not define in CTFL the idea of truth or satisfaction in a structure (system, interpretation)? Hodges diagnosis is that in the 1930s he was still held captive by Frege-Peano conception of logical languages as meaningful formalisms – properly regimented fragments of natural languages for the exact purposes of science whose expressions are divided into the formal-logical part and the non-logical part. He therefore did not have at his disposal the notion of an uninterpreted formal language that gets various interpretations by being attached to various mathematical structures appropriate to its type, which is based on the notion of an uninterpreted non-logical constant as a category to be distinguished from variables and interpreted constants as well as logical constants. Hodges diagnosis seems correct to me, so far as it goes. In the 1930s, and likely in the 1940s, Tarski shared Frege's view - propounded vigorously by his teacher Lesniewski - that logical languages are meaningful languages. In this respect, once again, Tarski was not an adherent of the conception of logic as calculus reinterpretable *ad lib* in structures, though he would strongly oppose the view that there is just one language and the related idea that the distinction of object-language and metalanguage is philosophically utterly misguided. What this shows is that, for philosophical reasons, Tarski was not prepared to take what Burgess would call "a formally very short step" from defining the notion of

A being a true sentence of $\langle L, I_{\mathfrak{R}}\rangle$

to defining the notion of

A being a true sentence of L in $\mathfrak{R}$,

since, in the first place, he was not prepared to think of an uninterpreted language L and it therefore did not occur to him to conceive of an interpreted language on the model of $\langle L, I_{\mathfrak{R}}\rangle$.

This explains why it did not make good sense for Tarski to speak of truth or falsity of a meaningless sentence even when considered in relation to a

[179] Weyl (1910).
[180] Huntington (1904), (1905).
[181] Padoa (1901).

mathematical structure. Hodges does not want to go that far. He says that there is no good evidence that Tarski shared Frege's worries, voiced in his well-known correspondence with Hilbert. For Frege, it made no sense to speak of a formal sentence coming out true when interpreted in a mathematical structure. However, without the point stressed above, Hodges does not seem to have any reasonable explanation as to why Tarski did not embrace the notion of an uninterpreted non-logical constant, especially when, as he himself points out, the notion was widely in use in the 1930s. But if so, what could prevent Tarski from using that notion himself? Our interpretation, on the other hand, takes seriously Tarski's official adherence to the Polish view that truth is absolute in nature. Indeed, the view was dominant in the Lvov-Warsaw school under Twardowski's influence, according to which a language is comprised by determinately meaningful sentences that are either true or false, being so absolutely, and not relatively.

5.6 Logical consequence and truth

John Etchemendy challenged what was an even more popular view, according to which the standard model-theoretic definition of consequence is contained in Tarski's classic article (1936b).[182] Tarski starts his discussion with criticizing syntactic (formalized) accounts of logical consequence (of the sort that he himself used to champion in the early 1930s)[183] by considering an ω-incomplete system of arithmetic. With respect to the system under consideration he says that it case-by-case proves

$P(n)$,

for each given n, but it does not prove the universal generalization

For every natural number x: Px

which, intuitively follows from that infinite collection of premises - the corresponding ω-inference being truth-preserving, yet not derivable within the system. Against the possible suggestion to overcome such limitations by expanding the deductive part by new (structural, hence recursive) inference rules (such as a finitary version of the ω-rule), he objects that Gödel's 1st incompleteness theorem shows that this strategy is hopeless. So long as the original theory is recursively axiomatizable and new inference rules are recursive, Gödel's theorem assures us that the expanded theory is bound to contain undecidable sentences.

Tarski then proposes his own, semantic account of logical consequence in terms of models. Verbally at least, the account appears to be model-theoretic:

> "The sentence X follows logically from the sentences of the class K if and only if every model of the class K is also a model of the sentence X." (Tarski 1936a: 417).

[182] Hodges (1985/86); Etchemendy (1988) and (1990).
[183] Tarski (1930).

Tarski defines models for sentences of a formalized language in a non-standard way by today's criteria: namely via detour through sentential functions obtained by uniformly replacing their non-logical constants (fully interpreted, it goes without saying, in accord with Hodges's observations) in all their occurrences by variables of appropriate types, it being sentential functions and not the sentences themselves that are strictly speaking satisfied by sequences of entities of appropriate set-theoretic (type-theoretic) sort. Given the sentence X and the class of sentences K, we obtain, via such systematic replacements, the sentential function X* and the class of functions K*. Arbitrary sequence of objects of types appropriate to free variables of X* that satisfies X* is called a model (or realization) of the sentence X; much the same can be said, *mutatis mutandis*, of K and K*: an arbitrary sequence of objects of appropriate types that satisfies each sentential function in the class K* is a model of the class of sentences K. With these definitions of models, Tarski hopes to explicate the informal notion of consequence respecting two traditional intuitions associated with it: (1) logical consequence preserves truth from premises to conclusion, (2) in logical consequence truth is preserved in virtue to the form of premises and conclusion alone, owing nothing to their content. More precisely, logical consequence owes something to the contents of premises and conclusion, but it is only the contents of formal-logical expressions that *do* matter.

The first condition is traditionally spelled out as follows: it *can* never happen that premises of a valid inference are jointly true while the conclusion is false. But Tarski preferred to avoid the modality, opting rather for something along the following lines:

> X is logically follows from K iff X is true whenever all the sentences in K happen to be jointly true.

To be sure, this formulation leaves much to be desired, but the second condition shows us the way how to be more specific on what "whenever" amounts to. It distinguishes formal consequence from material consequences. The first attempt to flesh it out that Tarski considers is deeply rooted in the logical tradition (Tarski calls it F-condition):

> If we uniformly replace in the sentences of the class K as well as in the sentence X the non-logical constants by any other of appropriate types, the sentence X* obtained from X is true whenever all the sentences of the class K* obtained from K are true. [184]

Tarski hastens to qualify this by saying that there may be no strict dichotomy, unless there is a principled dividing line between formal (logical) expressions and non-formal (non-logical) signs. Bolzano urged a version of this account for logically analytic sentences, talking about variable elements of a sentence (proposition – *Satz an sich*) freely replaceable by elements of the same type.[185] However, its shortcoming (at least in the linguistic form – for Bolzano can appeal to an ideal realm of ideas and propositions *an sich*) is that it relies heavily on the actual richness of the language considered, on whether it contains,

[184] For original formulation see Tarski (1936b: 415).
[185] Bolzano (1837).

say, a name or each individual in its associated universe of discourse.[186] So, according to Tarski, the condition F is in general only a necessary but not a sufficient condition of valid inference. It turns out to be both only in those cases where we have a name for every object in the universe of discourse. In general, though, this is not the case, as when we want to talk about non-denumerable domains. This "substitutional" approach to logical consequence thus suffers from the same kind of defect as substitutional account of quantification.

For several reasons, not all of which are important to us, Etchemendy rejects the received view according to which the model-theoretic account of logical consequence as preservation of truth under all admissible interpretations in structures amounts to the same as the 1936-definition of Tarski in terms of "models", as these were conceived of by him around that time. For one thing, Etchemendy says, there is no mentioning of variable domains of models in (1936a) article, in terms of which logical consequence is standardly defined. So Tarski seemed to hold a fixed-domain conception of models, which, in combination with other peculiarities of his approach, produces results incompatible with the standard account. If so, he could not provide the standard definition of truth in a structure either, since this rests on the idea that the domain of a structure can come from anywhere, provided it is a non-empty set – and not, that is, from some fixed universal domain, as was then usual in type-theoretic frameworks. The model-theoretic definition of logical consequence presupposes the standard model-theoretic definition of model, so of truth (satisfaction) in a structure. If Hodges is right that there was no standard definition of the later notion in CTFL (or, generally, in the 1930s), then there was arguably no standard definition of the former notion either.

Hodges's and Etchemendy's analyses may reinforce each other, though I do not want to submit that the two would agree on all points.[187] Thus, Hodges, for instance, makes claims not compatible with Etchemendy's conclusion to the effect that Tarski held a fixed-domain conception of models, since he claims that Tarski did not explicitly mention the domain-variability of models not because he did not accept it but because his (1936a) article was addressed to philosophical ears. His own diagnosis of why CTFL does not contain the standard model-theoretic truth definition is rather the absence from Tarski's work in the 1930s of the notion of uninterpreted non-logical constant and his reluctance to adopt a conception of formal uninterpreted language receiving various set-theoretic interpretations making its sentences true and false. At any rate, if the verdicts of Hodges and Etchemendy are on the right track, what Tarski offered in the 1930s were at best non-standard relative definitions of

[186] Tarski criticizes Carnap's account of consequence and analyticity (1934) for relying on the richness of the language, but the accusation is in fact incorrect, as we showed in the section devoted to Carnap's ideas. Interestingly, Carnap's early attempt to define analyticity (around 1932), which was also based on the substitutional reading of quantifiers, was marred by a circularity pointed out by Gödel in their correspondence, along with the suggestion that the problem could be avoided by treating second-order variables as raging over any property whatever defined over the intended domain – whether or not there is a name for it. As a matter of fact, Carnap incorporated this suggestion into his official account in (1934). For more on this interesting exchange see Awodey & Carus (2007) or Procházka (2010).

[187] Etchemendy also raised the important question whether the standard model theoretic account (Tarski-Vaught 1957) is conceptually adequate with respect to the intuitive or informal notion of logical consequence.

truth, satisfaction or logical consequence.

The detour-account of logical consequence via sentential (or propositional) functions, albeit non-standard from the point of view of modern model theory, does not by itself yield undesirable results, and it could even be thought well motivated in case of Tarski, who wanted to make use of all mathematically useful model-theoretic methods, while adhering to Frege-Peano conception of formalized language, according to which it is sort of a category mistake to talk of truth or falsity of uninterpreted sentences. Properly worked out, Hodges and Etchemendy agree, the approach via satisfaction of sentential functions is equivalent to the modern model theoretic approach. In fact, as Mancosu carefully documents, when it came to talk about interpretations, systems, realizations or models, the detour through sentential (or propositional) functions was common owing to the work of Padoa and Huntington.[188] And we have seen that despite Hilbert's pioneering work, even the members of his school used to talk about (1st order) formulas containing variables, not uninterpreted logical constants in the 1920-30s. The fact is that Tarski adhered consistently to this paradigm and offered an especially lucid description of it in his (1937) introduction to logic. He describes there a simple system of axioms involving two primitive signs: '≅' for congruence and the class-sign 'S' for line segments. The specific axioms state the reflexivity of the congruence relation in the class S of line segments, and the property that two line segments congruent to the same segment are congruent to each other: if $x \cong z$ and $y \cong z$, then $x \cong y$. It is then noted by Tarski that in actual derivations of consequences of the axioms via accepted rules of inference of the system no appeal at all is made to the actual meanings (interpretations) of the two primitive signs.[189]

It suggests itself to say that whatever admissible interpretation of the axioms that verifies them one was to choose, the situation would not change. Tarski highlights this by replacing the two primitive signs in the axioms by variables R (for relations) and K (for classes of objects) and considers any sequence of objects that satisfies the resulting system of sentential functions, according to the standard definition of satisfaction from CTFL. Such a sequence is called a model of the original set of axioms, as we would expect, given Tarski's

[188] Mancosu (2006). See also Jane (2006) and Mancosu et al (2009).

[189] This point was of course emphasized by several mathematicians and logicians who pioneered formalizations of mathematics – we find it expressed for instance in Frege, Pasch, Peano, Pieri, Padoa, Hilbert, Huntington or Veblen, among others. I am not sure, though, if Frege would have joined others in talking about various interpretations or reinterpretations of axiomatic systems. The problem is closely connected to the one under consideration: to what extent did Frege-Peano conception of logical language as meaningful formalism (regimented fragment of natural language) is at odds with the idea of truth relative to an interpretation. A good discussion in relation to Frege's views is Demopoulos (1994), where it is argued - *pace* Goldfarb (1979, 2001), Ricketts (1986, 1996) and other proponents of the *logocentric predicament* interpretation of Frege (as well as of Russell and Wittgenstein) - that Frege was not afraid of metatheoretical notions and investigations. If so, one need not be a model theorist in order to be concerned with metatheoretical questions. Much the same can be said of Tarski himself in the period of 1920s-1930s. That Frege did not find himself in the logocentric predicament is argued by Heck (2010), Stanley (1996) or Tappenden (1997) on the ground that his informal semantic explanations in *Grudgezetze* I (1893) do not seem to serve the role of mere elucidations helping the reader to understand *Begriffschrift* (1879), but are supposed to form the basis for informal arguments for consistency and soundness of the system. Tappenden (1997) shows that Frege was well versed in geometrical techniques of demonstrating independence of an axiom on the remaining axioms via interpretations.

detour-account of logical consequence.[190]

Still, non-logical constants are treated as interpreted and the model-theoretic effect is achieved (just as in his paper on logical consequence from the same period) by detour through sentential functions obtained from sentences by replacing all their (interpreted) non-logical constants (in all their occurrences) by variables of appropriate logical type. The account is thus still non-standard, by today's lights. What about the domains of models? Are they allowed to vary from a model to another or not? Several commentators opposed to Etchemendy's exegesis, have made a heavy weather of the fact that Tarski uses class-predicates to restrict the range of quantifiers, taking this to indicate that such predicates determine domains, and different predicates determine different domains.[191] Since this view has meanwhile become a kind of new orthodoxy, I should like to mention that it seems to be seriously challenged by careful recent works of Bays and Mancosu.[192] Mancosu, in particular, argues convincingly that Tarski held the fixed-domain conception even in the 1940s, since he distinguished the universe of discourse U associated with (typically, type-theoretic) language from individual domains D which are subsets of it, to which quantifiers may be restricted via class-predicates.

One could question the diagnosis sketched above on the ground that Tarski pays attention in CTFL to the relative notion of correct (true) sentence in an individual domain, which he associates with Hilbert's Göttingen school, whose members elaborated some ideas common in the algebraic approach to mathematical logic, which was pioneered by Boole, Peirce and Schröder. To the algebraic school we can count Löwenheim and Skolem, who used the informal notion of a formula (set thereof) being satisfied in an individual domain (*Individual Bereich*). Indeed, the question of completeness of 1st-order logic (isolated within the system of type theory as the *restricted functional calculus*, at the time widely considered an interesting yet small fragment of general logic) was formulated as an open problem yet to be solved by Hilbert and Ackerman already in the first edition of their classic textbook (1928):

> "Whether the axiom system is complete at least in the sense that all logical formulas that are correct for every domain of individuals can be derived from it is still an unsolved question." (Hilbert & Ackermann 1928: 68).

In later editions the authors refer that the problem was positively solved by Gödel in his dissertation:[193]

> "The question here is whether all universally valid formulas of the predicate calculus, as defined at the beginning of § 5 of this chapter, can be proved in the axiom system. We actually do have completeness in this sense. The proof is due to K. Gödel, whose exposition we shall follow." (Hilbert & Ackermann 1950: 95)

---

[190] One possible model is given when R is specified as the identity-relation and K as the class of all individuals.
[191] Gomez-Torrente (1996), Somes (1999), Simmons (2009).
[192] Bays (2001), Mancosu (2005).
[193] Its revised version was published as Gödel (1930).

The notion of 'a universally valid formula' means a formula (sentential/propositional function) true in all admissible interpretations of the calculus in individual domains (with labelled elements, properties, relations, etc.):

> "This interpretation as to content is made as follows. We consider as given a domain of individuals, to which the individual variables and the universal and existential quantifiers refer. This domain is left unspecified; we assume only that it contains at least one individual. A formula of the predicate calculus is called logically true or, as we also say, universally valid only if, independently of the choice of the domain of individuals, the formula always becomes a true sentence for any substitution of definite sentences, of names of individuals belonging to the domain of individuals, and of predicates defined over the domain of individuals, for the sentential variables, the free individual variables, and the predicate variables respectively. The universally valid formulas of the predicate calculus will also, for convenience, sometimes be called simply valid." (Ibid: 68).

It should be remarked that in CTFL (viz. Part III) Tarski defines the notion of a correct sentence in an individual domain roughly as follows:

> *A* is a true sentence of (the language of the calculus of classes - LCC) in an individual domain *a* just in case *A* comes out true when its variables are restricted to range only over classes of individuals in *a* (and not over the universal class of individuals over which the simple type-theory is built),[194]

basing this definition on the accordingly relativized recursive definition of satisfaction (technical details are in Appendix 3). He then defines two related notions: (a) the notion of a correct sentence in every individual domain (a universally valid sentence), and (b) the notion of a correct sentence in an individual domain with *k* elements. Having stated a couple of lemmas and theorems concerning the relations between these and other notions,[195] he equates (in Theorem 26) absolute truth with the special case of truth in an individual domain *a* identical with the whole universe of individuals (on which the type-theoretic system is based).[196] As regards (a), he explicitly allows domain

[194] Tarski (1935: 200-1). See also Tarski (1935: 239).

[195] Interestingly, Tarski's investigations in this condensed part of CTFL (Tarski 1935: 200-209) culminate in the purely structural definition of truth (Theorem 28) plus a decidability criterion for LCC (a general structural criterion of truth for sentences of LCC). He notes that this is by no means always possible (not even in case of languages of finite order) and that in the case of LCC it is due to its peculiar structure. Compare also the discussion in (Tarski 1935, Part VI, pp. 237-241), where he says, *inter alia*, that (1) when the set *Pr* of provable sentences of a formalized theory of a finite order is complete, it is easy to show that it coincides with the set *Tr* of true sentences (of its language), and hence *Tr* can be defined via *Pr*, which is itself defined structurally (sometimes one needs to add certain axioms to the original theory, as is the case with LCC). Moreover, the general structural criterion of correctness in a domain with *k* elements is easy to obtain only for finite *k*, by using the parallel method to the method of Boolean matrices used with respect to propositional calculus (which is complete and decidable, of course).

[196] So, if *a* has *k* elements, then a sentence (of LCC) is (absolutely) true iff it is correct in an individual domain with *k* elements. For, by Definition 26 (1935:200): a sentence is correct in an individual domain with *k* elements iff it is correct in some individual domain *a* such that *a* has *k* elements.

variation to define a universally valid sentence, which is a feature of the model-theoretic account of logical truth (validity). Having this in place, he goes on to formulate and prove various meta-theorems in terms of the relative notion of truth, among them, a version of Löwenheim-Skolem theorem concerning the size of models of 1st order theories (every 1st order theory that has a model has a model with a countable domain)[197] and Gödel's completeness theorem for 1st order logic (every universally valid 1st order sentential function is 1st order provable), both of which belong to the basic model-theoretic results and he emphasizes that such theorems can be precisely stated and proved only on the basis of precise definitions of the sort that he provides.[198]

In view of this, one may argue that Tarski had in mind the variable-domain conception of models, since (1) this was the standard conception at the time, and (2) he could not obtain the above mentioned fundamental results without presupposing the variable-domain conception of models. But Mancosu shows that the fixed-domain conception was widespread in the 1920-30s and even in the 1940s (at least in Tarski's work), and Bays explains how the fixed-domain approach can accommodate the early 'model-theoretic' results:

> "[...] there is a relatively straightforward technical trick which allows the proponent of a fixed-domain conception of models to obtain all the mathematical advantages of a variable-domain conception. Given a collection of sentences Γ, he has only to introduce a new predicate *D* (for domain) and to explicitly relativize each of the quantifiers in Γ to the predicate *D*. Having done this, he will induce a natural correspondence between the collection of variable-domain models of the original Γ and the collection of fixed-domain models of the newly relativized Γ′. As a result, every theorem concerning the collection of variable-domain models for Γ can be translated into an equally interesting (and, indeed, essentially identical) theorem concerning the collection of fixed-domain models for Γ′. The Löwenheim-Skolem theorems, for instance, translate into theorems concerning the possible cardinalities of the sets picked out by *D* (when Γ and Γ′ are first-order)." (Bays 2001: 1711).

Indeed, what Tarski says about relative satisfaction and truth (in an individual domain) seems perfectly compatible with Mancosu's and Bays' analyses, as the varying domains are subsets of the single universal domain comprising all (arbitrary) individuals (so one can apply Mancosu's distinction between the universe of discourse and its sub-domains to which quantifiers may be restricted).[199]

In consequence of his adherence to the type theoretic framework and to Frege-Peano conception of language as a meaningful formalism, Tarski did not developed a fool-blooded model-theoretic take on logic in the 1930s, which he helped to establish in the early 1950s, although he was thoroughly familiar with

---

[197] Tarski (1935: 205, the footnote n. 1)
[198] Tarski (1935: 240).
[199] Moreover, as Etchemendy (1988) points out, there is a domain variation in CTFL, but the interpretations of non-logical signs are assumed to be fixed (in particular, the interpretation of the inclusion sign is fixed as the inclusion relation between classes).

the early "model-theoretic" result (and contributed actively to this area) and was able to formulate and establish them within his preferred system. His definition of logical consequence was meant to fit the established formal-axiomatic practice of his times (viz. the practice of Hilbert's school, Peano's school and the school of American postulate-theorists), whose members commonly thought of logical consequences of axioms (also called postulates, conditions,...) as those propositional functions that come out true under all interpretations under which axioms come out true, More precisely:

> *A* is a consequence of the axiom-class *Ax* iff *A* comes out true under every interpretation of (assignment of values to) its variables under which all the axioms belonging to *Ax* come out true.

Whereas logicians belonging to these axiomatic schools took consequence to be a relation between propositional functions, Tarski defined it in (1936a) as a relation between a sentence and set of sentences, albeit using the detour through sentential (propositional) functions. It was this concept that he called the "common" or "proper" concept of consequence. And it was this concept that he wanted to capture (explicate) in his definition in terms of models, which he thought to be superior to formalized accounts of consequence in the aftermath of Gödel's theorems.[200] Already in CTFL he comments on the matter

> "The reduction of the concept of consequence to concepts belonging to the morphology of language is a result of the deductive method in its latest stages of development...In the light of the latest results of Gödel it seems doubtful whether this reduction has been effected without remainder." (Tarski 1935: 252, n. 1)

But unlike other mathematical logicians who commonly worked with this concept of consequence Tarski was able to offer a mathematically precise definition of this relation, since he already had his mathematically precise definitions of satisfaction and truth.[201]

To conclude this section, let us return to the interesting problem of ω-inferences that Tarski introduces at the very beginning of his (1936a). Tarski obviously used the example of ω-inferences to show that formalized (syntactic) accounts of consequence are inadequate as explications of the "common" or "proper" concept of consequence characterized by the fundamental properties of formality and truth-preservation. Etchemendy interprets Tarski as having in mind 1st order theories[202] and he finds Tarski's argument highly puzzling for two reasons. First, such inferences are invalid on the standard model-theoretic account of logical consequence (there are non-standard interpretations of the ω-incomplete system under which all premises are true but the conclusion is not).

---

[200] See here especially the paper of Jane (2004), with which I find myself in agreement on many points. A similar view is urged by Edwards (2003).

[201] To be fair to Carnap (1934), he is mentioned by Tarski (1936a) as the first logician to provide a plausible account of logical consequence, but he says that Carnap's definition is too complicated and applies only to a restricted class of systems. Tarski's first complaint is surely correct, but the second complaint was recently discredited by de Rouilhan (2008), who argues that Carnap's definition of consequence is in fact equivalent to Tarski's.

[202] Etchemendy (1988).

Second, by Gödel's completeness theorem, consequence coincides with provability for 1st order logic. Etchemendy's diagnosis is that Tarski had a very flexible conception of logical constant, on which ω-inferences come out valid, if only one takes numerical expressions as fixed logical constants (non-standard models being out of question).

Gomes-Torrente and others have retorted to this diagnosis that the textual evidence does not support this bold conclusion: what Tarski had in mind was that ω-inferences are valid with respect to higher order systems of the type that Tarski describes in CTFL or in (1933b) article.[203] I agree with that. But, according to Gomes-Torrente, such inferences turn out valid once we eliminate all numerical expressions in favour of their (higher-order or set-theoretical) definitional equivalents in the logicist style, say, in the following way (0, 1,.... being defined by 2nd-order formulas as finite sets):[204]

(A) *A0*. The empty set possesses *P*

A1. Sets containing only one element possess *P*

....

An. Sets containing exactly *n* elements possess *P*

A. Every finite set possesses *P*.

However, there is no need to claim that Tarski's account of consequence demands the definition of all numerical constants in terms of logical basis.[205] If '0' and 's' are among the primitive sings of the higher order system,[206] we can define the predicate '*N*' (for: *being a natural number*) in Dedekind's well-known inductive manner as the smallest set that contains 0 and is closed under the successor operation.[207] This is enough to account for the validity of ω-inferences of the following type

(B) *A0:* $P(0)$,

A1: $P(s(0))$

...

An: $P(s(...s(0)...))$

A: $\forall x(N(x) \rightarrow P(x))$,

if '*N*' is replaced by its definitional equivalent. Indeed, Tarski emphasised himself

> "[...] the necessity of eliminating any defined signs which may possibly occur in the sentences concerned, i.e., of replacing them by

[203] See especially Gomes-Torrente (1996), Bays (2001), Edwards (2003), and Jane (2004).
[204] I owe this example to Bays (2001), who remarks that 2nd-order formulas defining natural numbers as such finite sets are to be found in Tarski (1933b: 278-88).
[205] See also Saqüillo (1997).
[206] As in the system on which Gödel (1931) focuses – the higher order type system based on natural numbers.
[207] Cf. Jane (2004). Edwards (2003: 56) argues that the set contains nothing but natural numbers, provided that the domain of 2nd order quantifiers is the powerset of the domain of 1st order quantifiers, and that Tarski was actually committed to the semantics of full models for which this assumption holds.

primitive signs" (Tarski 1936a: 415)

Since A-type and B-type ω-inferences are valid on Tarski's account of logical consequence, it is uncharitable to interpret his account as applying only to systems in which a couple of signs of general logical character are treated as primitive (fixed), all other being defined in terms of them. In this respect, then, Tarski's account of logical consequence, though non-standard in certain ways, does not produce counter-intuitive results.

[6]

Semantic Conception or not?

6.1 The question of adequacy

In the previous chapters I have tried to show that Tarski's method of truth definition and his approach to theoretical semantics in general has various logical, philosophical and mathematical aspects that need to be taken seriously is we are to evaluate properly its fruits. It is my main goal in what follows to show how these various aspects are interconnected, where they are, but also, where it is better to keep them separate. This turns out important when it comes to evaluate the significance and import of Tarski's conception, because it may be quite tempting to criticise Tarski for not fulfilling ambitions he never had. I shall argue that he had philosophical ambitions, but that they were rather modest, although I admit that some of his claims are misleading in that they might suggest more ambitious philosophical aims. If the interpretation that I am about to offer is on the right track, the contribution of Tarski's semantic conception of truth and truth-definition to mathematical logic lies in his systematic formalization, indeed, mathematization of informal metamathematical ideas of the semantic variety, among whose chief fruits was: (a) a greater precision in metamathematics (precise definitions of fundamental metalogical notions as well as exact formulations and proofs of fundamental metalogical results couched in terms of such notions), (b) the method of truth (via satisfaction) definition plus definability and indefinability results concerning truth (satisfaction) for classical languages, (c) laying down a basis for a full-blooded model-theoretic approach to logic and semantics (developed at the break of 1940s-50s). Its philosophical significance lies mainly in formulating a particularly clear interpretation of the classical conception of truth going back to Aristotle, with Convention T at its heart, and the very first formal-compositional semantics, based on the separation of mathematical (formal) and empirical (foundational) issues in area of semantics.

Tarski's definitions of semantic notions have been quickly accepted by mathematical logicians because of their mathematical preciseness, extensional correctness and meta-theoretical fruits. Indeed, according to a widespread view, coextensive mathematical expressions are coextensive come what may, which is why mathematicians need not bother about their definitions being intensionally correct, in addition to being provably extensionally correct. Moreover, it would not be in accordance with the actual practice of mathematicians to maintain that a definition of a mathematical expression is not adequate unless the *definiens*

and *definiendum* are cognitively equivalent in some strong sense.[208] It does not seem likely, at least to me, that λ-*definable function* is cognitively equivalent in any interesting sense with *function computable by a Turing machine* or with *general recursive function*, though mathematicians are widely agreed that the first notion can be equally accurately defined by any of the two extensionally equivalent notions.[209]

However, truth is not just a notion dear to logicians, but it is one of those notions that have provoked attempts at philosophical analysis since time immemorial. And, of course, philosophers are typically not interested in purely stipulative definitions, but in definitions that aim to capture or elucidate some interesting concept already in use. As I already pointed out, it does not make good sense to ask of a purely stipulative notion whether or not it is adequate, whether or not it gets things right. If, on the other hand, one gives a definition of a notion already in use, we can ask not only whether the definition gets the extension of the notion right, but also whether it does a good job in capturing its meaning. Not being concerned predominantly with internal problems of mathematical logic but rather with the old question whether, eventually how it is possible to analyse the notion of truth, philosophers have naturally focused more on the adequacy of Tarski's truth definition. Its extensional correctness being granted, their evaluations of it have differed significantly, and there has been no wider agreement on its philosophical value as a definition or analysis of the notion of truth.

A reasonable way of assessing a conception of something is to see what goal its author had in propounding it, then checking whether or not he succeeded in attaining the goal. The question under consideration is whether Tarski's goal was to work out the method of constructing truth definitions for particular languages that is (provably) extensionally correct, or whether he wanted to capture something more than the extension of 'true' with respect to a particular languages, and if so, what it was that he wanted to capture. And even if one eventually comes to the conclusion that Tarski successfully attained a more ambitious goal than extensional accuracy, one can still ask if this goal is well-conceived with regard to the aim of explicating the notion of truth.

---

[208] Such a relaxed attitude towards definitions seems quite reasonable, in view of the vague character of the notion of *cognitive equivalence*.

[209] My usage of *extension* and *intension* is inspired by Carnap (1956) and possible worlds semantics. It differs from the traditional usage, according to which the extension of a term T is the class of all those things that T applies to, while its intension (or comprehension) is the class of characteristic attributes that are necessary and sufficient for an entity to possess in order to belong to T's extension. Extensional definitions are contrasted with intensional, since in giving the first we proceed by listing all the things to which T applies, whereas in giving the second we specify the characteristic attributes that all and only those entities possess to which T applies (of which T is true). As it is typically impossible to specify the extension of a mathematical term by enumerating its members, intensional definitions, so understood, are common in mathematical practice. It should be clear that one and the same extension can be picked out by different intensional definitions associated with different but coextensive terms (viz. the pair of expressions 'equilateral triangle' and 'equiangular triangle' – the first is commonly associated with the intension *a plane figure enclosed by three straight lines of equal length*, the second with the intension *a plane figure enclosed by three straight lines that intersect each other so as to form equal angles*. For a mathematician it makes some difference whether *equilateral triangle* is defined via the former or via the later formula. Hence we have a pair of coextensive predicates that are not cointensive in this sense. That does not mean, of course, that the predicates are not cointensive according to Carnap's usage.

No easy answers are available to these questions. Some commentators claim that Tarski's goal was to capture the extension of 'true' for a range of formalizable languages, while others claim that he obviously wanted more. What seems to be rather uncontroversial is that Tarski did not want to give a direct analysis of the ordinary notion of truth, on the ground that that notion does not seem to apply to sentence-types only (or in the first place) but also to beliefs, utterances, statements, or propositions (if such things are recognized). Clearly, he did not even want to give a direct analysis of the ordinary notion of sentential truth, which can be applied to any sentence whatever of any language, because he argued that this notion cannot be consistently defined in its generality. Tarskian truth definitions are to be considered rather as partial explications (in Carnap's sense) of the ordinary notion of truth, being consistent approximations of the semantic concept of truth predicated of sentences. Convention T states in precise terms the necessary and sufficient conditions that a truth definition must meet in order to be a faithful explication of the semantic notion of truth. And if the definition is materially adequate in that it satisfies it, it is assured to be extensionally adequate. Tarski's semantic definitions are designed to do justice to the notion of truth that seems intuitive and clear (up to the point, when it is confronted with paradox) and agrees to considerable extent with the prevalent usage. Or so Tarski claimed.[210]

That said, it is not settled whether material adequacy was meant to amount to more than extensional correctness of a candidate truth definition. It is sometimes possible to define truth (for L) in ways that have little in common, except that they subsume all T-biconditionals (for L) as their consequences. Tarski pointed out that, in case of certain formalized languages, *structural* truth definitions based on the method of elimination of quantifiers are possible as well as truth definitions proceeding via recursive definitions of satisfaction relation.[211] He would call both truth definitions materially adequate (because entailing all T-biconditionals), in spite of the fact that, on the face of them, they appear to be merely coextensive. Further, as Carnap remarks in his (1942), there are finite languages whose sentences display syntactic structure (but do not contain iterative constructions), for which we might define truth either in a trivial list-like manner or in a (less trivial) compositional manner, in both cases being faithful to SCT. However, the two truth definitions do not appear to have much in common except the desired deductive consequences, hence the right extension.

In fact, Tarski takes T-biconditionals (for L) to be partial definitions of 'true' for particular sentences (of L), and because of that he is willing to accept

---

[210] Hodges (2008), the proponent of the extensional interpretation of Tarski's enterprise, criticizes the view on which Tarski wanted to offer an explication (in Carnap's sense) of the notion of truth but his arguments are unclear to me. For positions that agree with mine see Künne (2003), Garcia Carpintero (1996) or Soames (1999).

[211] I owe this observation to van McGee (1993), who refers to Tarski (1948) for a purely structural definition of truth, on which true sentences coincide with those accepted by a Turing machine of a sort. The definition is faithful to SCT in that it subsumes all T-biconditionals for the object-language as its deductive consequences. It is well-known that Tarski was a pioneer of the decision method via quantifier elimination developed by Skolem and Langford. The method is used in CTFL as an example of a variant definition of truth. Hodges (2008) even claims that it was his work in this area that gave Tarski the very idea of truth definition based on recursive definition of satisfaction relation.

list-like definitions for code languages as conforming to SCT. But what do such partial definitions of truth (for L - here a very poor fragment of English)

> ‘Snow is white’ is true iff snow is white,
>
> ‘Grass is green’ is true iff grass is green,
>
> ‘The sky is blue’ is true iff the sky is blue,

have in common, except the fact that the word ‘true’ occurs on their left sides (but each time attached to a different sentence)? We see that at their right sides - partial explanations of the meaning of ‘true’ for L - different sentences of English appear. If so, it does not seem that we could make them to have something in common if we just encapsulate the information that they provide into one compact explicit definition along the following lines:

> (For every sentence *s* of L): *s* is true (in L) iff (*s* = ‘Snow is white’ and snow is white) or (*s* = ‘Grass is green’ and grass is green) or (*s* = ‘The sky is blue’ and the sky is blue).

What is more, even with respect to full-blooded languages with quantificational structure it does not seem that *satisfaction of a sentence by all/some sequences* captures the meaning of the term *true sentence*. First, it is not particularly plausible to say that ‘true’, in its actual usage, means ‘satisfied by all/some sequences’. Second, are we ready to say that ‘true’ has a different meaning when defined for a code-fragment (of English, say), as it has when defined for a quantificational fragment (of English)? Third, there are alternative ways of defining truth even for quantificational languages that do not appeal to satisfaction by sequences at all: (a) truth-conditions of quantified sentences being defined in terms of truth of (all/some of) their substitution-instances, truth for L can be defined by recursion on sentences (for L(PA), say); (b) allowing infinite disjunctions, truth could be defined for L even non-recursively, in an infinite list-like manner.[212]

One may argue that T-biconditionals are the crux of the matter when it comes to fix the meaning of ‘true’ – not just its extension - the idea of material adequacy being that T-biconditionals fix the meaning, *a fortiori* the extension of ‘true’ with respect to L, so that any two formally correct definitions are materially adequate definition of truth for L if and only if they entail all T-biconditionals for L. According this view, if we have a structural and a compositional truth definition, and both are formally correct and materially adequate, then we cannot say that they are “merely” coextensive. Or imagine that there is a formula semantically defining the set of L-truths so that we can construct a definition of L-truth in terms of that formula, which, while extensionally correct, does not have all T-biconditionals for L among its

[212] The point was made by Etchemendy (1988). On the other hand, we know that Tarski’s aim was to define ‘true’ w.r.t. L(T) on the basis of the metatheory in such a way that it makes it possible to prove important metatheorems about L with its built-in deductive theory T. Recursive definition of truth via satisfaction (and/or denotation) would thus appear to be an essential part of aim for reasonably rich languages, and not just an accidental feature (as, for instance, Etchemendy maintains)

deductive consequences (due to its deductive weakness).[213] Now, I want you to ask if the definition would be materially adequate by Tarski's lights? If your answer is "Yes", material adequacy coincides for you with extensional correctness. If your answer is "No", the two do not amount to the same, albeit material adequacy implies extensional correctness. Those commentators, who think that Tarski would have answered the question negatively, think that material adequacy assumes a certain conception of truth, however minimal, to which adequate truth definitions are expected to conform. It speaks for their interpretation that Tarski used to defer to the Aristotelian conception of truth and repeatedly claimed that T-biconditionals 'explain the meaning' of 'true' for particular sentences. However, one may well retort that this is a speculation, as Tarski did not consider the possibility of constructing an extensionally correct definition of truth that would not subsume all T-biconditionals as its deductive consequences. For all he said on the topic, it does not seem excluded that he required an adequate truth definition to prove T-biconditionals merely as a means to assure its extensional correctness. To be sure, the talk about "explaining the extension of 'true'" is gibberish; but it may well be that, for Tarski, *explaining meaning* amounts to *fixing application conditions*, hence the *extension* of an expression. For note that, according to Tarski, particular T-biconditionals are partial definitions of truth with respect to particular sentences, explaining the meaning of 'true' with regard to them. So, we may view such a biconditional as specifying the application conditions of 'true' with respect to a particular sentence: the condition under which it falls into the extension of 'true'.

Nevertheless, it seems to me that, on balance, the available evidence favours the interpretation according to which SCT is Tarski's way of making precise the Aristotelian platitude (conception), on which

> *A sentence is true iff what it says is as it says*,

or

> *A sentence is true iff it says that things are a certain way and things are that way*.

This is how Kotarbinski interpreted Aristotelian platitudes (along with other members of the Lvov-Warsaw school) and Tarski apparently subscribed to this analysis. Moreover, judging from his often quoted claim

> "We must first specify the conditions under which the definition of truth will be adequate from the material point of view. The desired definition does not aim to specify the meaning of a familiar word used to denote a novel notion; on the contrary, it aims to catch hold of the actual meaning of an old notion." (Tarski 1944: 13).

it would appear that his definitions aimed at more than extensional correctness, that is, coincidence with 'true' when restricted to particular object-languages of right type. Unfortunately, it is again by no means clear what *catching the actual meaning of an old notion* amounts to, since Tarski leaves us in the dark as to what notion of *meaning* he personally favours. At any event, he distinguishes

[213] See Gupta & Belnap (1993) or Patterson (2008b)..

extension and intension of a notion.[214] But, as regards extension, he only wishes to make it precise what kind of items the semantic notion of truth to be defined applies to – sentences-types, as opposed to utterances, beliefs or propositions. However, regarding intension he significantly said that his definitions aim to conform to the classical Aristotelian conception of truth, which he glossed in terms of *correspondence to* or *agreement with reality*. Elsewhere he says about materially adequate definitions that they capture "the current meaning of the notion as it is known intuitively."[215] This turn of phrase presupposes that, in general, the *current meaning* of a notion is something that can be said to be intuitively known by those people who understand the notion in its current usage. But it makes no good sense to say that, in general, the *current extension* of the notion is something that can be said to be intuitively known by those who understand the notion. Competent concept users certainly do not in general know the extensions of their concepts.[216]

While appropriateness of the talk about *correspondence* or *agreement with reality* with regard to Aristotle-type platitudes is questionable, and Tarski himself deemed such formulas vague, he realized that such platitudes aim to be general and point in the right direction. Still, they were imprecise, by his standards. First, once we attempt to express their content in the logical symbolism we get something along the following lines:

(For every sentence *x*): *x* is true iff ∃*p* (*x* expresses *p* and *p*).

How are we to read the apparatus of quantification on the right side? If in the standard objectual style, then, quite apart from the worry as to what kind of entities we quantify over, the sentence does not make good sense, since the second occurrence of '*p*' calls for a sentence to yield something grammatical, but objectual variables occupy nominal places. If, on the other hand, we interpret quantification in the substitutional style, then one may worry that the notion of truth is presupposed, as truth of quantified sentences is explained in terms of truth of substitution-instances of their matrixes. Furthermore, in its unrestricted form, it would likely give rise to a version of semantic paradox.

Still, those platitudes capture to some extent the powerful intuition that truth of a sentence depends both on what it says and how things are. It seems to me that Convention T is Tarski's attempt to spell out in more precise terms what this intuition amounts to with respect to a given language, taking into account also his observations on paradoxes and formal correctness in general. Indeed, applied to a given language L, T-biconditionals for L, as specified in Convention T, come close to being instances of such general platitudes. In my view, Tarski conceived of such platitudes as informal and imprecise (but not valueless) attempts to generalize what is obvious on particular T-biconditionals of the form:

'p' is true (in L) iff p

---

[214] Tarski (1944:14).
[215] Tarski (1931: 128-129).
[216] Note that one who understands a particular T-biconditional may not know whether the sentence mentioned on its left side falls into the extension of ‚true'. One only knows under what conditions it falls in its extension.

or, more generally,

> X is true (in L) iff p,
>
> where 'X' stands in for a syntactically perspicuous ML-designator of an L-sentence and 'p' for an ML-translation of that sentence.

In light of this, we can appreciate his demand that a satisfactory definition of truth for L subsumes all such instances as its special cases (deductive consequences), and why he describes such definitions at places as (possibly infinite) 'logical products' of all partial definitions. It seems that Tarski wanted to capture the meaning of the notion of truth at least to some extent via capturing the fundamental Aristotelian intuition. But it was in the nature of his explicative definition that he focused only on well-behaved objects that are theoretically more tractable than natural language – namely, formalized and extensional fragments of natural languages, free of context-sensitive, ambiguous or vague expressions. Explications of informal and imprecise notions should yield precise *explicata*, which are theoretically fruitful in that they could play the explanatory role – and could play it even better owing to their precise character - that their informal counterparts play.

However, several thinkers have argued that the notions defined via Tarski's procedure cannot play that we might expect from our informal, if imprecise notion of truth. Let us therefore see what the objections state and what merit they eventually have.

6.2 Is semantic conception of truth semantic?

We should first note that Tarski's method is considerably limited in its scope. First, it is designed to apply only to a certain family of properly restricted extensional languages, without any suggestion whether, or how, Tarski's criteria and techniques might be supplemented or modified to cover syntactically richer languages, including natural languages or languages approximating their complexity (one immediately thinks of so-called intensional or hyper-intensional constructions). Second, the method is designed to apply only to languages in which there is no context-sensitivity, no ambiguity or vacuous expressions. On the other hand, natural languages, on which formal semanticists focus their attention, abound in such phenomena. Third, in view of semantic paradoxes, Tarski argued that no consistent definition or even theory of truth that meets his adequacy condition can be given for a language that contains or can express its own notion of truth (or satisfaction) unrestrictedly applying to all its sentences (predicates), provided that we assume that classical bivalent logic holds. But natural languages certainly do seem to contain such notions of truth (satisfaction), and so Tarski concluded that neither natural languages nor properly regimented languages that approximate them in expressive power can be given consistent truth-definitions in his style.

In spite of the fact that Tarski's method is so limited in its scope, Davidson, Montague and others have persuaded many theorists that Tarskian truth definitions (or their model-theoretic extensions) provide at least a *Muster* to guide constructions of more sophisticated semantic theories for richer languages, including substantial fragments of natural languages. It is natural to call a theory

semantic if it systematically articulates the truth-conditions of sentences on the basis of the semantic properties of their significant syntactic parts and the syntactic mode in which the parts are combined. Sentences thus need to be thought of as constructible from a finite stock of simple expressions according to a couple of *syntactic* rules of admissible combination.

The semantic theory first fixes the interpretations of each simple expression, by assigning to it a semantic property of the type appropriate to its syntactic category, and then gives a couple of *semantic* rules that determine the semantic properties of complex expressions, given the semantic properties of their simpler constituent parts and their syntactic mode of combination. Finally, all this is to be arranged in such a way that the semantic properties delivered by the rules for sentences turn out to be the conditions under which the sentences are true. Plausible accounts along these lines have to cover elementary as well as more complex sentences, the details depending on the complexity of the language.

Tarski's absolute and relative method of truth definition influenced respectively the two most influential truth-conditional approaches to formal semantics: the truth-theoretic of Davidson and his followers and the model-theoretic of Montague and others. In fact, it has even been claimed - although Tarski himself would not have gone so far - that plausible semantic theories for substantial fragments of natural languages could or should have *something like* Tarski-style truth-definitions as their basis, though modified or supplemented to accommodate features not present in languages with simple logical syntax. Davidson went even so far as to suggest that we can use Tarski's method to give a compositional theory of meaning for a given natural language L:

> "There is no need to suppress, of course, the obvious connection between a definition of truth of the kind Tarski has shown how to construct, and the concept of meaning. It is this: the definition works by giving necessary and sufficient conditions for the truth of every sentence, and to give truth conditions is a way of giving the meaning of a sentence. To know the semantic concept of truth for a language is to know what it is for a sentence - any sentence - to be true, and this amounts, in one good sense we can give to the phrase, to understanding the language. […] Indeed…a Tarski-type truth definition supplies all we have asked so far of a theory of meaning […]." (Davidson 1984: 24).

In view of this, the significance of Tarski's method of truth definition would seem to be beyond any question. But several philosophers argued that strictly Tarskian truth definitions are of no use as theories of meaning or even as semantic theories. In general, these arguments aim to show that there is more to the notion of truth than we can read off from particular Tarskian truth definitions given for particular languages.

6.3 The incompatibility objection

The scepticism can be pressed from various more-or-less related directions. As regards Davidson's influential program, at least in its early stages (in the

1960ths), one obviously pertinent observation is that we cannot expect from a Tarskian truth-definition for L (even in its recursive form) that it tells us, via its clauses specifying denotation-conditions, satisfaction-conditions, or truth-conditions anything revealing about the meanings of expressions and sentences of L, for the simple reason that we have to know them in advance in order to be in a position to define truth for L in Tarski's style, or to evaluate its material adequacy as a definition of truth for L.[217] To take a simple example, we cannot expect from the following sentence

'Der Schnee ist weiss' is true (in German) iff snow is white,

that it explain to us both what the sentence mentioned on the left side means and what 'true' means. That clause cannot serve as a partial explanation of 'true', unless we already know that the sentence quoted on the left side means the same as the sentence used on the right side that we are suppose to understand. At the same time, it cannot serve as an explanation of the meaning of the quoted sentence, unless we already know what 'true' means.

In a similar way, then, the typical recursive clause for truth-functionally compounded sentences such as:

… *A und B* is true (in German) iff *A* is true and *B* is true,

cannot explain to us both what 'true' means and what 'und' means. Paul Horwich put the problem accurately when he said that: there are too many unknowns in such 'equations'.[218]

Someone may be tempted to think otherwise, focusing not on heterophonic but on homophonic T-biconditionals (or recursive clauses framed in a metalanguage that contains the object-language) such as the following:

'Snow is white' is true (in English) iff snow is white;

'Snow is white and grass is green' is true (in English) iff 'Snow is white' is true (in English) and 'Grass is green' is true (in English).

These may appear to encapsulate the information about the meaning of the quoted sentence and the connective 'and'. In principle, however, the situation is as before, only more vivid. In order to explain what 'true' means by means of such homophonic sentences, one has to understand, in the first place, what the quoted sentence means (in the first sentence - otherwise one will not understand its right side, hence the sentence itself) or what 'and' means (in the second sentence - otherwise one will not understand its right side).

The diagnosis sketched above ought to be even clearer when we replace both putative unknowns in the heterophonic equivalences with arbitrary symbols (assuming that the putative meaning-explanations, if successful, should confer the right meanings on them) so as not to be seduced by any antecedent intuitions that we may have about their meaning (alternatively, we could leave the German

[217] This observation is commonly attributed to Dummett (see 1973, 1978b), but Tarski was arguably fully aware of it (cf. Tarski 1940. See also Quine (1970) and Künne (2003).
[218] Horwich (1998).

sentences in their place and imagine the equivalences to be given to someone who knows English but no German):

a) *S* is T (in German) iff snow is white,

b) *A x B* is T (in German) iff *A* is T and *B* is T.

Ad (a): if we understand its right side and know in addition that *S* means what the right side means, then we can figure out what 'T' does not mean (e.g. that it does not mean 'false'), or we can even surmise that it might mean 'true'.[219] However, unless one knows at least so much, one cannot say anything at all about the meaning of 'T', much less to judge the material adequacy of (a) as a partial definition of 'true' w.r.t. *S*. What one can assert is at best conditional: (a) is a partial definition of 'true' w.r.t. *S* (a sentence of German), if *S* (= 'Der Schnee ist weiss') means that snow is white (or: ...if 'Snow is white' is a translation of *S*).

The situation is much the same with (b), though here I can imagine someone to claim that it suggests itself to render 'x' as 'and' and 'T' as 'true', since we can read off from that clause that a sentence compounded from two sentences by means of the operator 'x' has the property T just in case both its component sentences have that property, which condition is obviously satisfied when we take 'x' to mean 'and' and 'T' to mean 'true'. However, a moment reflection should persuade such a person that this won't do, since many other pairs ⟨operator, predicate⟩ satisfy the condition equally well: we can just as well take 'x' to mean 'or' and 'T' to mean 'a sentence' or even 'false'.

John Burgess notices that the same observation applies, *mutatis mutandis*, to recursive clauses of the relative definition of *truth-in-a-structure* along the lines:[220]

$M \models \neg A$ iff it is not the case that $M \models$ A

$M \models (A \wedge B)$ iff $M \models$ A and $M \models$ B

For, if we want such clauses to explain (define) the semantic-turnstile for our object- language, we have to rely on our antecedent knowledge of the meaning of logical operators of the object-language as given by the corresponding meta-linguistic locutions on the right side (as Tarski did in his absolute definitions). This should be immediately clear once we specify the clauses in homophonic style:

$M \models \neg A$ iff $\neg\, M \models A$

$M \models (A \wedge B)$ iff $M \models A \wedge M \models B$

To be sure, we can use those clauses to (partially) explain the meanings of the operators of the object-language, but we can do this only if we already assume it to be given that $M \models A$ means that *A is true in a structure M.* But we

[219] Note the hedge 'might': given only as much information, the condition expressed on the right side is equally satisfied when we let 'T' to mean something more bizarre e.g. 'is true and 2+2=4'.
[220] Burgess (2008b).

cannot hope to explain both – the double turn-style as well as logical operators - in the same stroke. To draw the point home, it is again helpful to replace both putative unknowns by symbols utterly foreign to us:

M --- /*A* iff it is not the case that *M* --- *A*

M --- (*A* x *B*) iff *M* --- *A* and *M* --- *B*

The lesson we should take to heart is that the project of explicitly *defining* truth for L in Tarski style is not compatible with the project of providing a truth-conditional theory of meaning for L, as conceived of by Davidson and his followers. Davidson came to accept this diagnosis. From the early 1970s onwards he has talked about reversing Tarski's strategy: instead of taking the meanings of L-expressions for granted (in the guise of translation) and defining truth-in-L on the basis of them, he proposes to take truth for granted, as a primitive notion, and employ it in interpreting an infinite number of L-sentences via a recursive theory of truth for L, based on the recursive characterizations of satisfaction and reference for L-predicates and terms respectively, and modified so as to cover apparently non-extensional contexts and the ubiquitous phenomenon of context-sensitivity (truth-conditions being specified for a sentence as potentially uttered by speaker at a time *t* and place *p*).[221] It is a matter of controversy whether the program can be successfully carried through, but there has been a fairly wide agreement even among Davidson's severe critics that his approach to semantics needs to abandon Tarski's ambition to provide an explicit definition of truth for L in favour of an axiomatic approach in terms of the primitive notion of truth.[222] Davidson came soon to champion the view that truth is a fundamental and indefinable notion – the idea that plays a crucial role in his unified theory of language, mind and action.[223]

6.4 The modal objection

Another widely discussed type of objection against Tarski's conception of truth is the so-called modal objection. It starts with the claim that the modal status of informal T-biconditionals changes significantly when 'true' is replaced in them by its Tarski style *explicatum*. And because the two notions exhibit such significant differences, the latter cannot possibly provide an accurate explication of the former, even when the former notion is restricted to a given L.[224]

The objection goes back to Moore and Lewy,[225] but in its modern version it has been presented by Putnam and Etchemendy, on whose arguments we shall mainly focus. However, the essential ingredients of their arguments were put forward already by Pap:

"Now, it is consistently thinkable that, while the moon is indeed

[221] See various classic papers collected in Davidson (1984), e.g. (1967), (1973a), (1973b). See also Davidson (1990), which is a fine statement of his view.
[222] Cf. Soames (1999).
[223] See Davidson (1990) and (2005).
[224] See Putnam (1985), (1994 a, b) or Etchemendy (1988).
[225] Moore (1953), Lewy (1947).

> round, the sentence "the moon is round" is not true for the simple reason that it does not express the proposition that the moon is round, but instead some false proposition. From the proposition that the moon is round we can logically deduce such propositions as that the earth's satellite is round (assuming the identity "the moon = the earth's satellite" to be analytic, or that there exists at least one round celestial body, not however, the proposition that the sentence "the moon is round" expresses the proposition that the moon is round. In other words, the truth-value of the semantic proposition that "the moon is round" is true depends on what proposition is expressed by this sentence, while the truth value of a proposition of astronomy hardly depends on semantical facts." (Pap 1954: 25).

Etchemendy repeats essentially this argument when he invites us to consider the following pair of claims:

a) Snow is white,

b) 'Snow is white' is true.

The two claims are related via the contingent (*sic!*) circumstance that 'Snow is white' means that snow is white, but they are different non the less, because their truth-values differ with respect to other possible worlds. Thus, we can conceive of a possible world *w* much like our actual world except that the sentence 'Snow is white' means something different in *w* than it means actually; e.g. while "snow" means still the same stuff in *w*, "white" has a meaning in *w* in virtue of which it applies to black things and 'black' has a meaning in *w* in virtue of which it applies to white things. It follows that 'Snow is white' is false in *w*, assuming that snow remains white in *w*. Clearly *a*-claim holds but *b*-claim fails to hold in *w*, and the whole biconditional made up from them thus fails to hold in *w*. And we can conceive of a possibility *w** such that snow is no longer white in *w** while 'Snow is white' means something different than it means actually, which is true in *w*, e.g. while "snow" means the same stuff in *w** which it means actually, "white" applies to cold things in *w** and snow is still a cold stuff in *w**. If so, *b*-claim holds but *a*-claim fails to hold in *w**, and the biconditional made up from them thus fails to hold in *w**.[226]

Etchemendy thus follows Pap in claiming that "semantic" claims such as

'Snow is white' is true iff snow is white,

are contingent. Recall now our first Tarskian truth-definition (D1). We showed how to deduce the T-biconditional for the sentence 'Der Schnee ist weiss' of $L_{0,}$

[226] The view that Moore, Lewy, Pap and Etchemendy all subscribe to is that sentences – even those that are determinate and eternal in Quine's sense - do not possess their semantic properties hence truth-conditions essentially, but only contingently. On the other hand, the claims expressed by 'It is true that snow is white' and (b) 'Snow is white' have the same truth-value in whatever possible world, the biconditional 'It is true that snow is white iff snow is white' holding of necessity. On this basis, one could argue that it is more plausible to take propositions (qua things named by *that*-clauses) as primary truth-bearers, adopting a propositional variant of T-schema as governing the meaning of 'true' as a predicate of propositions. Cf. Horwich (1990) or Soames (1999), who propound propositional versions of deflationism.

with help of elementary logical and syntactical laws assumed in the metatheory:

> 'Der Schnee ist weiss' is a true sentence of L_0 iff snow is white.

According to Putnam and Etchemendy, since definitions, logical and syntactical laws are necessary, and whatever can be deduced from necessary premises is necessary, the T-biconditional is necessary - holding come what may. Indeed, as the biconditional follows from the metatheory containing only obvious syntactical and logical laws plus (D1), it is a logical (or, perhaps, logico-syntactical) truth, because whatever follows from logical (...) premises and definitions is a logical (...) truth.[227] Intuitively, however, it makes an empirical and contingent claim, as counterfactual considerations seem to show. Had snow been white but 'weiss' was used to apply to black instead of white things, 'Der Schnee ist weiss' could have been false. But then the whole biconditional would have been false. Thus, semantic properties of expressions – truth included - depend on their meanings, which in turn depend on how the expressions are employed by speakers or communities. Since the definition of the truth-predicate à la Tarski does not capture the dependence of truth and related semantic properties (such as nominal denotation or predicative satisfaction) on linguistic usage, it is not a semantic predicate at all!

Putnam does not go as far as to deny that Tarski's definition, which validates all T-biconditionals for L (which are, as it were, 'true by definition') is a useful tool in mathematical logic.[228] But he thinks that it is no use as a philosophical account of truth:

> "A property that the sentence 'Snow is white' would have (as long as snow is white) no matter how we might use or understand that sentence isn't even doubtfully or dubiously 'close' to the property of truth. It just isn't truth at all […]." (Putnam 1985: 333).

According to Putnam, Tarski's formal truth-definitions extensionally agree with 'true' with respect to particular formalized languages, but they cannot capture its intension, much less its meaning, failing to reveal the semantic dimension connecting truth to meaning and linguistic practices. To drive the point home, Etchemendy invites us to consider the definitional variant of our T-biconditional, obtained by replacing 'is a true sentence of L_0' with its Tarskian *definiens*:

> ('Der Schnee ist weiss' = 'Der Schnee ist weiss' and snow is white) or ('Der Schnee ist weiss' = 'Das Grass ist grün' and the grass is green) or ('Der Schnee ist weiss' = 'Der Himmel ist blau' and the sky is blue)

---

[227] I take 'logic' here to include also what others might consider to belong to mathematics, in particular, set theory.

[228] Etchemendy says that, owing to this property, the predicate defined in Tarski style is a powerful device of "semantic ascent" (viz. Quine, 1970), serving the logician to express generalizations by means of which one can affirm or deny an „infinite lot of sentences" (or simulate infinite conjunctions and disjunctions). Thus, by affirming "Every sentence of the form *A or not A* is true" one affirms, in a way, each of an infinite number of sentences of the form *A or not A*. We shall see in Chapter 7 that this idea animates *disquotationalism* – a sentential brand of the deflationist approach to truth championed by Field (1994), Leeds (1978) or Williams (1999), among others. See Chapter 7 for more details.

iff snow is white.

This sentence is provably equivalent – given the logical and syntactical laws of the metatheory – to the sentence

Snow is white iff snow is white,

which is a logical truth failing to deliver any semantic information about expressions of L_0.

The modal objection trades on the assumption that Tarski's definitional framework renders T-biconditionals necessary (as consequences of definitions, logical laws and syntactical laws), whereas outside of the framework we would intuitively take them to be contingent. Tarski's method of truth definition thus seems to face an unpalatable dilemma. Either

i) the definition is necessary, but then T-biconditionals following from it (plus the logico-syntactic part) are necessary,

or

ii) the definition (plus the logico-syntactic part) entails contingent T-biconditionals, in which case it is not necessary, since what has contingent consequences is contingent (assuming that the logico-syntactic part is necessary).

According to the objector, either the definition fails to provide a satisfying definition of truth, since it fails to do justice to the contingent status of T-biconditionals, or it is merely materially true which is an unacceptably weak standard of definitional adequacy.

If valid, the argument threatens not just Tarski's conception of truth but any theory of truth that takes T-biconditionals as, in some sense, definitional of truth.[229] In particular, it threatens various deflationary conceptions of truth (of which more later) that construe T-biconditionals (at least those that are non-paradoxical) as definitional, axiomatic or analytic of truth, though they usually part company with Tarski in that they do not require that truth be explicitly definable in higher-order metatheory. The objection should work equally well for a single T-biconditional such as:

'Snow is white' is true iff snow is white,

supposed to partially define 'true' w.r.t. 'Snow is white'. The objector should now say that since it is implausible that a sentence has an axiomatic/definitional status unless it is at least necessary, we should reject the claim that the biconditional is definitional or axiomatic of 'true', because it is contingent.

It is perhaps clear that what is at stake is Tarski's Convention T, because it makes the allegedly problematic demand that T-biconditionals be consequences of the truth definition (as framed in the metatheory containing

[229] See also Chapter 7.

logical and syntactical laws).

6.4.1 The modal objection rebutted

According to the most typical defence-strategy against the modal objection, Tarski takes L to be individuated in part by the fact that particular expressions that belong to it are equipped with particular meanings.[230] Consequently, if there is an expression *e* whose meaning is different in L and L* respectively, then L and L* must be different languages, even though they may be syntactically indistinguishable. It follows that by adding a single expression to L or by changing the meaning of a single expression of L, we no longer have L but a different language L*. If we are imagining that a sentence or expression belonging to L could change its meaning contingently on its use by the community actually using L, we are imagining, strictly speaking, that a different, though closely related language L* would be used by the linguistic community. Once we individuate languages in this manner, we see that even informal T-biconditionals turn out to be necessary. If the following principle holds (at least for context-insensitive sentences):

> If *S* means (in L) that *p*, then *S* is true (in L) iff *p*,

and if the antecedent holds of necessity (as when we individuate L semantically), then the consequent should also hold of necessity.

It is fair to say that Putnam is well aware of this manoeuvre. He reports to remember that when he talked about this problem with Carnap, the immediate reaction of Carnap was to distinguish two notions of language, namely: *language as a system of communication* and *language as a semantical system*. This distinction is explained in his *Introduction to Semantics* (1942). Now, here is what Putnam remembers Carnap to have said:

> "Everything depends on the way the name of the language—'German' or whatever—is defined." If by "German" we mean "the language spoken by the majority of the people in Germany" or "the language spoken by the people called 'Germans' in English," then it is only an empirical fact that "Schnee" refers to the substance snow in German, and only an empirical fact that "Schnee ist weiss" is true in German if and only if snow is white ... . But in philosophy, Carnap urged, we should treat languages as abstract objects, and they should be identified (their names should be defined) by their semantical rules. When "German" is defined as "the language with such and such semantical rules" it is logically necessary that the truth condition for the sentence "Schnee ist weiss" in German is that snow is white." (Putnam 1988: 63)

In order to extract from this a response to the modal objection, we are to think of abstract languages as semantically individuated, that is, as interpreted in Carnapian semantical systems that fix or stipulate the denotations of their primitive expressions via the rules of denotation and, on this basis, determine the

[230] See, for instance, Davies (1981), Garcia Carpintero (1996), Künne (2003).

truth conditions for all their sentences (via compositional rules of truth). The specific character of such stipulations assures that the semantical system S for L, based on the syntactic theory for L, entails Tarskian adequacy conditions in the form of T-biconditionals w.r.t. L. Indeed, Carnap says that the semantical rules not only fix the meanings of sentences of L (via fixing their truth conditions) but they also *define* the notion of truth for L, similarly as rules of denotation interpret non-logical constants as well as define the notion of denotation for L.

The trouble is that Carnap wants the semantical rules of S to do a double duty. On the one hand, they are to interpret L in the way we have just described; on the other hand, they are to "define" the semantic notions of designation and truth that occur in them. Putnam complains that he begs the question at issue when he invokes his abstract-language conception, since that conception invokes semantic notions that are to be explained:

> "What I thought but did not say was: And, pray, what semantical concepts will you use to state these 'semantical rules'? And how will those concepts be defined?" (Ibid: 63)

The problem that Putnam seems to have in mind is that Carnap wants both to have his cake and eat it, when he wants his semantical rules to play both roles – interpretive (stipulating the meanings of L-expressions in S) and definitional (stipulating the meanings of semantic expressions for L). However, if the semantical rules are to play the first role, the meanings of L-expressions being treated as so far "unknown" or "unsettled" (it is the rules that confer on them their meanings), they have to make use of already understood semantic notions that cannot themselves be "unknown" or "unsettled". And if the semantical rules are to play the second role, the semantic notions being treated as so far "unknown" or "unsettled" (it is the rules that confer on them the meanings), they have to make use of already understood L-expressions that cannot themselves be "unknown" or "unsettled".

The problem reminds us of the incompatibility objection against Davidson's early program in truth-conditional theory of meaning or Burgess' objection targeting the idea that recursive clauses of the model-theoretic truth definition could fix the meanings of L-expressions (including logical constants) occurring in them as well as the intended meaning of the semantic turnstile. Once again: there are too many unknowns but not enough equations to help us to calculate their values. This observation animates also Etchemendy's claim,[231] approved by Davidson,[232] that semantics (model-theoretic or truth-theoretic) needs to use an undefined metatheoretic notion of truth. Interestingly, another distinguished semanticist could have made essentially the same point a couple of decades before:

> "...in discussing the semantical rules of a formalized language, we thought of the concepts of denoting and of having values as being known in advance, and we used the semantical rules for the purpose of giving meaning to the previously uninterpreted logistic system. But instead of this it would be possible to give no meaning in advance to the words "denote" and "have values" as they occur in

[231] Etchemendy (1988).
[232] Davidson (2005).

> the semantical rules, and then to regard the semantical rules, taken together, as constituting the definitions of "denote" and "have values" (in the same way that the formation rules of a logistic system constitute a definition of "well-formed). The concepts expressed by "denote" and "have values" as thus defined belong to theoretical syntax, nothing semantical having been used in their definition." (Church 1956: 64)

Church has in mind semantic definitions in a more powerful metatheory free of semantic notions, but he muddies the water by claiming that they belong to theoretical syntax. Gödel and Tarski showed that semantic notions for a reasonably powerful L(T) cannot be reduced to syntactic notions for L(T). The former can be defined in a higher-order metatheory MT, but the latter can be defined within T itself, provided that T contains elementary arithmetic.[233] At any rate, it seems that he wants to distinguish two ways of looking at a semantical framework. We may view it as using primitive semantic notions (of *designation*, *truth* or *having values*) to specify the interpretation of an uninterpreted L, or we may view it as fixing the meaning of semantic notions for a fully interpreted L. The former procedure seems better suited to reveal the nature of semantics, as distinct from syntax:

> „…in order to maintain the distinction of semantics from syntax „denote" and „have values" should be introduced as undefined terms and treated by the axiomatic method…And in fact Tarski's *Wahrheitsbegriff* already contains the proposal of an axiomatic theory of truth as an alternative to that of finding a syntactical equivalent of the concept of truth." (Church 1956: 66)

I tentatively suggest that Carnap could have had dimly in mind a middle way between the two alternatives mentioned by Church. Semantic rules of the type

> Let '*a*' denote Chicago;
> Let '*P*' denote the property of being a large city;
> Let *PR(in)* be true iff the designation of *in* has the property designated by *Pr;*
> etc.,

fix the interpretation of L, and, by the same token, the extensions of semantic notions featuring in them (for the object-language under consideration). In this sense, the semantic rules can be said to define the semantic notions, even though they presuppose our grasp of them required for them to play the interpretative role. As a matter of fact, one can sometimes see the same idea at work in

[233] To be fair to Church, Carnap talked about semantic notions being definable in a sufficiently powerful syntactic metalanguage, and even Tarski used to speak of defining semantic notion on the basis of morphology (his label for syntax – a theory of structural properties of expressions). What was clear to them was that the metalanguage needs to be logically stronger, with higher order variables (or stronger set-theory). Hence, syntax plus higher-order (stronger) logic is enough to define semantic notions, but it does not follow that the higher order apparatus itself is to be reckoned to syntax (what about the substantial ontological commitments of the apparatus?), unless one claims that any semantics-free theory is *eo ipso* syntactic. Which use of 'syntax' seems perverse to me.

expositions of model-theoretic truth definitions: by stipulating the application conditions of semantic notions one interprets expressions of the object language via fixing their extensions in a given structure, by the same token, fixing the extensions of the semantic notions that are already assumed to be understood to some extent (it being assumed to be understood e.g. that '$M \models A$' means that *A* is true in *M*).

So perhaps there is no problem with Carnapian semantic approach after all. Be that as it may, Tarski's own semantic account seems to be immune to Putnam's objection, since Tarski did not conceive of languages as interpreted via rules involving the notion of truth (or designation, satisfaction). He explicitly stressed this difference:

> " ...regard the specification of conditions under which sentences of a language are true as an essential part of the description of the language" (Tarski 1944: 373, n. 24)

Fernandez Moreno is without doubt right to say:

> "In Tarski's semantics the interpretation of a language principally results from the co-ordination of the basic constants of the object-language with their metalinguistic translations; in the process no appeal is made to the definition of truth. In contrast, in Carnap's semantics the interpretation of a definite language calls for the application of the definition of truth which is to be found in the corresponding truth-rules ..." (Fernandez Moreno 1992: 38)

The first defence-strategy blocks the modal objection by claiming that T-biconditionals are necessary, provided that we individuate languages semantically so that e.g. a sentence "Der Schnee ist weiss", qua a sentence of L_0, (or German) cannot but mean that snow is white, hence cannot but be true iff snow is white. But there is n alternative way of answering the modal objection that consists not in individuating languages semantically, but rather in distinguishing two different ways in which we can modally truth-evaluate sentences of a given (interpreted) language, that give rise to two notions of truth. Thus, Gupta and Belnap wonder how it can be that we can interpret both of the following sentences (numbered as (19) and (20) respectively) as true, given that one claims that 'Snow is white or snow is not white' is necessarily true, while the other seems to deny this:

> The sentence 'snow is white or snow is not white' is necessarily true,
>
> If 'or' had meant what 'and' means then the sentence 'snow is white or snow is not white' would not have been true.

Their solution consists in distinguishing two notions of truth along the following lines:

> "To determine whether a sentence falls, in a world *w*, under the first notion of truth – the notion that is employed in (19), and which we shall call the logical notion – we determine whether the sentence is true in *w* with the meaning it has in the actual world. On the other

> hand, to determine whether it falls under the second notion of truth in *w* – the notion that is employed in (20), and which we shall call the non-logical notion – we determine whether it is true in *w* with the meaning it has in the world *w*." (Belnap & Gupta 1993: 21).

Truth, they say, depends both on meaning and facts. But in the first case the meaning has as "frozen" in our actual world so that we do not have to take into account its variability across possibilities, whereas in the second sense it can vary across possibilities, and this variability is reflected in truth-evaluation of a sentence w.r.t *w*. The authors then make two important comments. First, the distinction is significant if not all languages are individuated semantically (*sic!*), otherwise all languages would possess their semantic properties necessarily, and the distinction would make no difference, as there would be no meaning variations across possibilities to take into account. Second, Convention T yields a reasonable adequacy criterion only for the logical notion of truth: only then the definition yields consequences that have the right modal properties (necessarily true T-biconditionals).

Having the two defence-strategies in place, let me now voice some misgivings about them. Both defence-strategies attempt to defend Tarski's conception against the modal objection by denying the intuition to the effect that T-biconditionals are contingent, arguing that they are necessary either under the conception of language as individuated by semantic properties or under the logical notion of truth (viz. Gupta and Belnap). Since they take the truth definition plus logical and syntactical laws to be necessary, its consequences (as appended to logico-syntactical axioms) are necessary too. So there is no modal problem after all. In Chapter 7 we shall have an occasion to see that similar defence strategies against the modal objection have been proposed by truth-deflationists who take the truth-schema '*p*' *is true iff p* (or better, its non-paradoxical instances) to be somehow definitional or axiomatic of the notion of truth. It seems to me, however, that one may accept the intuition that T-biconditionals are contingent and still deny that there is a modal problem for Tarski's conception of truth definition.

How so? It should be remarked that Tarski nowhere says anything about the modal status of T-biconditionals (and the biconditional-forming connective itself is material) or of the truth definition itself. He does not even say that logical or syntactical laws are necessary! His view seems to be that definitions are sentences expanding deductive theories so that they can be treated as additional axioms - sentences we accept as true without requiring any further proof of them (axioms plus their deductive consequences are then "asserted sentences" of the deductive theory). He seems to think that the content or meaning of a notion as it occurs within a deductive theory is fixed by accepted sentences in which it features, hence by asserted sentences (specifically, axioms) of the theory. It can be said that the meaning of a notion is its inferential role or potential within the deductive theory. Patterson rightly emphasises[234] that Tarski thought that the meaning (qua inferential role) of certain notions – namely of the primitives – cannot but be fixed in this axiomatic way; other notions, however, can be shown to be "reducible" to other notions, which fact can be codified in the form of an explicit definition. The definition $A = df.\ B$ is thus a sort of

[234] Patterson (2008b).

license to substitute *B* for *A* in all sentential contexts and inferences of the background theory, so that *A* can always be eliminated.[235] The fact that some sentences as opposed to others are accepted as true confers on them a specific status in the deductive theory, but that does not mean that it confers on them specific modal status. The same applies, *mutatis mutandis*, to definitions. It seems likely to me that Tarski would have detested the traditional metaphysical talk about contingent and necessary truths. He was well known for his mistrust to intensional notions (or operators) and for his preference for extensional languages (as object-languages as well as meta-languages). Like Quine – with whom he shared quite a few ideas in this area - he was quite sceptical about the possibility of making a principled distinction between analytical and factual truth,[236] and he would therefore have seen little hope in drawing a principled dividing line between definitional and non-definitional truths, or between necessary and contingent truths.[237]

It is thus doubtful whether Tarski would have been moved by counterfactual considerations conducted in terms of possible world. He could reject them out of hand, on the ground that they are formulated in all too unclear terms, or he could say that since the definitions in an extensional metatheory are construed as merely materially true, what follows from them in combination with the logical and syntactical laws of the metatheory is expected to be only materially true. Granted, had the world been different, the words and sentences of L could have changed their meanings, and the sentences that actually do duty as T-biconditionals for L could have ceased to be T-biconditionals. But Tarski did not mean his formal truth definitions to predict how the property of L-truth would behave in such counterfactual situations; he wanted to capture the application conditions of 'true' for L – hence the set of true sentences of L - given L's actual semantic properties.[238] Had L been different in its semantic aspects, another truth definition in Tarski style would have applied to it, reflecting its changed semantic properties. There is thus no need to interpret Tarski as adopting a conception of language as having its semantic properties *necessarily*. L may retain its identity even if it expands or subtracts its vocabulary over time or its expressions change their meanings and semantic properties (contingently on how L-speakers use them). Consequently, sentences that actually do duty as T-biconditionals for L can turn out false, in which case other sentences would do duty as T-biconditionals for L. So, no sentence that has the status of a T-biconditional for an L-sentence is false, so long as it has that status. But it can lose this status, in which case it might be false. What Tarski requires is that the object-language (of a reasonable complexity) has (a) fixed vocabulary of context insensitive words with unambiguous (and perhaps non-vague) meanings, (b) precise (extensional) syntax recursively fixing its set of context insensitive (declarative) sentences. Such a language is artificial to a significant extent indeed (though, according to Tarski, it is a regimented fragment of a natural language). But this does not mean that he thought that it possesses its semantic properties *necessarily*.

---

[235] For an excellent modern account of definitions see Belnap (1999).
[236] See Frost-Arnold (2006) for a rewarding discussion.
[237] Recall here Quine's *Two Dogmas* (1953c), where Quine took for granted the notion of logical analyticity (necessity) but Tarski deems even this notion to be controversial to the extent that it is controversial what words or constructions count as logical.
[238] See Patterson (2008b).

There is thus a clear sense in which truth, as defined by Tarski, *depends* on meaning (*pace* Putnam and other modal-objectors): the notion of truth does not apply to a sentence no matter how its meaning varies (diachronically or across possible worlds). It just does not offer - and does not pretend to offer - any deep explanation of this dependence (going beyond Aristotle's platitude to the effect that *S is true if things are as S says they are*), which would show how truth and related semantic properties supervene on facts about use of expressions by speakers or communities in their socio-physical environments.

6.5 List-like character of Tarski's truth definitions

A *prima facie* reasonable philosophical desideratum on an adequate definition of the concept *F* is that instead of enumerating (all and only) *F*-instances it should state something common to all and only *F*-instances that explains why they instantiate *F* and not some other concept. As Socrates put it: "…I am seeking that which is the same in all these cases…" (*Meno* 75a). Thus, consider the following list-like definition of the concept of *chemical element*:[239]

> (For all *x*): *x* is a chemical element iff (=df.) *x* = Hydrogen or *x* = oxygen or *x* = Nitrogen, or ...,

specifying, case-by-case, all the 253 known chemical elements. Its obvious shortcoming is that someone familiar with it could well know what elements there actually are without knowing what makes them to be *chemical elements* (except the fact that they are on the list). Such a person would not be able to tell why Hydrogen, Oxygen, ..., appear on the list; he/she would be completely at a loss to determine whether the so defined concept applies to a newly synthesised element not yet on the list. Shortly: the definition is non-extendible. Or, to put it slightly differently: the definition does not give us any hint as to how to go on in new cases.

For some concepts, to be sure, we can frame extensionally correct definitions enumerating all their instances. Thus, we can define the notion of *a solar planet* in the following easy manner:

> (For every *x*:) *x* is a planet orbiting the sun iff *x* is = Earth or *x* = Mars, or *x* = Venus, or *x* = Saturn, or *x* = Jupiter, or *x* = Mercury, or *x* = Neptune, or *x* = Uranus, or *x* = Pluto,

But we cannot define in this manner many concepts of philosophical interest to us such as *being a person*, *being a machine*, *being good*, *being just*, *being virtuous*, etc. The problem is not primarily that we cannot hope to enumerate all their instances, when there is an infinite number of them, but, rather, that we might have a perfectly accurate explanation of what being *F* amounts to, without being able to specify all *F*-instances, because it is one thing to know the application conditions of *F*, and quite a different thing to know what particular items actually instantiate *F*. The following is a correct definition, if anything is:

---

[239] I owe this example to John Searle's unpublished manuscript 'Truth'.

(For every *x*:) *x* is a bachelor iff *x* is an unmarried man,

since whatever falls under the concept of *being a bachelor* falls under the concept of *being an unmarried man*, and vice versa. Moreover, it provides a sort of criterion (or rule, if you like) that can be applied in any given case: to find out whether *x* is a bachelor, check whether *x* is a man and, if so, whether or not *x* has a wife. But we are surely not required to know, of any given *x*, whether or not *x* is a bachelor (passes the criterion supplied by the *definiens*), in order to have the concept of *bachelor*.

Now, a whole bunch of more-or-less related objections to Tarski's conception of truth definition concerns the fact that Tarski's truth definitions, whether simple or complex, implicit or explicit, rest in one way or another on list-like or enumerative definitions of semantic notions (truth, satisfaction, denotation).

6.5.1 The epistemic objection

Consider, for instance, the explicit definiens of (D1):

> (*s* = 'Der Schnee ist weiss' and snow is white) or (*s* = 'Das Grass ist grün and the grass is green) or (*s* = 'Der Himmel ist blau' and the sky is blue)

we see at once that there is nothing obviously *semantic* about it, even though it meets Convention T , being faithful to the semantic conception of truth. It does not seem to throw any light upon how truth-conditions are determined by the semantic properties, including semantic relations objects, of their significant constituents and their manner of composition (it is no use to say that it relates sentences to extra-linguistic entities, facts or states of affairs, since it obviously does not do that). In light of this, the truth definition for L_2 certainly appears to be more illuminating of the semantic structure of L_2 than the two previous truth-definitions are of the semantic structures of L_0 and L_1 respectively.

What this shows is that a definition of truth for a language (at least for a simple language) might satisfy the demands of Tarski's semantic conception of truth without being *semantic* in any natural sense of that notion. That the truth-predicate defined in this way does not have the right connections to semantic facts becomes transparent, says the objector, once we realize that if we did not understand German but had reliable information that the following informal T-biconditional is true

> * 'Der Schnee ist weiss' is true (in German) iff snow is white,

we would have at least some information about the meaning that 'Schnee ist weiss' has as a sentence of German, hence as a sentence of L_0 (assuming we understand the meta-language in general, and 'true' in particular). We could infer, for instance, that it does not mean that snow is not white, that snow is black and other things incompatible with the fact that snow is white. Now it is standard to assume that an explicitly defined predicate can be replaced by the definiens without any loss (throughout extensional contexts). But, once again, when we replace 'true' in the informal T-biconditional for 'Der Schnee ist

weiss', by its explicit Tarskian definiens and perform admissible simplifications licensed by logical and syntactical laws for L_0 , this is equivalent to the triviality

> Snow is white iff snow is white,

which, obviously, does not tell us anything interesting about the meaning or semantic properties of the sentence 'Der Schnee ist weiss'. One can understand what the list-like truth-definition for L_0 states without knowing anything at all about the semantics of L_0.

We have said that the truth definition for L_2 appears more illuminating of the semantic structure of L_2 than the truth-definitions for L_1 is of the semantic structures of L_1, and this in turn appears more illuminative than the truth definition for L_0 (owing to its use of recursion). But the appearance may be delusive. In fact, the objector argues, essentially the same considerations apply, *mutatis mutandis*, to the truth definition for L_1 and L_2. Especially if they are put in their explicit forms (and only these meet all Tarski's strictures), it becomes clear that one could understand what they state without knowing any semantic fact at all about L_1 or L_2. As Etchemendy and Soames have pointed out,[240] when we replace in the T-biconditional for 'Der Schnee ist weiss' (as a sentence of L_1), the predicate 'true' by its explicit Tarskian definiens, we will obtain the long claim:

> There is a set TR of sentences of L_1 to which 'Der Schnee ist weiss' belongs such that...

It could be shown that after admissible simplifications of this claim we get something that, once again, does not state any information at all about the meaning or semantic properties of the sentence 'Der Schnee ist weiss' in L_1.[241] Even at the intuitive level: we understand this claim, but unless we know that TR is the set of true sentences of L_1 (which information is not stated in the definition), we cannot, solely on the basis of this claim, infer anything concerning the meaning that 'Der Schnee ist weiss' has as a sentence of L_1.

Moving finally to the truth-definition for L_2, we observe that the part of it that takes care of the satisfaction conditions of simple sentential functions is trivially list-like or enumerative:

> *p* satisfies *f* iff (*f* = 'x_k ist ein Mann' and p_k is a man) or (*f* = 'x_k ist eine Frau' and p_k is a woman) or (*f* = 'x_k liebes x_l' and p_k loves p_l).

But then, it would appear, essentially the same line of argument can be used to show that the whole definition has in itself no semantic import. Or so the objector claims.

6.5.2 The objections from *non-extendibility* and *no commonality*

There is a set of related objections to the effect that Tarski's truth-definition are non-extendible, each particular truth definition for a particular language being

[240] Etchemendy (1988: 56-57); Soames (1999: 102-105).
[241] Cf. Soames (1999: 104).

based on the language-specific clauses for basic cases (be it simple sentences, predicates or terms of the object-language). Consequently, a particular truth definition for a given object-L does not contain anything to guide one in extending it to new cases (if a new sentence, predicate or term is added to L, or a new language is considered).[242] Thus Max Black complained in his critical review (1948) that Tarskian truth-definitions are language-relative, differ in extension from one language to another, and inevitably fail to reveal what they have in common - what makes them to be definitions of truth and not of something else. Dummett put it slightly differently in his classic article on truth: Tarski's truth definitions introduce extensionally adequate truth-predicates but they do not tell us anything about what the point of so introduced predicates is.[243] The worry is not that we have been given no hint how to define truth for languages containing other than extensional constructions. Rather, the formal truth definition gives us no hint as to how to extend it to new cases that are logically familiar.

Once again, it will be useful to illustrate this objection on the elementary truth definition (D1). Let L_{0*} be just like L_0 except that it contains in addition the sentence 'Die Sonne ist gelb' (if we allow the phenomenon of language expansion or change, we might instead talk about L_0 at two different temporal or counterfactual stages of it). It may seem that (D1) instructs us how to go on in defining truth also for L_{0*}, namely that we are to add just one extra-clause for 'Die Sonne ist gelb':

D1*: *s* is a true sentence of L_{0*} iff

(a), (b) and (c) as in (D1),

(d) *s* = 'Die Sonne ist gelb' and the sun is yellow.

However, nothing in the *definiens* of (D1) - supposed to explain all the meaning of the defined notion - dictates that we extend (D1) in this particular way, pairing 'Die Sonne ist gelb' with the condition expressed by the sentence translating it, and not, say, in the following way:

D1:** *s* is a true sentence of L_{0*} iff

(a), (b) and (c) as in (D1),

(d) *s* = 'Die Sonne ist gelb' and Venus is pink.

In this case, truth for L_{0*} is not adequately defined, since its consequence is that

'Die Sonne ist gelb' is a true sentence of L_{0*} iff Venus is pink.

But we know this, because we know in addition something not stated in (D1), namely that (a) (D1) is intended to meet Convention T - to be materially adequate - and English sentences expressing the conditions paired with sentences belonging to L_0 give their meanings (are their correct translations);

[242] The point was made by Field in his classic paper (1972) and by Dummett (1978b).
[243] Dummett (1978a: xx-xxi).

and (b) 'The sun is yellow' gives the meaning (is a correct translation) of 'Die Sonne ist gelb'. That this information is not stated in (D1) should be clear from the fact that one does not need it in order to understand what (D1) says. Someone who knows English but no German at all could very well understand (D1) – stated in English - without possessing this information, and hence without having the slightest idea as to how to go on in defining truth for L_{0*}.[244]

A closely related objection is that Tarski showed at most how to define 'true' w.r.t. L_1, 'true' w.r.t. L_2,…, or 'true' w.r.t. L_n, each time obtaining an extensionally correct truth definition for a particular language L_i, but he did not explain what the various truth-predicates - 'true' as defined for L_1, 'true' as defined for L_2,.... - have in common, how they are related to our pre-theoretical notion of truth. Tarski's truth definitions thus fail to tell us anything about *truth in L*, for variable 'L'. This objection and the non-extendibility objection are indeed two sides of the same coin, since both trade on the observation that Tarskian truth definitions are based on language-specific clauses that specify, case-by-case, application-conditions of semantic notions to non-logical primitives. Since each particular truth definitions is so intimately tied to just one language with its specific vocabulary, none of them can tell us what they have in common (indeed, why all of them deserve to be called definitions of *truth*), and none contains enough information to guide us in framing truth definitions for different languages.

The objector of this callibre complains that Tarski did not explain the general or relational notion of *x is true in L*, for a variable 'L'. And this holds good even if 'L' is restricted to range over properly restricted and formalized languages. At this point, it may be interesting to consider a very similar objection that Quine levelled in his *Two Dogmas* against Carnap's recursive definition of analyticity introduced in his *Meaning Postulates*:[245]

> A sentence *S* is analytical in L iff
>
> (a) *S* is a meaning postulate of L, or
>
> (b) *S* follows logically from the meaning postulates of L,

all the meaning postulates of L being enumerated. Quine complains that Carnap's definition gives at best the definition of analyticity for one particular language L, that is, the definition of the non-relational notion *x is analytic-in-L*. What it fails to deliver is the explication of the general relational notion *x is analytic in L*, for a variable '*L*'. Moreover, Carnap's definition of analyticity is inadequate, even if '*L*' is restricted to formalized languages that Carnap (like Tarski) works with:

> „The notion of analyticity about which we are worrying is a purported relation between statements and languages: a statement is said to be *analytic for* a language *L*, and the problem is to make sense of this relation generally, that is, for variable '*S*' and '*L*' […] By saying what statements are analytic in L_0 we explain 'analytic-

[244] *Mutatis mutandis*, analogous arguments apply to truth definitions based on the notion of denotation and satisfaction, since these are also defined case-by-case, in a list-like manner.
[245] Carnap (1947).

> for-L_0' but not 'analytic', not 'analytic for'. We do not begin to explain the idiom '*S* is analytic for *L*' with variable '*S*' and '*L*', even if we are content to limit the range of '*L*' to the realm of artificial languages." (Quine 1953c: 33-34)

It should be clear that a particular Carnapian definition of analyticity gives no hint as to how to go on in new cases, owing to the circumstance that its base clauses enumerate L-sentences that are to count as meaning postulates of L. It is hard to overlook a close parallel between Tarski's truth definition for a particular L and a Carnapian analyticity definition for such L. Marian David argues that Quine's critique of Carnap is well-taken,[246] provided that the *explicandum* is the general relational notion *x is analytical in L*, for variable 'L'. Carnap acknowledged himself that the relational notion is the *explicandum* as the following letter of Quine documents:

> „The main illumination for me, in our joint performance at Chicago, was that your "analytic-in-L_0", and "analytic-in-L_1" etc., which I have represented as mutually irrelevant and irrelevant to "analytic-in-L" (for variable 'L'), do have a principle of unification precisely in the sameness of the explicandum. The issue therefore becomes: is it a reasonable explicandum?"[247]

Granting this clarification of what is at issue between the two thinkers, Quine's point stands untouched, in spite of the fact that he shows a characteristic tendency to evade the issue at hand by proposing to focus instead on the adequacy of the definiendum itself (the general notion of analyticity for variable 'L'). No wonder, David duly points out, because if Quine's objection really discredits Carnap's definitions of analyticity, a parallel objection would seem to show that Tarski's truth-definitions cannot provide an adequate *explicatum* of the general relational notion of *truth in L*, for variable 'L'. Clearly, Tarski's definition of a monadic predicate 'analytic in L_0', or better, of the hyphenated predicate 'true-in-L_0' does not throw any light on the general relational predicate 'true in L', for variable 'L', and it obviously does not matter at all how many such restricted monadic truth predicates we have defined in Tarski's style.

Did Quine think that we should not expect of truth definitions that they elucidate our general relational notion of truth (at least as it is restricted to formalized languages)? If so, his works from the period in question do not contain any argument for such a radical disproportion in approach.[248] Or was

---

[246] What is somewhat puzzling, in view of the fact that the aim of *Two Dogmas* is to discredit the very notion of analyticity, is that Quine seems to accept - but perhaps only for the sake of argument - that there is a relational and translinguistic notion of analyticity, of which we can make at least so much sense that we can see that Carnap's attempt to explicate it is a blatant failure.

[247] Quoted according to David (1996: 284).

[248] A few years later he would argue that general relational semantic notions had better be abandoned. Already in *Two Dogmas* Quine rejected the notion of analyticity, his main reasons being contained in the second part of it. In *Word and Object* he extended the attack to the notion of meaning and synonymy using his radical translation argument (partly in response to Carnap's *Meaning and Synonymy in Natural Language* (1955)). He started to talk about immanent (intra-linguistic) notion of truth that applies to sentences of one's mother (home) language only, contrasted, presumably, with transcendent (translinguistic) notion of truth (reference, etc.) that applies also to sentences of other languages (even those that we do not understand). But such

Quine unaware of the striking parallel? No, he surely was well aware of it. By way of praising the merits of Tarski's truth definitions, he says that:

> "In Tarski's technical construction, moreover, we have an explicit general routine for defining truth-in-L for individual languages L which conform to a certain standard pattern and are well specified in point of vocabulary. We have indeed no similar single definition of 'true-in-L' for variable 'L' [...]" (Quine 1953b:138)

But a few pages before he says:

> "Thus it will be recalled that the problem of construing 'analytic was recognized as the problem of construing 'analytic in L;' for variable 'L'." (Ibid: 134)

The explanation of Quine's reluctance to treat truth and analyticity on a par can be found in the same work. He praises there Tarski's semantic definitions as extensional semantics at its best (qua *theory of reference*), claiming that this brand of semantics is definitely to be preferred to intensional semantics (qua *theory of meaning*). He further notes that for notions belonging to extensional semantics we have the following principles governing their application that he calls "paradigms of clarity" (henceforth I keep his numbering of them):

> (7) "'____ ' is true-in-L if and only if ____
> (8) '_____' is true-in-L of every _____ thing and nothing else.
> (9) '_____' names-in-L _____ and nothing else. (Ibid: 135)

> "[…] which, though they are not definitions, yet serve to endow 'true-in-L' and 'true-in-L of' and 'names-in-L' with every bit as much clarity, in any particular application, as is enjoyed by the particular expressions of L to which we apply them. Attribution of truth in particular to 'Snow is white', for example, is every bit as clear to us as attribution of whiteness to snow." (Ibid: 138)

However, we have no such glaring paradigms of clarity for the notions of intensional semantics, in particular, for analyticity.

It should be remarked that Quine's disquotational paradigms do not feature a variable 'L'; L either coincides with or is a restricted fragment of the metalanguage - here of English. Indeed, in order to avoid semantic paradox, L had better be a proper part of English such that (7), (8) and (9) do not belong to it. Consequently, (7) cannot cast light on the relational notion of *truth in L*, for a variable 'L', since is an English paradigm for *truth in a (restricted) fragment L of English*. The same applies, *mutatis mutandis*, to predicative application (satisfaction) and nominal denotation (reference, designation). Quine does not

---

translinguistic notions depend on the notion meaning or interlinguistic translation, which he questioned. Immanent (interlinguistic) semantic notions, on the other hand, are quite safe; and, of these, immanent truth, denotation and application have the considerable merit (not possessed by intensional notions like analyticity) that their meaning is governed by obvious disquotational principles of the type: '*p*' *is true iff p* (which, properly restricted, do not give rise to semantic paradox), endowing these notions with a useful expressive role in infinite generalizations (making it possible to accept or reject in a short manner "an infinite lot of sentences").

maintain that (7) throws light on it; what he claims is that Tarski gave us "an explicit general routine for defining truth-in-L for individual languages L". He sums up:

> "See how unfavourably the notion of analyticity-in-L, characteristic of the theory of meaning, compares with that of truth-in-L. For the former we have no clue comparable in value to (7). Nor have we any systematic routine for constructing definitions of 'analytic-in-L', even for the various individual choices of L; definition of 'analytic-in-L' for each L has seemed rather to be a project unto itself." The most evident principle of unification, linking analyticity-in-L for one choice of L with analyticity-in-L for another choice of L, is the joint use of the syllables 'analytic'." (Ibid: 138)

Quine seems to have a point when he says that there is a difference between truth and analyticity in that for the former we have a glaring paradigm of clarity in (7), whereas for the later notion we have nothing of the sort. What (7) provides can be regarded a paradigm of clarity only for sentential truth for fragments of English, itself framed in English (or in a more comprehensive fragment of English). And Tarski gave us a 'general routine' for defining such restricted notions for a whole class of English fragments, provided that they are formalized and semantically open, and fragments serving as metalanguages are logically stronger than object-fragments for which truth is defined. In this sense, (7) is an English disquotational paradigm of clarity for *truth in L*, where 'L' can even be treated as a variable ranging over such sub-fragments of English.

To draw the point home, let '*L*' range over semantically open fragments of English that are finite. Tarski showed that there is a general (if trivial) procedure for defining *truth* for any *L*:

A) For every sentence *x* of *L* write down the biconditional of the form:

'____' is true in L iff ____,

in which both blanks are filled in by *x*.

Alternatively:

B) Let all sentences of *L* be enumerated in an *n*-termed sequence *s* without repeating terms. Then the following biconditional defines truth for L:

(For every sentence *x* of L): *x* is true sentence of L iff (1) *x* = '___' and ___, or (2) *x* = '___' and ___ , ..., or (*n*) *x* = '___' and ___,

in which the 1st sentence of *s* fills in both blanks of the 1st clause, the 2nd sentence of *s* fills in both blanks of the 2nd clause, and so on, for any finite *k*, the *k*-th sentence of *s* filling in both blanks of the *k*-th clause.

As regards English fragments with a more complex but formally

manageable structure that contain an indefinite number of sentences, general routines for defining their truth predicates in more powerful English fragments will be more complicated (their adequacy being judged in light of the paradigm (7)). Tarski's great contribution was that he succeeded in making them precise via his recursive techniques. The main difference is that the crucial role is not played by base-clauses for sentences (as in finite or propositional languages) but by clauses specifying satisfaction conditions for simple predicates and/or denotation conditions for simple terms. For such predicates and terms we possess paradigms of clarity (8) and (9) respectively (or something equivalent).

According to Quine, a certain step towards generality was thus taken, that has no parallel when we consider the notion of analyticity. The idea is that the general Tarskian routine (B) gives us a hint as to how to go on in new cases or indicates what Tarski's truth definitions for languages belonging to the range of 'L' have in common. But what we still lack is a corresponding paradigm of clarity for translinguistic truth to guide us in constructing truth definitions for languages other than sub-fragments of English. To be sure, for fragments of other languages there are analogous paradigms of clarity and analogous general routines for defining their restricted disquotational truth predicates. Thus, for German fragments, we have this paradigm of clarity

(7*) '____ ' ist wahr-in-L wann und nur wann ____,

and we could readily formulate an analogous general routine for defining *wahr-in-L* in German, where L is a semantically open sub-fragment of German. Yet, and here comes the crux of the matter, neither (7), nor (7*) nor anything of the sort gives us a paradigm of clarity for the general notion of sentential truth, there being no general routine to define truth for one language in a different language, which would be on a par with (B). One may think that T-schema is the desired paradigm of clarity for translinguistic notion of sentential truth (restricted, perhaps, to semantically open languages), Convention T providing a hint of a corresponding general definitional routine, when combined with Tarski's enumerative-cum-recursive techniques. But Quine should not accept this suggestion, because T-schema and Convention T rely heavily on the notion of interlinguistic sameness of meaning, in the guise of (correct) translation, which, by his own lights, is problematic. Later, Quine came to acknowledge this, as he came to emphasize that the notion of truth is reasonably clear (only) in its disquotational and immanent use.

For obvious reasons, Tarski's truth definition for a language in a different language cannot employ disquotational base-clauses (whether for sentences, predicates or terms). The routine (B), on the other hand, owes its projectibility to the disquotational feature doing its work in base-clauses. For instance, given an English truth definition for a restricted fragment L of English, framed according to (A) or (B), (B) instructs us how to extend the definition to a new English sentence by properly disquoting it. For more complex general routines using, say, recursion on the complexity of L-predicates, a new case would typically be a simple English predicate not yet in L, and the general routine would instruct us how to deal with it by properly disquoting the predicate.

However, there is no analogous mechanic procedure for the

translinguistic notion of truth. Even if one is armed with Convention T, one has to know, in addition, the meaning (translation) of a newly added sentence (predicate or term) in a meta-language, in order to be able to extend the original truth definition to it. As David succinctly put it:

> "...knowing how to construct the base clauses for L1 does not in general help construct the base clauses for other languages. So for each of L1, L2, etc., constructing their base clauses will be "a project unto itself" to borrow a phrase from Quine. There is, then, no good reason for saying that 'true-in-L1', 'true-in-L2', etc. share a "principle of unification", hence no good reason for saying that they will serve as adequate explications of the general notion of truth." (David 1996: 293)

To sum up our discussion: if Quine's objection against Carnap's definition of analyticity-in-L as an explication of the general relational notion of analyticity is on the right track, then Tarski's truth definitions do not fare any better as explications of the general notion of truth (not even when it is restricted to formalized languages). For one thing, such definitions are based on language-specific base clauses (supplemented or not by recursive clauses), and fail to capture the relational notion of truth for much the same reason as Carnapian definitions fail to capture the relational notion of analyticity. For another thing, particular Tarski's truth definitions do not suggest to us any general routine - comparable to (A) or (B) - for defining translinguistic truth predicates. More precisely, to the extent that there is a general routine, it essentially involves Convention T plus techniques of enumeration-cum-recursion, in which case, however, it relies on the notion of interlinguistic translation, and assumes knowledge of the meaning (translation) of each new case to be considered.

6.6 Concluding remarks

In my view, the foregoing considerations do not diminish the value of Tarski's theory. As Davidson put it: Tarski "made it thunderously clear" that it not possible to define our general (pre-theoretical) notion of truth, there being no formal way of consistently capturing this notion. He has in mind Tarski's famous argument to the effect that truth, in its translinguistic cannot be defined in a formally correct and materially adequate manner. According to Tarski, there is no hope of giving a fully general definition of the sort:

> For every language *L* and sentence *S*: *S* is true in *L* iff ...*S*....*L*,

for variable 'L' and 'S'. Although he suggested that it is possible to frame a generalized definition for formalized languages at least, such a definition could not be fully general, since the language in which it would be framed could not itself belong to the range of 'L' (on pain of paradox). Compare this statement of Tarski:

> "There will be no question at all here of giving a single general definition of the term ['true sentence']", (Tarski 1983: 133)

So, truth will have to be defined always for a particular (properly

formalized) language, and Tarski showed how to so for a properly restricted and formalized L in meta-L, assuming the (correct) translation of L into meta-L. More precisely: if (a) L is an arbitrary semantically open object-language with the exact structure of the right logical type (i.e. a formalized language - finite or extensional), and if (b) meta-L is a logically stronger metalanguage, and if (c) we know the meanings (translations) of L's expressions in meta-L, then Convention T plus the techniques of enumeration-cum-recursion instruct us how to define truth for such L in meta-L.[249] This suggests to us a certain general procedure for defining truth predicates, albeit not completely routine or mechanic.

What about the epistemic objection? Must we grant the critics their radical claim that Tarski's method of truth definition has no semantic import? Granted, once satisfaction or denotation and, in terms of them, truth is so defined, there remain no semantic terms in their definiens, since they all disappear in favour of the terms of a semantics-free metalanguage (meta-theory). Now, this is all to the good, if one is after extensionally adequate definitions of semantic notions in terms of set-theory, logic, syntax and the vocabulary of the object-language to which the definitions are relativized (which was arguably Tarski's main logical aim, for reasons discussed in the earlier sections). However, set-theoretical definitions, like those definitions that are trivially list-like, can capture only the extensions of restricted semantic notions but they cannot possibly explain, explicate or reduce them to terms that are conceptually, ontologically or epistemologically more fundamental.

While mathematical logicians tend to emphasise that Tarski showed us how to do formal semantics (model theory) by precise mathematical methods (indeed, within set theory), philosophers are naturally not so impressed by this aspect, as is clear from the fact that they do not widely accept Tarski's claim that his method of truth definition captures the actual meaning of an old notion (at least for L). It must be granted, I think, that the fact that Tarski showed how to interpret the truth theory in a more powerful mathematical theory does not yet mean that he showed something philosophically (as opposed to mathematically) important about truth and related semantic properties. The critics are right in so far as they claim that Tarski's method of truth definition does not tell us everything there is to the notion of truth by way of a philosophical account. Consequently, Tarski's claim that his truth definitions catch hold of *the actual meaning* of our intuitive notion of truth is unfortunate and misleading, to say the least.

Yet, I would like to say that all this is of a limited importance as a critique of his method of truth definition. Tarski could have been confused or simply careless in his claims on this matter, but even though his truth definitions do not in fact *catch hold of the actual meaning of an old notion of truth*, his method might still provide a different sort of insight into the notion of truth than its analysis, showing us how to reconstruct the truth conditions of sentences of L (of a certain type) as systematically depending on the semantic properties of their significant parts, based on their syntactic structure. What should be clear but is often overlooked is that there is more to Tarski's method or conception of truth definition than particular formal definitions for particular formalized

[249] For Tarski, the final step would be to turn the definition to a fully explicit one.

languages. Its heart is Convention T with the material adequacy condition, and the standard technique for a reasonably rich L is a recursive characterization of satisfaction (or denotation). In tandem, these two aspects indicate where the semantic significance and value of Tarski's theory of truth. It is its inherent part that the truth definition for L is intended to be a materially adequate definition of truth for L, faithful to one powerful intuition about our common notion of truth: *S is true iff things are as S says they are*. This intention takes the form of the requirement of its implying all T-biconditionals for L. And it is the fact that it has such consequences that confers upon it the status of the truth definition for L, for then the claim that all and only true sentences of L satisfy the definition holds. Tarski's truth definition for L thus cannot be taken as a stipulative definition, although it cannot be construed as a meaning or concept giving definition either. Furthermore, the fact that it is intended to be materially adequate shows that it is not divorced from meaning at all, but depends in a way on it (viz. Convention T and its appeal to the notion of translation), since every change in meaning of sentences calls for a brand new truth definition entailing a new set of T-biconditionals.[250] The semantic import of a Tarskian truth definition for L, as based on the recursive definition of satisfaction for predicates (or denotation for terms) of L, rests simply on the claim that the definition is a materially adequate definition of truth for L.

True, in making this crucial claim we must use our ordinary notion of truth. But there is no vicious circularity involved, because the claim itself is not part of the definition framed in the metalanguage, but a higher level claim about our success in achieving what has been our goal all along. Incidentally, this is the reason why Tarski's truth definitions, without further ado, look semantically uninformative. But once we know that the definition is materially adequate, we can read it as containing relevant information about the semantic properties of L, based on its compositional structure (it is here where the recursive technique plays its role). Indeed, it is remarkable that Tarski did not hesitate to use the notion of truth in his original version of Convention T:

> "A formally correct definition of the symbol '*Tr*', formulated in the metalanguage will be called an adequate definition of truth if…[…]." (Tarski 1935: 187-188).

This shows, as Davidson noted,[251] that *we are not wrong to interpret* Tarski's method of truth definitions as a method of fixing the extension of 'true' for particular languages (properly formalizable), that takes full advantage of our pre-theoretical grasp of the notion of truth in the form of the semantic conception of truth. Drawing on the observation of Etchemendy, Davidson and Heck argued[252] that the claim that Tarski's truth definition for a quantificational language L is materially adequate in that it satisfies Convention T makes the definition equivalent to an axiomatic theory of truth for L, whose axioms mimic the clauses of the recursive truth definition (viz. e.g. (D5)), with semantic notions construed as its primitives:

---

[250] This holds, whether we consider this as a change of language (in the standard response to the objection of Putnam and Etchemendy), or not (in the non-standard response that I favour, sketched in 6.4).
[251] Davidson (1990).
[252] Etchemendy (1988), Davidson (1990), Heck (1997).

Base axioms (satisfaction-conditions for simple predicates):

p satisfies 'v_k is a man' iff p_k is a man;

p satisfies 'v_k is a woman' iff p_k is a woman;

p satisfies 'v_k loves v_l' iff p_k loves p_l;

Recursive axioms (satisfaction-conditions for complex predicates):

p satisfies $\neg A$ iff *p* does not satisfy *A*;

p satisfies $A \wedge B$ iff *p* satisfies *A* and *p* satisfies *B*;

p satisfies $A \vee B$ iff *p* satisfies *A* or *p* satisfies *B;*

p satisfies $\forall v_k A$ iff every sequence *p** which differs from *p* at most in its *k*-th member satisfies *A.*

Truth axiom (truth-conditions for sentences - 0-argument predicates):

(For every sentence *x*:) *x* is true iff *x* is satisfied by all sequences.

For various reasons, the question whether axiomatic theories along these lines can serve as empirical theories of understanding/interpretation, as Davidson famously claimed,[253] is rather controversial. But, at the very least, such theories *do* seem to cast some light upon the semantic structure of quantificational languages: showing how the truth conditions of sentences systematically depend on the truth-relevant semantic properties of their significant parts based on their logico-syntactic structure (thereby revealing the inferential structure of quantificational languages). And that is no mean achievement.[254]

Tarski was well aware of the fact that his method of truth definition had this dimension, when he said that in order to capture the truth conditions (T-biconditionals) for each of the indefinite number of sentences of a reasonably rich language L, the most *simple* and *natural* way is to proceed through a recursive characterization of the satisfaction conditions for complex predicates in terms of the satisfaction conditions of simpler predicates that are their immediate significant constituents (if L has complex terms, a recursive characterization of the denotation conditions is added).

---

[253] Davidson (1990).

[254] Heck (1997) further argues that such a theory can be an empirical theory of truth for L, given that the axioms have an empirical substance - the modal-objection is then irrelevant, since the consequences of empirical (hence contingent) axioms are contingent. Against Etchemendy's objection to the effect that axiomatic semantic theories using primitive semantic notions are hard to reconcile with Tarski's aim of providing a provably consistent theory of truth, Heck argues that axiomatic theory is interpretable in a higher order metatheory – via Etchemendy's "connecting principles" – this being the proof of their relative consistency. He points out, rightly to my mind, that already Tarski established this in CTFL, though not for compositional truth-theories which mimic his recursive definitions (but employ primitive semantic notions). What this shows is that Tarski's methods are not incompatible with the project of empirical semantics (not, at least, for the reasons that Etchemendy mentions).

> "[…] it turns out that the simplest and the most natural way of obtaining an exact definition of truth is one which involves the use of other semantic notions, e.g., the notion of satisfaction […]." (Tarski 1944: 345)

He deemed it natural, I dare say, because it is a rather intuitive thought that semantic properties of complex expressions, *a fortiori* truth conditions of sentences, depend on the semantic properties of their significant constituents. Logicians have always honoured intuitions such as:

> a sentence of the form *name+predicate* is true iff the predicate is true of what the name denotes (or: what the name denotes has the property that the predicate denotes/signifies).

Such intuitions have been elaborated and generalized in various ways, but the essential idea remained: truth or falsity of sentences (of at least certain forms) depends on the semantic properties of their significant syntactic parts, so that it is possible to specify the condition under which a sentence (of an appropriate form) is true in terms of the semantic properties of its parts.[255] Tarski arguably took truth to be the central semantic notion. Admittedly, he said that when it comes to formulate semantics for complex languages it is most convenient and natural to define satisfaction first, since it is easy to define remaining semantic notions – truth included - as special cases of satisfaction. His work on semantic definability of *n*-dimensional sets (in a given structure – e.g. of real numbers) played an important role here, because truth is there defined as a limit case of satisfaction-relation: *satisfaction by 0-term sequences* (in CTFL: satisfaction by all/some sequences).[256] At the intuitive level, however, we would explain satisfaction in terms of truth, rather than *vice versa*:

> *a* satisfies 'F' iff '*a* is F' is true.

Tarski had an original idea of how to use this intuition in the formal definition of truth, without presupposing the notion of truth for an object-language, but defining it instead in terms of satisfaction. In order to define adequately the notion of truth for a quantificational language that has an indefinite number of sentences each of which has a certain exactly specifiable logico-syntactic form, it is natural to take full advantage of the fact that truth or falsity of sentences of increasing logico-syntactic complexity depend on the semantic properties of their immediate constituent parts, and ultimately on the semantic properties of their simple parts. For languages worth of that name, a satisfactory characterization of truth conditions for their sentences is naturally going to be framed in terms of the relations of satisfaction or denotation (or something analogous) between expressions and objects; typically, such relations will have to be defined recursively on the logico-syntactic structure of expressions. Such a definition then displays the contribution of that structure to

---

[255] Semantic ideas and methods anticipating those that he came up with had been in the air since at least Plato's semantic analysis of simple predications (to be found in the *Sophist*), in which *something is said of something*, according to which (when we generalize the original idea of a verb signifying an action, by taking *property* to be a generic term covering everything that can be predicated/said of something): a predication of the form *name+verb* is true iff the denotation of its nominal element possesses the property (or falls under the concept) signified by its verbal element.

[256] Cf. Tarski (1931).

the truth-conditions of sentences of L_2, and provides the basis for a precise account of logical consequence. This might well be the main import and value of Tarski's method of truth definition, in which also its philosophical importance largely consists.

On the other hand, Tarski repeatedly stressed that adequate definitions (or axiomatizations) of semantic notions for L should be general formulas (or, as he also put it, logical products) subsuming all instances of relevant schemata (w.r.t. L) such as:

'F$x_1,...,x_n$' is satisfied by $\langle a_1,...,a_n,\rangle$ iff F$a_1,...,a_{n;}$

'*N*' denotes *a* iff $a = N$;

X is true iff p.

In view of this, it would seem that his considered approach to semantics does not give pride of place to the idea that an adequate explanation of truth for L must render it a "correspondential" notion (*truth* =df. correspondence to facts, or designation of facts, or something of the sort).

That is not to say that there are no traces of this idea in his conception. We mentioned that Tarski said that his definition aim to conform to the classical Aristotelian conception of *truth as correspondence*. However, we also pointed out that he explicated the imprecise idea via his Convention T, which has nothing at all to do with correspondence, as traditionally understood. Still, one may argue that his theory of truth is correspondential, provided that correspondence is construed as based on (or derived from) the relations between sub-sentential expressions and (typically) extra-linguistic items. Nominal denotation and predicative satisfaction are surely understood by Tarski to be paradigmatic such relations – being expressed by what we call *directly relational semantic notions*. So understood, Tarski's truth definition for a reasonably complex L can be interpreted by someone as an explication of a "correspondential" notion of truth for such L, in which the informal and imprecise notion of object-based correspondence is replaced by the notion of predicative satisfaction (and/or nominal denotation), which, in turn, can be explained in precise mathematical terms. Now, Tarski made it clear that whether an adequate definition of truth for L takes this "object-based" form depends on the logico-syntactic complexity of L. Thus, we saw in Chapter 3 that for impoverished languages with a finite number of sentences it is possible to define their adequate notion of truth in a trivial manner, by enumerating all instances of the T-schema, whereas with respect to more complex languages we are forced to make use of a more devious apparatus, taking full advantage of the idea that the truth-value of sentences of certain logical forms are determined by semantic features of its components - such as names and predicates - in accordance with compositional rules. This may be read as a sign that, at least in such paradigmatic cases, Tarski's theory of truth is a version of object-based correspondence theory of truth, truth being explained (reduced to) word-to-world relations. However, all depends on how the notions of satisfaction and denotation are accounted for.[257] As Tarski fixes such relations via lists (i.e.

[257] Davidson (1969), Fernandez-Moreno (2001) and Field (1972) argue that Tarski's theory of

equates their definitions with logical products of instances of the relevant schemata), many commentators have argued that his conception is not sufficiently robust to do duty as a full-blooded theory of truth or semantics. While some people take this to be its obvious shortcoming (there is more to truth than Tarski's theory reveals), others take it to be its laudable, deflationary feature (there is less to truth - and semantic notions in general - than philosophers traditionally thought).

---

truth is correspondential in character. However, Davidson (1990) abandoned this view in later works (arguing that if sentences correspond to anything at all in Tarski's theory, then they correspond all to the same thing).

[7]

Robust conception or not?

7.1 Field's battery of objections and the physicalistic-naturalistic agenda

In a widely read paper on Tarski's theory of truth,[258] Hartry Field offered a concise if a bit idiosyncratic exposition and critical evaluation of it, especially of its alleged claim to have solved the riddle of truth by way of reducing it (related semantic notions) to the ideology acceptable to the physicalists. What Field argues, in a nutshell, is that the base clauses of Tarski's recursive definitions that define satisfaction (or denotation) are only extensionally correct, since they proceed by enumeration of cases. This, however, is not enough for a genuine reduction, by any standards common in serious science. Along the way Field levelled a battery of objections to Tarski's theory of truth as it was originally expounded, which, in one form or other, are still discussed in the literature.

Field's evaluation of Tarski's theory of truth is by no means exclusively negative. He credits Tarski for having showed us how to reduce the truth (at least for a range of formalizable languages) to what he calls ***primitive denotation***, which involves the notions of *an object being denoted by a name*, *a predicate applying to an object*, and *a functional symbol being fulfilled by pairs of objects*. What he criticizes is, first, Tarski's alleged claim to have rehabilitated truth in particular, and semantic in general, by showing us how to reduce semantic notions to the ideology acceptable to physicalists (i.e. general logical plus physical notions); second, what he takes to be Tarski's unfortunate and misleading exposition of his basic semantic ideas, which, in Field's view, encouraged the first mistaken claim, and, moreover, made his truth definition seem more restricted that it needs be, once it is properly exposed.

In order to keep track of Field's argumentation, we need to extend our familiar quantificational language L_2 by a few names, e.g. {'Günter Grass', 'Helmut Kohl', 'Angela Merkel'}, together with a few term-forming expressions, e.g. {'Der Vater von', 'Der Bruder von'}. Let 'L_3' be a name for the so extended language – a bit richer, but still very poor 1st order fragment of German. In view of this fact, we have to revise also the syntactic description, adding the recursive definition of terms (since iterative term-forming functors generate an indefinite number of complex terms), and accordingly also the

[258] Field (1972).

definition of atomic formulas and sentences. *The set of terms of L_3* is the smallest set such that (i) all variables are terms, (ii) all names are terms, (iii) every expression of the form *f*(*t*) is a term, where *f* is a 1-place function-symbol and *t* a term (*closed terms of L_3* form the subset of this set to which all and only those terms belong that do not contain any variable). *The set of atomic sentential functions (Asf) of L_3* is the smallest set such that (i) every expression of the form *tP* is *Asf*, where *P* is a 1-place predicate and *t* a term, (ii) every expression of the form t_iPt_k, is *Asf*, where *P* is a 2-place predicate and t_i,t_k are terms (*atomic sentences of L_3* form the subset of this set to which all and only those sentential functions belong that do not contain any variable). *The set of sentential functions of L_3* is the smallest set containing atomic sentential functions, being closed under the operations of negation, disjunction and universal quantification. Finally, *the set of sentences of L_3* is the subset of this set containing all and only those sentential function that do not contain any free variable. Having this in place, it is quite straightforward to define denotation (for terms) and satisfaction (for sentential functions) in the following:

Variant A

I The denotation of a term *t* of L_3 w.r.t. to the sequence *s* - Den(*t*)$_s$

(a) Den(v_k)$_s$ = s_k, if v_k is the *k*-th variable;

(b) Den('Günter Grass')$_s$ = Günter Grass;

Den('Angela Merkel')$_s$ = Angela Merkel;

Den('Helmut Kohl')$_s$ = Helmut Kohl;

(c) (i) Den(*Der Vater von* t_k)$_s$ = *a* iff t_k is a term, and there is a *b* such that Den(t_k)$_s$ = *b*, and *a* is the father of *b*,

(ii) Den(*Der Bruder von* t_k)$_s$ = *a* iff t_k is a term, and there is a *b* such that Den(t_k)$_s$ = *b*, and *a* is the brother of *b*.

II A sequence *s* satisfies a sentential function *f* of L_3

(a) If *f* = *t ist eine Frau*, where *t* is a term, then*:*

s satisfies *f* iff Den(*t*)$_s$ is a woman;

If *f* = *t ist ein Mann*, where *t* is a term, then*:*

s satisfies *f* iff Den(*t*)$_s$ is a man;

(b) If *f* = t_i *liebes* t_k, where t_i, t_k are terms, then:

s satisfies *f* iff Den(t_i)$_s$ loves Den(t_k)$_s$

.....

The remaining recursive part of the definition of satisfaction (w.r.t. *s*) for non-atomic sentential functions, and, in terms of it, of sentential truth, is the same as in (D6) for L_2.

To be accurate, we should note that Tarski's official strategy would be to stick to the satisfaction part of definition, reducing nominal denotation to

sequence-relative satisfaction pursuing the following strategy:

> "To say that the name *x* denotes a given object *a* is the same as to stipulate that the object *a* ...satisfies a sentential function of a particular type. In colloquial language it would be a function which consists of three parts in the following order: a variable, the word 'is' and the given name *x*." (Tarski 1935: 194)

Implementing this observation, what we get is the definition of nominal denotation for L3:

> **A name *n* of L_3 denotes an object *a*** iff one of the following conditions is satisfied
>
> (a) *n* = 'Helmut Kohl' and *a* satisfies *x = Helmut Kohl*;
>
> (b) *n* = 'Angela Merkel' and *a* satisfies *x = Angela Merkel;*
>
> (c) *n* = 'Günter Grass' and *a* satisfies *x = Günter Grass*

In this way, Tarski "reduced" nominal denotation to predicative satisfaction. However, as it is obvious that the sentential function '*x = n*' is satisfied by the object *a* iff *a* is *n*, we can further simplify the definition of denotation:

> **A name *n* of L_3 denotes an object *a*** iff one of the following conditions is satisfied
>
> (a) *n* = 'Helmut Kohl' and *a* is Helmut Kohl;
>
> (b) *n* = 'Angela Merkel' and *a* is Angela Merkel;
>
> (c) *n* = 'Günter Grass' and *a* is Günter Grass,

This simplified definition of nominal denotation for L_3 is equivalent to the clause (Ib) of A-variant truth definition.

Now, had Tarski considered languages containing function symbols, by means of which complex terms are formed, he would have urged an analogous definition of *a being denoted by a complex term of the form f*(*t*) in terms of

(i) *t* denoting *b*, for some object *b*,

and (ii) *a* satisfying the sentential function '*x* is *f*(*b*)'.

Since *a* satisfies '*x* is *f*(*b*)' iff *a* is *f*(*b*), we can just as well use the definition:

> **A complex term of L_3 of the form *f(t)* denotes an object *a*** iff one of the following conditions is satisfied
>
> (a) *f* = 'Der Bruder von', and there is a thing *b* denoted by *t*, and *a* is the brother of *b*;
>
> (b) *f* = 'Der Vater von', and there is a thing *b* denoted by *t*, and *a* is the father of *b*,

which is equivalent to the clause (Ic) of A-variant truth definition.[259] What is

[259] An alternative could be to eliminate names and functors of L3 (hence complex terms) via contextual definitions in Quine's style (cf. Quine 1970) in favour of quantified variables,

important is that the notion of denotation featuring in the definition can be reduced to the notion of satisfaction. In this way, Tarski could avoid discussing syntactically more complex languages – for which the truth definition would be more cumbersome – being content to point out that they can in principle be dealt with in much the same spirit.

Field takes care to make it clear that just as satisfaction-conditions are defined for simple predicates by enumeration of basic cases (pairing each simple sentential function with a certain condition so that the material adequacy criterion for satisfaction is satisfied), the denotation-conditions for simple/closed terms are defined by pairing each name with its denotation via its meta-linguistic translation (so that the material adequacy criterion for denotation is satisfied). The base clauses enumerate basic cases of the defined notion, in terms of which other cases are defined. Field then makes a couple of comments on A-variant, intended to question its claim to be a plausible semantic theory, rehabilitating semantic notions by reducing them to scientifically acceptable ideology.

Apart from the fact that particular truth definitions in A-variant style apply only to formalized languages of a certain specific (1st order) structure supposed to be free of context-sensitive features, a particular such definition defines truth (and related notions) only for a particular L. Moreover, it can define truth for L, only as L happens to be in a given temporal stage *t*, where it has a specific non-logical lexicon, which fact is reflected in list-like clauses specifying denotations (etc.) of its primitives. Let me sum up what Field says about A-variant truth definition for L:

(a) Owing to the requirement that the sense of every expression be unambiguously determined by its form (thus eliminating ambiguity and context-sensitivity), A-variant cannot be applied to (reasonably rich fragments of) natural languages;

(b) Owing to the language-specific clauses reflecting the specific lexicon of L, A-variant is so intimately tight to L that it is inapplicable to other languages, not even to languages of the same logical type (i.e. 1st order languages with the same semantic categories occupied by different items);

(c) Owing to the language-specific clauses reflecting the specific lexicon of L at a temporal stage *t*, A-variant cannot be applied to L, as it happens to be at different temporal stages, in which L has modified lexicon.

And, last but not least:

(d) Although it may encourage a misleading appearance of success in reducing semantic notions to the physicalist ideology, owing to its list-like clauses, it completely trivializes the worthwhile project of reducing semantics to the ideology acceptable to physicalists.

---

predicates and relations, and then applying Tarski's original method that needs only the notion of sentence-relative predicative satisfaction (whose limiting case is sentential truth).

I am not about to discuss in detail (a), as it is not essential to Field's critique of Tarski's program. I have made some remarks on it earlier, noting that various modifications of Tarski's method of truth definition can be made that allow us to deal with the phenomenon of context-sensitivity (more serious problems might be raised by the phenomena of lexical ambiguity and vagueness, which seem to be ubiquitous in natural languages), which would call for some revision of the criterion of material adequacy spelled out in Convention T.[260] At any event, the point can hardly count as a serious criticism of Tarski, since he excluded unregimented natural languages from the scope of his method. (b) implies that A-variant truth definition for L is no use when it comes to define truth for a structurally similar yet lexically different languages. Furthermore, (c) implies that A-variant is no use even when it comes to define truth for L, a single word being added to or subtracted from L at a different temporal stage. From a certain perspective, it is one and the same objection, because we can say that when a single word is added to (or subtracted from) L, a new language L* results. We discussed this objection in detail in Chapter 6.

Let me first mention another point made by Field that is worth stressing in this connection. He observes that truth-predicates defined in A-variant style – each predicate being defined for a given particular language with its specific lexicon at a particular temporal stage - differ in extension, for the simple reason that different languages contain different words, hence sentences. Accordingly, they differ also in meaning as well, on the assumption that difference in extension entails difference in meaning.[261] Consequently, Tarski's method as exemplified in A-variant has little to do with

> "...explaining the meaning of the word 'true'....the definition works for a single language only, and so if it "explains the meaning of" the word 'true' as applied to that language, then for any two languages L1 and L2, the word 'true' means something different when applied to utterances of L1 than it means when applied to utterances of L2!" (Field 1972: 356)

As it is not plausible that 'true' has different meanings applied to different languages, it is only charitable not to interpret Tarski as wanting to give an analysis of the notion of truth via meaning-explaining definitions.

One may wonder, in view of these comments of Field, why Tarski championed A-variant style of truth definition. Field asks this question and gives the following answers: (1) since Tarski could not be concerned with meaning-explaining definitions of semantic notions, his reason should be closely connected to his ambition to reduce semantic notions to the respectable conceptual basis; but (2) the strategy he chose to achieve that goal, namely A-variant truth definition, was misguided, since no genuine reduction could be

---

[260] Cf. Davidson (1984), Lepore & Ludwig (2005), Kaplan (1989), or Larson & Segal (1995).

[261] There is at least one commentator who would disagree here. David (2008) suggests the possibility that Tarski took 'true' to be a contextually-sensitive word of a sort, whose extension depends on chat contextually salient language it is applied to, but its meaning remains the same across such varying contexts (analogy: 'I' changes its reference but not meaning, depending on who utters it). But there is little evidence to ascribe this view to Tarski.

achieved in that manner. If Tarski thought more about the matter, he would have realized that what he showed was how semantic properties of expressions reduce to semantic properties of their simpler constituents (based on syntax). On the other hand, he said nothing illuminating at all regarding reduction of semantic properties of primitive expressions themselves.

To make his criticism particularly vivid, Field constructs another definition of truth, of which he says that although Tarski did not give it, he "should have given it", because it displays more accurately what he achieved. In our reconstruction of Field's favourite variant of truth definition, we shall focus on L_3. Minor calligraphic variants aside, the truth-definition for L_3, as Tarski "should have given it", runs as follows:

Variant B

I The denotation of a term of L_3 w.r.t. to the sequence *s*

(a) $\text{Den}(v_k)_s = s_k$, if v_k is the k-th variable;

(b) Den('Günter Grass')$_s$ is what 'Günter Grass' **denotes**;

Den('Angela Merkel')$_s$ is what 'Angela Merkel' **denotes**;

Den('Helmut Kohl')$_s$ is what 'Helmut Kohl' **denotes**;

(c) (i) Den(*Der Vatter von* $t_k)_s = a$ iff t_k is a term, and there is b such that $b = \text{Den}(t_k)_s$, and '*Der Vatter von*' is **fulfilled** by $\langle a,b\rangle$

(ii) Den(*Der Bruder von* $t_k)_s = a$ iff t_k is a term, and there is a b such that $\text{Den}(t_k)_s = b$, and '*Der Bruder von*' is **fulfilled** by $\langle a,b\rangle$

II A sequence *s* satisfies a sentential function *f* of L_3

(a) If f is of the form tP, where P is a 1-place predicate and t is a term, then: s satisfies f iff 'P' **applies to** $\text{Den}(t)_s$;

(b) If f is of the form t_kRt_k, where R is a 2-place predicate and t_i, t_k terms, then: s satisfies f iff 'R' **applies to** $\langle \text{Den}(t_i)_s , \text{Den}(t_k)_s\rangle$;

.....

Again, the recursive part of the definition of sequence-relative satisfaction for non-atomic sentential functions, and, in terms of it, of truth for sentences, is the same as in (D6) for L_2.

B-variant definition looks familiar, but a few comments are in order. Since we deal with sentential functions, it is convenient to define sequence-relative denotation in a uniform manner both for open and closed terms, and on this basis to define recursively sequence-relative satisfaction (as in the model-theoretic account we first define denotation (or value) for each term w.r.t. assignment of values to variables – taken from the domain of a structure – and then employ it in the recursive definition of satisfaction for formulas). As for the sequence-relative denotation of the k-th variable, things are exactly as in A-variant: it is the k-th term of s. The main difference, compared to A-variant, is

that the denotation of a name is not directly specified, since, according to Field's view of the matter, what bearer the name actually has "depends on the facts we have not yet been given" about L_3 and its usage by L_3-speakers.[262] We are to say, Field says, that the name denotes what it denotes. The same goes, *mutatis mutandis*, for predicates (each is said to apply to objects that it in fact applies to), and function symbols (each is said to be fulfilled by pairs of objects that in fact fulfil it). The rationale for this move is the same: we have not yet specified the facts (under their physicalistic descriptions) in virtue of which such expressions possess the semantic features they in fact possess.

Our choice of L_3 is somewhat unfortunate in this respect, since we all very well know what bearers the three names occurring in it have (the same holds of application- and fulfilment conditions of predicates and term-forming expressions respectively). But Field's main point applies even in this case: that we know what 'Angela Merkel' denotes in our German fragment does not mean that we are in possession of a genuine explanation as to what facts make it the case that that sequence of sings denotes something, and that it has exactly that particular denotation in German speaking community.

Field's exposition does not suffer from this defect, because he proceeds schematically. Instead of listing expressions of a particular language, he uses indexed letters for names ('c_k'), predicates ('p_k'), and function symbols ('f_k') (we proceeded similarly in giving our schematic description of the general model-theoretic truth-definition). Officially, he proceeds as if devising the truth definition for a given interpreted language L, construing indexed letters as different non-logical constants of various types. But the procedure is in fact highly schematic so that nothing prevents us from viewing it as a general framework applicable to any given interpreted language of 1st order type containing basic semantic categories of names, *n*-place predicates and *n*-place function-symbols. Field suggests the following generalization that makes B-variant independent on the lexicon of a particular language at a particular temporal stage:

1. Den(*k*-th variable)$_s$ = s_k;

2. If e_i is a name, Den(e_i)$_s$ is what e_i denotes;

3. If e_i is a singular term and e_k is a function symbol, Den(e_k(e_i)$_s$) = a iff

   (i) there is b such that Den(e_i)$_s$ = b

and (ii) e_k is fulfilled by $\langle a, b\rangle$

.....

Sequence-relative satisfaction can be accounted for a similar way. B-variant truth definition thus does not suffer from the limitations of A-variant. Moreover, Field says, B-variant, as opposed to A-variant, can accommodate context-sensitivity, being applicable to sentence-tokens.

Let us now compare A-variant and B-variant to get a better grip on the

[262] Field (1972: 349).

question of which of them fares better in light of Tarski's goals. One difference should be obvious. The first, but not the second, takes full advantage of the fact that we know meta-linguistic translations of names, predicates and function symbols of L_3 to specify the denotation conditions of its terms and satisfaction-conditions of its predicates, thereby saving any appeal to the notions of *nominal denotation*, *predicative application* or *functional fulfilment*, that feature essentially – indeed, non-eliminably - in Field's B-variant. A-variant is recursive in nature; it does not by itself eliminate semantic notions from every sentential context of the metalanguage. But we have seen that such elimination can be effected in a higher-order language (or assuming richer set-theory allowing quantification over all subsets of the domain of the object-language). B-variant might also be turned to an explicit definition that eliminates the notions of sequence-relative satisfaction and denotation. It is important to keep in mind, though, that it does not eliminate the three semantic notions of primitive denotation (which, for this reason, I have written in bold letters).

It was, of course, vital to Tarski's project of providing method of constructing adequate truth definitions that they do not contain any ineliminable semantic notions. This was not just because of formal correctness but because of material adequacy as well. B-variant (even turned explicit) is thus not of the same interest as A-variant. The recursive A-variant does not allow us to eliminate semantic notions from every context of the meta-language, but its power dwells in the fact that it allows us to derive all T-biconditionals for L_3, as well as all biconditionals that serve as conditions of adequacy on definitions of denotation and satisfaction. Recall that the right-hand sides of such biconditionals do not contain any semantic garbage! In this limited sense at least, even the recursive A-variant is kind of eliminative, since such biconditionals are paradigms in which applications conditions of a semantic notion with respect to a particular expression are explained in non-semantic terms (provided that the object-language does not contain semantic notions). No such biconditional is a consequence of B-variant, even when it is turned to an explicit form; its consequences always contain some semantic garbage on their right-hand sides, as witnessed by the following examples:

> *s* satisfies 'Helmut Kohl ist ein Mann' iff 'ist ein Mann' **applies to** DEN('Helmut Kohl')$_{s.}$ iff 'ist ein Mann' **applies to** what 'Helmut Kohl' denotes;
>
> *s* satisfies 'Helmut Kohl ist ein Mann' iff 'ist ein Mann' **applies to** what 'Helmut Kohl' **denotes**.

It should be obvious that B-variant, as it stands, is not materially adequate, and does not satisfy Tarski's demands; in particular, we have no criterion of its extensional correctness.[263] One may also worry, with Tarski, that because it contains unreduced semantic notions, it is problematic. This, however, is not the issue between Tarski and Field, Field concedes that this aspect makes his favourite variant of truth definition only partially satisfactory. Before turning to this problem, let me mention one possible objection to the effect that not even A-variant truth definition is completely free of semantic notions. It assumes, so

[263] The point made forcefully by McDowell (1978).

the virtual objector say, the notion of translation that arguably involves semantic elements, at minimum, the element of *preservation of meaning*, since Tarskian truth-definition for L in ML is adequate only *modulo* a correct translation of L into ML, not *modulo* any translation. Thus, when dealing with interpreted languages in use, we are not free to stipulate the meanings of their expressions via an arbitrary "translation-manual" from L to ML.[264] To be sure, we can define a function whose domain is the set of L-expressions that takes values from the set of ML-expressions (so that each L-name is mapped to an ML-name, 1-place predicate to a 1-place predicate, etc.), and call the function a *translation of L into ML.*[265] However, there will be many such functions, but not every will do when we want to give a correct truth definition of L in ML.

For instance, we would offer a grossly incorrect definition of sequence-relative satisfaction for our tiny fragment of German, if our translation manual licensed the following clauses:

> *s* satisfies 'x_1 ist eine Frau' iff s_1 is a cat;
> *s* satisfies 'x_2 ist ein Mann' iff s_2 is a robot;
> *s* satisfies 'x_1 liebes x_2' iff $s_1$hates s_2.

Tarski did not bother to explain what he understood under *(correct) translation*; he tacitly assumed that it is possible, that *the* (or, perhaps, *a*) correct translation of L into ML can be settled. But Field suggests the adequacy criterion:

> "An adequate translation of a primitive e_1 of L into English is an expression e_2 of English such that
>
> (i) $e_1$and e_2 are coreferential, and
>
> (ii) e_2 contains no semantic terms," (Field 1972: 355)

where two expressions are coreferential just in case they have the same extension. The clause (i) concerns material correctness of translations and it spells it out, modestly enough, as preservation of extensional meaning; the clause (ii) concerns formal adequacy and is designed to block question-begging translations of the type

> "Helmut Kohl" → "What 'Helmut Kohl' denotes",

that would reintroduce into the definition unreduced semantic notions. Now, the clause (i) reveals the semantic character of translation, containing as it does the notion of coreferentiality. The question arises whether Tarski's A-variant truth definition really eliminates all semantic notions. Field considers this objection – by the way, a very popular one - but he disposes of it in the following way.

---

[264] Note that our intuitive notion of translation - involving as it does the notion of (at least partial) meaning preservation - does not support talk about translating uninterpreted languages into other languages - interpreted or not - since, in such cases, there is nothing to be preserved (not even partially). Incidentally, that is one reason why material adequacy criterion spelled out in Convention T does not make sense for relative truth definitions for uninterpreted languages.

[265] This sort of a formal-syntactic translation is perfectly applicable to formal-uninterpreted languages; once a translation function is well-defined for such a language, there is no further question of correctness or incorrectness.

Admittedly, the notion of correct translation (preserving at least extensional meaning) is a part of Tarski's methodology or meta-metatheory (invoked in Convention T), but there is no trace of it in Tarski's particular truth definitions.[266]

For all that I have said up to this point, we seem to have little reason to prefer B-variant to A-variant, but we seem to have some good reasons to prefer A-variant to B-variant. Only with A-variant we have a criterion that we got things right (viz. material adequacy in Convention T); secondly, only A-variant allows us to eliminate semantic terms (at least from the metatheory if not from the meta-metatheory), as soon as it is turned to an explicit form. Why, in spite of this, does Field think that Tarski would have done better to give B-variant, and not A-variant that he actually gave, which is both materially adequate (hence extensionally correct) and does not contain any unreduced semantic notions in its wake (after being turned explicit)?

To answer this, we should first observe that Field approvingly quotes Tarski's contention to the effect that an adequate truth definition should not invoke any unreduced semantic notions in the definiens:

> "We desire semantic terms (referring to the object language) to be introduced into the meta-language only by definition. For, if this postulate is satisfied, the definition of truth, or of any other semantic concept will fulfil what we intuitively expect from every definition; that is, it will explain the meaning of the term being defined in terms whose meaning appears to be completely clear and unequivocal." (Tarski 1944. 351)

On this basis, he formulates the adequacy criterion that, he thinks, would have been approved also by Tarski:[267]

> "M) Any condition of the form
>
> (2) $\ldots\forall e[e \text{ is true} \equiv B(e)]$

[266] We shall see that Field sees a problem here. To anticipate his worries, could one presuppose, without further ado, a semantically loaded notion of translation when one's programme is to reduce semantic notions to the ideology acceptable to physicalists? Well, perhaps there is a way of specifying what correct translation amounts to that turns out to be non-semantic, but it is just not clear what it is. Another alternative could be to work with homophonic translation and define truth only for an object-language that is a proper part of the metalanguage. This, however, would seriously restrict the application of Tarski's method.

[267] McDowell (1978) shows that there is a subtle yet important incorrectness in this thought. Tarski's adequacy criterion is of course Convention T, not Criterion M. The fact that Field attributes the later to Tarski is a further evidence that he is under-impressed by the desideratum of material adequacy, which, however, is all-important for Tarski. In the footnote where Field comments on the connection of Convention T with Criterion M he downplays the role of Convention T by saying that its only function is instrumental with respect to Criterion M – assuring that the truth definition is extensionally correct. It is doubtful, however, whether this is a correct diagnosis of Tarski's position. Admittedly, a part of the appeal of Convention T is that if a truth definition satisfies it, it is extensionally correct. But we have had an occasion to see that Tarski likely believed that Convention T captures something important about the very concept of truth, something closely connected with T-biconditionals qua "partial definitions" of the concept of truth.

> should be accepted as an adequate definition of truth if and only if it is correct and '*B*(*e*)' is a well-formed formula containing no semantic terms." (Field 1972: 361)

Under "correctness" he understands extensional correctness and claims that it is all right to demand that a definition of truth deliver an extensionally correct formula free of semantic terms. Consequently, he must take his own favourite B-variant to be inadequate (at least partially):

> "It [namely the criterion M] rules out the possibility of T1 [namely B-variant] by itself being an adequate truth definition; and it is right to do so, if the task of a truth definition is to reduce truth to non-semantic terms, for T1 [namely A-variant] provides only a partial reduction." (Ibid: 362)

Prima facie at least, A-variant (when turned explicit) fares considerably better in this respect. Why, once again, does Field think B-variant to be superior to A-variant, given that *extensional correctness + elimination of semantic terms* seems to be satisfied by the later but not by the former? Well, Field does not endorse Criterion M *in toto*, but only as a necessary condition on an adequate truth definition. The core of his argument is that extensional correctness (to which material adequacy is only instrumental according to him) plus elimination of semantic notions from the *definiens* is only necessary but not sufficient for *genuine reduction* of truth in particular and semantic notions in general. Granted, certain idiosyncratic features allow A-variant to eliminate semantic terms, but Field thinks he can show that these very features are responsible for is being philosophically cheap and uninteresting. Elimination of semantic terms in A-variant only masquerades as a genuine reduction of semantic to non-semantic, and, ultimately, to physicalistic properties.

Field's plea for B-variant is motivated by his view that the goal of a *genuinely* scientific semantics splits into two sub-goals, of which only one was successfully accomplished by Tarski: namely, showing how to construct a compositional semantics for a particular (1st-order) language, which reduces truth-relevant semantic features of complex expressions to those of primitive expressions (based on 1st-order syntax). Such an account may be divided into three parts:

(a) classifying primitive expressions of a language into basic (logical and non-logical) categories;

(b) assigning each basic non-logical category of expressions a semantic property of a type appropriate to it (denotation-conditions for names, application-conditions for predicates, fulfilment conditions for function symbols, etc.);

(c) laying down compositional (recursive) rules determining the semantic properties of complex expressions, given their syntactic (logical) form, fixed meanings of logical constants and semantic properties of simple expressions forming them.

With B-variant truth definition, Field says, this part of the agenda of scientific semantics is successfully finished, at least for 1st-order languages (or languages formalizable in this manner). The generality attained in B-variant truth definition may remind us of the model-theoretic framework, except that there is no parameter for structure buried in Field's definition, as object-languages are fully interpreted (have their intended interpretations). Nevertheless, the parallel is striking. In the model-theoretic account we have a general skeleton and we flesh it out by applying it to a particular 1st-order uninterpreted language L with a fixed signature and an L-structure, thereby obtaining a particular definition of truth for L-sentences relative to that L-structure. B-variant is a general skeleton and we flesh it out by applying it to a given interpreted L (in which we can discern the relevant 1st-order structure and reckon its words to right categories). We thus obtain a particular truth definition for L in B-variant style. The difference is that while the first procedure starts with an uninterpreted L and then specifies its interpretation (and defines truth relative to it), the second starts with an interpreted L, but it leaves it open what interpretation it possesses, because it leaves it open what interpretations its primitive expressions possess (for the programmatic reasons spelled out above).

Field is well aware of the parallel. B-variant truth definition with its unreduced semantic notions is useful, since it allows us to deal with typically model-theoretic questions concerning what happens to the truth-value of a sentence (set of sentences) when we conceive of its non-logical primitives as changing their interpretations (relative to various domains attached to quantifiers), the interpretations of logical constants being fixed. Perhaps the chief purpose of the model-theoretic semantics for 1st-order or other languages is to formulate the principles of compositional assignment of truth-relevant features to an indefinite number of expressions of various categories and explicate on this basis the notions of logical consequence, validity, (un)satisfiability, etc. Semanticists ask what meanings do, how the meanings of simple words compose in various truth-relevant ways into still larger wholes, attempting to find right types of set- or function-theoretic entities that model these roles (be they extensional, intensional or hyper-intensional). Here are two representative passages from authors widely known for their work in model-theoretic semantics (interestingly, the first author quotes explicitly Field's diagnosis of the matter):[268]

[268] I first learned about these connections from LePore's article (1983), where he offers an interesting critique of the model-theoretic approach (indirectly of Field's favourite approach), defending the truth-theoretic approach of Davidson. It is unsurprising that the model-theoretic semantics shows little interest in the question of what facts, if any, make it the case that simple expressions acquire the interpretations they do, or which among the many possible interpretations of a language is the standard or intended one (or epistemologically speaking: how can we know what the interpretation of it is). Indeed, a view seems widely held that it is not the business of semantics in the model-theoretic style to come up with such a story. Interpretations can be thought of as attached to words, say, in Tarski's pairing manner or via interpreting functions. The nature of the connection obtaining between words and their actual interpretations is not their concern, however interesting issues it may raise. Semanticists work at the theoretical level, where such foundational questions do not yet arise, taking to heart the advice of Carnap (1942), who proposed to distinguish pure (abstract) from descriptive (empirical) semantics.

> "The real work of the truth definition and similarly for a Montague-style possible world semantics, comes in the specifications of how the interpretations of the infinite set of sentences can be determined by a finite set of rules from the interpretations of the primitives." (Partee 1977: 321-22).

> "A central goal of (semantics) is to explain how different kinds of meanings attach to different syntactic categories; another is to explain how the meanings of phrases depend on those of their components
>
> .... But we should not expect a semantic theory to furnish an account of how any two expressions belonging to the same syntactic category differ in meaning. "Walk" and "run," for instance, and "unicorn" and "zebra" certainly do differ in meaning, and we require a dictionary of English to tell us how. But the making of a dictionary demands considerable knowledge of the word." (Thomason 1974: 48-9).

In a similar vein, B-variant truth definition, if properly understood, explains how truth-conditions and satisfaction-conditions of sentences and predicates of a 1st-order language can be compositionally specified on the basis of denotations of names, application-conditions of simple predicates and fulfilment-conditions of function symbols. What it does not explain – and does not pretend to explain - is what it is for primitives of the language to have the semantic powers they actually have. Such full-blooded semantics should tell us not only how complex expressions - most importantly, sentences - compositionally depend for their representational powers on semantic powers of their simpler components and their syntax; it owes us some explanatory story as to what physical or at least physically based facts (of a causal-social-historical breed) confer on expressions representational powers that they possess. Presumably, it is facts about linguistic practices and habits of a community situated in a historico-physical environment that determine what semantic properties its expressions possess. Without such a naturalistic story, we are puzzled about the nature of semantic properties, hence about the nature of our language.

Puzzling they are, as many distinguished thinkers were painfully aware long before, asking questions such as how it is possible to represent physically absent or remote things, etc. Indeed, so puzzling they appear, if no explanation is available in scientific terms, that sceptical voices might propose to eliminate them altogether from scientific language (in fact, semantical eliminativism was a strategy very much in fashion among some members of Vienna circle; that Quine toyed with it at times is also well known):

> "But how could we ever explicate in non-semantic terms the alleged fact that these utterances are true? Part of the explication of the truth of "Schnee ist weiss und Grass is grün" presumably, would be that snow is white and grass is green. But this would only be part of the explanation, for still missing is the connection between snow being white and grass being green on the one hand, and the German

> utterance being true on the other hand. It is this connection that seems so difficult to explicate in a way that would satisfy a physicalist, i.e., in a way that does not involve the use of semantic terms." (Field 1972: 359-260).

On the one hand, semantic talk is not so easy to abandon, since it has a useful role to play in our linguistic practices. Thus, for instance, the notion of truth enables us to form beliefs about reality on evidence of reports of our fellow speakers. The pattern of reasoning is as follows:

> I have no evidence whether or not P is the case, but then a reliable member of my community tells me that P is the case. Believing that what he/she says is *true*, I form myself the belief that P is the case.

Other considerations speaking in favour of semantic properties could be made. Why do we aim at having true rather than false beliefs? Because possession of true, as opposed to false beliefs increases our chances to achieve our ends. And how are we to explain the point of scientific inquiry and its predictive success, if not by saying that it aims to discover what is true about reality – to cut nature at its joints?

On the other hand, Field rightly points out that there is no remotely plausible account available of semantic properties reducing them to scientific ideology. In view of this, some thinkers who take seriously the considerations in favour of semantic properties propose to treat them as irreducible properties *sui generis* (semantic facts – e.g. the fact that "Helmut Kohl" denotes the person that it actually denotes, namely, Helmut Kohl – being brute facts), rather than abandoning them altogether. This position is called by Field *semanticalism*:

> "This doctrine, [which] might be called 'semanticalism', is the doctrine that there are irreducibly semantic facts. The semanticalist claims, in other words, that semantic phenomena (such as the fact that 'Schnee' refers to snow) must be accepted as primitive, in precisely the way that electromagnetic phenomena are accepted as primitive (by those who accept Maxwell's equations and reject the ether); and in precisely the way that biological and mental phenomena are accepted as primitive by vitalists and Cartesians. Semanticalism, like Cartesianism and vitalism, posits nonphysical primitives, and as a physicalist I believe that all three doctrines must be rejected." (Ibid: 358)

The opposite position is occupied by semantic eliminativists, whose conclusion is: common sense and intuitions aside, science has the last word, and if it turns out impossible to reduce semantic properties to physicalist ideology, we better abandon them (that is not to say that a plain man is discouraged from using semantic terms in ordinary life).

Field finds neither of the two extreme positions particularly attractive but agrees with the eliminativists that if it turned out to be impossible to reduce semantic notions to scientific ideology, we better abandon them. Fortunately, the

situation is not as hopeless as semantic eliminativists would have us believe. In fact, the recursive machinery of B-variant truth-definitions is a promising start, since it "reduces" semantic properties of complex expressions to those of primitives. It remains to give a plausible, general explanation as to what broadly physical facts confer on non-logical primitives (of this or that semantic category) their specific semantic properties (e.g.: what facts are responsible for the fact that 'Helmut Kohl' denotes in the mouths of speakers of German (L_3) the person it actually denotes). In virtue of what facts do they posses such specific properties and not others.

The problem here is not epistemological but metaphysical, in the sense of the question: how do facts about meaning and semantic properties depend on (are determined by) facts about usage and environment (physical facts)? Such questions are called metasemantical or foundational. It is questions with which Quine, Davidson, Kripke, Lewis and others wrestled when discussing the problem of indeterminacy of (radical) translation (interpretation), inscrutability of reference, rule-following, actual-language relation, etc. Quine, Davidson and Lewis are proponents of semantic holism in that they claim the priority of sentences or utterances or beliefs (or totalities of sentences, utterances or beliefs) over sub-sentential expressions, when it comes to foundational questions (this is closely connected to their interpretivist methodology). Field, on the other hand, propounds a version of semantic atomism: it is primitive denotation of simple words where language and world get into contact – where words and language ultimately acquire their representational powers.

This is claimed by Field to be the grain of truth in correspondence conceptions of truth, though they were imprecise and typically invoked the problematic category of facts (states of affairs). Field champions an atomistic, object-based correspondence theory, which does not invoke facts or states of affairs (though it does not deny their existence either, and is perhaps compatible with them) but explains truth-conditions of sentences in terms of narrowly semantic notions of primitive denotation.[269] Now, once Tarski finished the compositional part, semantics must explain, in physicalistic terms, what primitive denotation is. To Field, the most promising research strategy is to seek a naturalistic theory of primitive denotation, along the lines initiated by Kripke (and Putnam and Donnelan – the grandfathers of the so-called *new theory of reference*) in his "picture" of causal-historical character of reference for singular and/or natural kind terms. Eventually, one may go further, as Field proposed in other articles, urging a kind of language of thought hypothesis to the effect that representational properties of linguistic expressions derive from representational properties of mental states, and ultimately, from representational powers of inner sentence-like tokens, which, it is hoped, can be explained naturalistically (perhaps combining some elements of conceptual-role semantics with elements of informational-causal semantics). To think of semantics along these lines might look like a promising strategy, if one is after a reductive theory that attempt to explain semantic properties of a language by showing them to be reducible to or supervening on facts about linguistic practices and environment.

What Field has in mind is neatly illustrated by his own example of

[269] See McDowell (1978) for a good discussion. Against Field, he urges kind of non-reductionist semantical holism inspired by Davidson.

chemical valence. A chemical valence is an integer assigned to a chemical element indicating its combinatory potential with respect to other elements, in terms of which the chemists explain what elements combine with what elements in what proportions. Now, the analogy is calculated to cast light on primitive denotation is this: also configurations of elements (radicals) have valences, which, however, are determined by valences of elements that make them up together with their structure. Not so, however, valences of chemical elements themselves, since elements are not made up of other elements that have valences (to be sure, they are made of atoms, which in turn are made of ..., etc., but these elements do not themselves possess valences). Field envisages a recursive definition of *valence* resting on the classification of structures of possible configurations of chemical elements (analogy: the syntactic description of a language) and characterizing valences of structurally complex configurations of elements in terms of valences of less complex configurations making them up, down to the valences of simple elements (analogy: the definition of satisfaction, by recursion on the logical-syntactical complexity of a formula). We might turn such a recursive definition to an explicit form

> *c* has valence *n* iff *B*(*c*; *n*),

which explains valences of configurations in terms of valences of their elements based on their structure, but still contains the notion of valence as attached to elements. What about them, now? Do they not call just as much for a genuine reduction? Let us see if it would be enough to supply the following basis for recursion (it being assumed that a correct valence is assigned to every existing element):

> (∀*E*)(∀*n*)(*E* has valence *n* iff *E* is potassium and *n* is +1, or ..., or *E* is sulphur and *n* is -2) ?

It is obvious that this elimination of valence is a pseudo-reduction by enumeration, albeit it is extensionally correct. If scientists could not provide anything better by way of its explanation, the notion of valence would have to have no place in serious science and should be dropped from it. Fortunately, it was discovered that valences of elements reduce to certain structural properties of atoms, and its use in science was vindicated.

Much as the recursive definition of valence reduces valences of configurations of elements to valences of elements (based on their structure), recursive clauses in B-variant truth definition reduced semantic features of complex expressions to semantic features of primitive expressions (based on their syntax). What about these primitive semantic properties? Are we to consider them irreducibly primitive? Once again, Field claims that this would go against the naturalistic stance of science, and it would be the grist on the semantic eliminativist's mill.

As before, we seem to have two alternatives. One is to eliminate semantic notions via trivially-looking clauses:

> A) 'Günter Grass' denotes Günter Grass;

‘Angela Merkel’ denotes Angela Merkel’
‘Helmut Kohl’ denotes Helmut Kohl;

B) ‘Der Vater von’ is fulfilled by ⟨*a*,*b*⟩ iff *a* is the father of *b;*
‘Der Bruder von’ is fulfilled by ⟨*a*,*b*⟩ iff *a* is the brother of *b;*

C) ‘ist eine Frau’ applies to *a* iff *a* is a woman;
‘ist ein Mann’ applies to *a* iff *a* is a man;
‘liebes’ applies to ⟨*a*,*b*⟩ iff *a* loves *b*.[270]

Tarski’s A-variant truth definition pursues this strategy. If Field’s analogy is on the right track, this is a pseudo-reduction as trivial as elimination of valence by means of the enumerative definition. Pursuing this strategy, one has done nothing to answer foundational questions. Despite delusive appearances, Tarski did not *explain* primitive denotation by providing his favourite A-variant truth definition.

But if there is anything philosophically really interesting in semantics, it is the question how language hooks on the world, how or in virtue of what facts the first can express various things about the second that are true or false, as the case may be. Despite misguided remarks of Tarski, he left us as puzzled regarding such matters, as we were before. Observe that by adding (A), (B) and (C) to Field’s preferred B-variant truth definition, we get a definition equivalent to A-variant: we can (a) deduce from so extended B-variant truth definition all T-biconditionals, and (b) all semantic notions, including primitive denotation, can be eliminated from it, once we turn it to an explicit form. However, we have just seen that elimination achieved in this way is fairly cheap, not deserving the appellation of a genuine reduction. Field concludes that the so extended B-variant has no interest beyond what the original B-variant states, hence A-variant has no interest beyond it either, as it is just B-variant plus the trivializing clauses (A), (B) and (C). Tarski made no doubt one important step towards explaining semantics and the perennial problem of the relation of language and reality. But perhaps the more important step is yet to be made: a general, reductive explanation of primitive reference, which would be physicalistic (naturalistic) in spirit. Tarski could circumvent unreduced notions belonging to the circle of primitive denotation because he assumed that the list-like definitions along the lines of (A), (B) and (C) are available. These definitions were available, because he took for granted that a correct translation (homophonic or heterophonic) of object-language (L_3) into the meta-language (English) is available to us, as the definition-constructors. We are familiar with the denotations of names,

---

[270] Equivalently:
(A*) *n* denotes *a* iff (*n* = ‘Günter Grass’ and *a* = Günter Grass) or (*n* = ‘Angela Merkel’ and *a* = ‘Angela Merkel’) or (*n* = ‘Helmut Kohl’ and *a* = Helmut Kohl;

(B*) (i) *P* applies to *a* iff (*P* = ‘ist eine Frau’ and *a* is a woman) or (*P* = ‘ist ein Mann’ and *a* is a man) or (*P* = ‘liebes’ and *a* loves *b*);
(ii) *P* applies to ⟨*a*, *b*⟩ iff *P* = ‘liebes’ and *a* loves *b*;

(C*) *f* is fulfilled by ⟨*a*, *b*⟩ iff (*f* = ‘Der Vater von’ and *a* is the father of *b*) or (*f* = ‘Der Bruder von’ and *a* is the brother of *b*).

fulfilment-conditions of functional expressions and application-conditions of predicates of the object-language, because we know how they translate into ML, in which we are competent. And we know, by disquotation, the denotations of names, fulfilment-conditions of functional expressions and application-conditions of predicates that are their translations in his (home) ML. Proceeding in this manner, Tarski swept under the carpet all (or almost all) philosophically interesting questions that one may raise regarding meaning, translation and semantic properties in general.

The second strategy preferred by Field is different. Suppose that we had a physicalistic theory on which primitive semantic properties of names, predicates and function symbols supervene on complex physical properties D, A, and F respectively. Such properties would apply to any name, predicate or function-symbol whatever in any language containing such expressions. But once we apply the general physicalistic theory of nominal denotation, predicative application and functional fulfilment to a given language with the right structure, say to L_3, it has consequences of the following sort:

(a) The fact that a name *n* denotes what it denotes supervenes on *n*'s instantiating D (its instantiating D explain why *n* denotes what it denotes);

(b) The fact that a simple predicate *P* applies to what it applies to supervenes on *P*'s instantiating A (*P*'s instantiating A explains why *P* applies to what it applies to),

(c) The fact that a functional-expression *f* is fulfilled by what it is fulfilled by supervenes on *f*'s instantiating F (*f*'s instantiating F explains why *f* is fulfilled by what it is fulfilled.

Supplementing the recursive B-variant truth definition for L_3 with the applied theory, we carry out a thorough reduction of semantic notions to non-semantic notions. In this way we could hope to provide a scientific semantics for L_3.

7.2 In defence of Tarski's approach: a division of theoretical labour

Field notes that Tarski's strategy was to introduce semantic notions into metalanguage through explicit definitions that contain no semantic notions that could not be themselves explicitly defined in purely non-semantic terms, so that they are eliminable in principle from any sentential context of the metalanguage. Now, of course, B-variant truth definition, with its unreduced notions of primitive denotation, violates this desideratum blatantly. Where Field went astray is in supposing that Tarski's only or principal motivation for this desideratum was his desire to show that scientific semantics is possible in the sense of being incorporable into the unified scientific outlook by reducing (via definitions) semantic notions to physicalistic ideology. Now, there is a trace of this motive in Tarski's work, but only at one place in his published writings (in the lecture directed at the positivist auditorium). In view of this, it seems that Field attaches too much weight to this motive and fails to emphasize other, arguably more important motives Tarski had for the desideratum, better

documented in his works.

Tarski was not confused to choose A-variant truth definition. He wanted to give a definition entailing all T-biconditionals for the object-language with quantificational structure. Field admits, in a way, that it was no aberration on Tarski's part that he chose A-variant with its enumerative base-clauses, and not B-variant, employing unreduced semantic notions of primitive denotation. But he muddies the water twice. First, when he does not attach due weight to the all important goal of satisfying Convention T (in fact, he tends to downplay it); second, when he attaches too much weight to what he takes to be Tarski's primary aim in CTFL, namely the reduction of semantic notions to the physicalist basis. Another motive that he omits to give a due weight to is that only with the help of eliminative definitions could Tarski hope to persuade his mathematical fellows that the part of metamathematics that concerned the semantics for logico-mathematical languages can be conducted in entirely logico-mathematical terms (syntax – theory of concatenation - can also be interpreted in mathematics) and using mathematical techniques, hence that it can itself be treated as a logico-mathematical discipline. The second motive could have been the principal one, as mathematical logicians worked informally yet safely with semantic notions. Moreover, the clue to Tarski's method of constructing consistent semantics is his distinction between a semantically open object-language L not containing its own semantic predicates and a richer meta-language ML containing the semantic predicates of L, though not its own – being semantically open as well.

The eliminative part is thus not the heart of Tarski's method of truth-definition: except for the programmatic reduction to (interpretation in) logico-mathematics (set theory), all one might want from semantics is already at the recursive level and works well there (and, as Tarski himself said, more intuitively than at the explicit level). The principled distinction between L and ML, on the other hand, gives us a well-based hope that the method will not involve us in semantic paradox. For it is the recursive machinery that allows us to derive fundamental principles governing truth such as excluded middle, non-contradiction, etc.; moreover, it allows us not only to state but prove fundamental theorems concerning the question of completeness and consistency of the class of theorems of the deductive theory T in L and L-truths respectively, as well as theorems concerning the relation between the two classes (is the first included in the first and/or *vice versa*?). The role of the explicit part of the method consists in the fact that it provides us with a kind of formal assurance that the method does its work without involving us in a paradox. And once we have formally assured ourselves that the method works, we have thereby proved that the recursive definition (which is more intuitive than the explicit) is itself perfectly in order.

Putnam and Field agree that Tarski's definitions tell us nothing particularly illuminating about the question of how expressions (or languages) owe their semantic properties to linguistic practices in social-physical environment. Curiously enough, Putnam calls Tarski's theory of truth "non-semantical', because of *reducing* semantic terms to non-semantic (logical, syntactical, mathematical) terms. Field, on the other hand, blames it for not reducing the semantic notions in a proper manner to the physicalist ideology. But there is no conflict between them, since both can recognize two senses of

reduction. In a sense, Tarski succeeded in reducing semantics to non-semantics, since he showed us how to eliminate semantic notions to a non-semantic basis via explicit meta-theoretical definitions that are extensionally adequate. In a sense, he did not succeed in reducing semantics to non-semantics, because his definitions have, at bottom, a trivially list-like or enumerative character, hence fail to connect semantic notions in any relevant way to the facts about meaning and, in last instance, to the facts about usage, on which all semantic properties presumably depend (supervening on them).

Field's and Putnam's worries are interesting in their own right, challenging thinkers with deflationist tendencies – to be discussed in the next chapter - who claim that Tarski's truth definitions tell us (nearly) everything there is to know about truth and semantic in general.[271] Still, it is very much open to doubt whether Tarski wanted to carry out a robust reduction of semantics to non-semantics, and hence whether he is to be blamed for not carrying it out successfully. One of the main theses of this work is that Tarski's main contribution was to separate formal-semantic from meta-semantic questions. One of those who urge this view is Scott Soames, whose work on Tarski's theory of truth was in part a polemical reaction to Field's article.[272]

According to Soames, a Tarskian language is an abstract (typically 1st-order) language represented by an ordered triple $\langle S_L, D_L, I_L \rangle$, where:

> (a) S_L is a set of disjoint sets of basic syntactic categories of expressions of L (it may include a set of L-terms, a set of 1-place L-predicates, a set of 2-place L-predicates, etc.);
>
> (b) D_L is the domain of individuals associated with L;
>
> (c) I_L is an interpretation function (mapping).

I_L assigns each expression belonging to one of the basic categories in S_L a certain set-theoretical entity of appropriate type as its interpretation in the usual model-theoretic style: D_L to quantifiers, individuals of D_L to terms, subsets of D_L to 1-place predicates, sets of *n*-tuples on D_L to *n*-place predicates, etc.

Soames attributes this conception of abstract languages to Lewis and Kripke. And, of course, it is reminiscent of the model-theoretic account. The main difference is that in the model-theoretic semantics we start with an uninterpreted 1st-order language L (or with a class of such languages with the same logical structure) and then give for it a Tarski-style definition of *truth w.r.t. M*, where *M* is represented by the ordered pair $\langle D, I \rangle$, *D* and *I* being explained in the same way as we have done above. Soamsian languages, on the other hand, are semantically individuated languages. What he urges is a purely abstract-mathematical perspective *à la* Carnap, minus Carnap's (alleged) confusion of assuming semantical rules to do the double duty of defining semantic notions for L and interpreting L-expressions. Soames takes to heart Tarski's lesson that for

---

[271] Viz. Leeds (1978) or Horwich (1984). Field has become a deflationist over years, but he does not show tendency to claim that Tarski said nearly everything there is to say about truth; rather, he tends to say that Tarski said more about truth that there is to say – having in mind the compositional clauses of Tarski's truth definitions.

[272] Soames (1984).

definitions of semantic notion to be so much as materially adequate – that is, satisfying Convention T and related conventions for satisfaction and/or denotation – there must be something with respect to which they are judged adequate or not. In particular, the conventions that supply adequacy-criteria require that certain sentences (expressions) of ML correctly translate certain sentences (expressions) of L. But we can hardly talk about correct translation when L is an uninterpreted language to be interpreted in ML.

Like Church, Soames does not deny that it is possible to interpret L in Carnap's style, but claims that it does not make sense to say that we thereby define semantic notions in Tarski style. Rather, we are employing semantic notions as already well-understood primitives, and by stipulating the conditions under which they apply to L-expressions of various types we interpret those L-expressions themselves in such a way that, in the end, every L-sentence is provided with its unique condition of truth. So construed, semantical rules are more like axioms than definitions. Since this means to take the notion of truth as already well-understood and more or less unproblematic, it could not be an option for Tarski. Soames proposes to interpret semantical rules as defining restricted semantic notions w.r.t. L in Tarski style, which requires that L is already interpreted via the interpretation function I_L built-in L = $\langle S_L, D_L, I_L\rangle$. Soames' abstract languages share with Carnapian languages the property that they posses their semantic properties essentially – as I_L is built-in L, we cannot change it in $\langle S_L, D_L, I_L\rangle$ without changing L for another language L*. So whereas the model-theoretic account defines *denotation*, *satisfaction* and *truth* for an uninterpreted L relative to *M*, Soames proposes to define truth for L that is semantically (model-theoretically) interpreted, so that we do not have to refer to *M* in its truth definition. This is the double way in which mathematicians operate with languages such as L(PA): sometimes, when it is clear from the context what they talk about (i.e. the standard model), they work with L(PA) as if already interpreted in the standard model and are happy not to mention the standard model at all, giving an absolute truth-definition in Tarski's CTFL style for it; other times, however, e.g. when they wish to be precise or want to discuss the problem of various interpretations of L(PA) including non-standard models, they make the reference to interpretation in the standard model explicit, there being a trace of it in the truth definition.

Soames' account is sketchy but it suggests to us at least two sorts of truth-definitions. Let L be a 1st-order language $\langle S_L, D_L, I_L\rangle$, where S_L contains the set T_L of terms {*m*; *n*} and the set P_L of 1-place predicates {*P*}, D_L is {New York; Chicago} and I_L is such that

$I_L(m)$ = New York
$I_L(n)$ = Chicago
$I_L(P) = \{x \in D_L: x \text{ is a city}\}$.

The interpretation function I_L is construed as a purely mathematical object: i.e. a function-in-extension, qua a set of ordered pairs of primitive expressions of L and set-theoretical entities of appropriate type defined on D_L. In a way, we can define I_L in a list-like manner:

(For every expression $e \in S_L$ and every object *o* such that $o \in D_L$ or $o \subseteq D_L$):

$I_L(e) = o$ iff e = 'm' and o = New York or e = 'n' and o = Chicago, or e = 'P' and $o = \{x \in D_L: x \text{ is a city}\}$.

Assuming the standard recursive definition of the set of $Sent_L$ of L-sentences (with the base clause covering atomic L-sentences of the form *Pr(t)* and the recursive clauses covering negations, conjunctions, and universally quantified L-sentences), we define absolute notions of denotation and truth for L in the following way (as both objects to be found in D_L have a name in L, we define truth for complex, including quantified L-sentences, inductively in terms of truth of less complex sentences):

(I) (For every term $t \in T_L$ and every object $a \in D_L$):

t denotes a (in L) ff $I_L(t) = a$;

(i.e. iff $t = m$ and o = New York, or $t = n$ and o = Chicago);

(II) (For every sentence $s \in Sent_L$): S is true (in L) iff

(i) $s = Pr(t)$, where $Pr \in P_L$ and $t \in T_L$, and $a \in I_L(Pr)$, for some a such that t denotes a (in L); or

(ii) $s = \neg A$, where $A \in Sent_L$, and it is not the case that A is true (in L); or

(iii) $s = A \wedge B$, where $A \in Sent_L$ and $B \in Sent_L$, and both A and B are true (in L); or

(iv) $s = \forall vA$, where $A \in Sent_L$, and $A(t/v)$ is true (in L), for any term $t \in T_L$.

We have not introduced the notion of predicative application, but it should be clear that it poses no special problems. We could define it in a parallel way:

(II) (For every predicate $Pr \in P_L$ and every object $a \in D_L$):

Pr applies to a (in L) iff $a \in I_L(t)$

(i.e. iff $Pr = P$ and $a \in \{x \in D_L: x \text{ is a city}\}$).

We can see that via the clauses directly listing the values of I_L function for non-logical primitives of L we can get rid of semantic notions, so that our truth definition is non-semantic. This feature enables us to deduce all T-biconditionals for L (*modulo* the metatheory containing logico-mathematics and syntax). This, I take it, is how Soames would define semantic notions of denotation and truth for a given interpreted L. What he explicitly mentions is a generalized definition of truth for a variable 'L', where every language in the range of 'L' belongs to the set J of such abstract languages that have a similar 1st-order structure. Since the truth-definition aims at generality, it lacks any language-specific clauses directly listing values of non-logical primitives of this

or that interpreted language belonging to *J*. Instead, it contains only language-unspecific clauses:

I* (For every language $L \in J$, term $t \in T_L$ and object $a \in D_L$):

t denotes a in L iff $I_L(t) = a$

II* (For every language $L \in J$ and sentence $s \in Sent_L$):

s is true in L iff

(i) $s = Pr(t)$, where $Pr \in P_L$ and $t \in T_L$, and $a \in I_L(Pr)$, for some a such that t denotes a in L; or

…....

As this general truth definition does not specify the interpretations of primitives, it does not license derivation of T-biconditionals for particular sentences of languages belonging to *J*. However, if we instantiate the generalized truth definition with respect to a particular language from *J* – say, the language dealt with above - what we need to derive T-biconditionals for its sentences is just the information about the interpretations of its primitives, supplied by the language-specific clauses. Soames' account has obvious parallels with Field's. It focuses on interpreted as opposed to uninterpreted languages. The generalized truth definition reminds us of Field's own generalized B-variant that abstracts from lexical idiosyncrasies of particular languages at particular temporal stages. Its instantiation with respect to a particular language L yields a definition close to what we would obtain applying the generalized B-variant to a particular language. Furthermore, once we add to the instantiated truth definition clauses specifying the interpretations of L's primitives, we get something close to Field's A-variant truth-definition, though Field would now complain that such clauses are trivializing.

Soames calls this framework 'Tarski's theory of truth'. However, had Tarski thought of formalized languages along these lines, he would have been well on the way towards the full-blooded model-theoretic account**.** As I argued that in CTFL he did not yet embrace the model-theoretic approach, I have to reject any suggestion to the effect that he would have endorsed this interpretation of his project, had he been confronted with it. The truth is that the generalized definition of Soames has no precursor in Tarski's CTFL, where he provides only a particular truth-definition for a particular language (and hints how to extend it to other languages from a large group), remarking that a generalized version of the method would be rather complicated. Moreover, though Soames' generalized truth definition deals with interpreted languages, I would hesitate to call it *absolute,* since it does not imply T-biconditionals for particular sentences. In order to generalize Tarski's case-by-case procedure, we have to drop all language-specific clauses that directly specify denotations of names or satisfaction-conditions of simple predicates.

However, Soames is quite explicit that his reconstruction of Tarski's theory of truth does not aim at historical faithfulness; rather, it is meant to preserve its laudable "deflationary" spirit, while making it immune to the modal

and related objections discussed earlier in Chapter 6. By "deflationary" spirit Soames means that Tarskian truth-definitions were designed to serve other aims than to shed light on substantial questions about semantics of real-life languages. According to Soames, Tarski conceived of his truth definitions as tools designed to serve the needs of metalogic; there was no need for him to bother about the deep question of what makes expressions and sentences of logico-mathematical languages to mean what they do. He sidestepped that question by assuming that we have a language that we understand, ML, and that ML either contains the object-language L or we are able to (correctly) translate L into ML. Now, one could complain with Field that, at this juncture, the semantic notion of (correct) translation has entered Tarski's methodology after all, which is no less obscure (perhaps it is even more obscure) than the notion of truth; moreover, it hides from us the fact that we lack any adequate explanation of what facts bring it about that L-expressions manage to possess their semantic properties, where such properties come from, or, indeed, how they are possible in this socio-physical world of ours. But Soames would retort to this that even Field's favourite B-variant does not explain to us what facts (about usage) confer on logical expressions their semantic properties, as they have been specified in the recursive clauses.[273]

There are passages in Tarski's work that suggest that he did not consider translation, synonymy or analyticity as *bona fide* semantic notions (narrowly semantic – in our preferred terminology), on the ground that they do not involve word-to-world but only word-to-word relations. But this is hardly satisfying, given that translation is to be correct, hence, at minimum, truth-preserving (extension-preserving in general). The question of what translation of a given language L into ML is correct (in virtue of what facts) is substantial to the extent the question of what L-expressions mean (in virtue of what facts) is substantial. It should be obvious that semantics in Tarski-style cannot tell us anything particularly interesting on such matters, except of how complex expressions depend for their truth-relevant properties on truth-relevant properties of less complex components. This becomes only more vivid when we set out to define truth in Tarski's manner for a sublanguage L of ML, assuming the translation of L into ML to be homophonic. Such a truth definition can hardly deceive us into thinking that it says something revealing about the meanings of L-expressions or the socio-physical origins of their meanings (the basis on which they supervene). For one thing, we need to know their meanings beforehand to understand the definition in the first place. For another, it is enough to know their meanings; but we do not need to know in addition what, if anything, they supervene on. That is to say, assuming a correct translation of L into ML, what Tarski's truth definition for L shows is how L-sentences depend for their truth on the world, given the semantic properties of their components. But it offers no story about how the connections between L and the world got established in the first place.

Soames suggests a similar defence of Tarski's definitional framework against the modal objection based on his different abstract-language conception. The objection trades on the following consideration: the semantics of a language L depends on the facts (presumably, naturalistic facts) about some folk F using

[273] Soames (1984: 420). Although Field (1994) thinks that this particular defect of his proposal could be fixed, he now thinks that it is better to abandon the reductionist project and urges a deflationary theory of truth and related semantic notions (of which more in next sections).

L; hence, had the folk F changed its linguistic practices and conventions in certain easily imaginable ways, L's semantics would have changed accordingly. According to Soames, the objection works only if we treat languages that Tarski applied his techniques to as empirical languages in the above mentioned sense: as possessing their semantic-properties contingently, that is, depending on the facts about usage of corresponding linguistic communities. However, Tarski was uninterested in empirical theory of meaning and semantics for empirical languages (he invoked repeatedly the notion of meaning or translation in Convention T, without making appearance of someone who would seriously bother to explain or reduce them to more fundamental notions). Indeed, he distinguished theoretical and applied semantics – the distinction much in the spirit of Carnap's distinction between pure and empirical semantics. Rather, what Tarski wanted to accomplish in CTFL was to show how to define in a mathematically precise manner theoretically useful (because materially adequate and inductively defined) and well-behaved (because consistent) notions of truth for a range of logical languages construed as abstract and semantically individuated entities. They are abstract in that they are represented as set-theoretical entities that are to serve certain theoretical purposes and they are semantically individuated since their representations via ordered triples of the type $\langle S_L, D_L, I_L \rangle$ involve interpretation functions construed extensionally as ordered pairs of expressions and set-theoretical entities (of the type appropriate to the semantic category of their correlated expressions). Granted, Tarski did not have this conception of language in the 1930s, but Soames' considerations make explicit his deflationary attitude to semantics in the following way. For Tarski, L is a meaningful formalism with an intended interpretation, but Tarski does not show any interest in the metasemantic question as to how (in virtue of what brute facts) did L acquired its intended interpretation: what facts are responsible for the fact that its expressions have the representational powers that they do have.

Soames wants to defend Tarski's deflationary approach to truth by showing how to sidestep the "substantial" problem lurking here in the guise of correct translation (and its factual basis, if any). Tarski's semantics does not offer us any interesting story about the foundations of linguistic intentionality, yet says that this is not its shortcoming but rather a laudable deflationary aspect. It shows that we can account for truth to a large extent independently of having answered such difficult questions. If a correct translation of L into ML can be assumed to be fixed and known, we can treat it abstractly as a mapping of L-expressions to ML-expressions, making it explicit that it is not part of our business to explain what correct translation (*a fortiori* synonymy or meaning) is based on, in much the same way as when we construe interpretation functions built-in Soamsian languages as mathematical objects. From this perspective, then, there is not much difference between the two approaches. None of them pretends to say anything substantial about linguistic intentionality and its nature.[274] Semantic individuation does not amount to anything robust at this

[274] Starting with translation of L into ML, we indirectly assign interpretations to expressions of L via understanding corresponding expressions of ML. Still, one may worry that even if this works with L-terms translated into ML-terms, by disquotation of which we grasp the interpretations of L-terms, it is not clear how to extend it to predicates of L. For Tarski-style truth-definitions do not explicitly assign values to predicates. This worry can also be answered. The definition of satisfaction implies that every n-place sentential function *f* determines a corresponding set of

level, since interpretation functions are construed as pairing functions. We are content to assume – this time, in the model-theoretic style superseding Tarski's translational style - that words are mapped to set-theoretical objects. Hence no interesting question can arise at this level as to what makes it the case that such-and-such a word is interpreted by (mapped to) such-and-such a set-theoretical object. It is up to us to define abstract objects. For abstract languages of this sort we can define truth in Tarski-style, using language-specific list-like clauses, or give Soames-style general truth definition for a variable L (dropping language-specific clauses). The two, closely related definitions are what we need in order to conduct meta-theoretical investigation of the usual sort. Compositional semantics and logical consequence are the hallmarks of this approach, both belonging to the abstract level, which can be approached in a mathematically precise manner. [275]

Foundational questions are to be attacked at a different level. Of course, as Carnap pointed out,[276] in devising abstract languages we might have in mind a particular language-in-use that we want to theoretically "approximate" or "model". Treating abstract languages in this way, we can ask what makes it the case that one such abstract model fits the actual linguistic practice of some language community (or approximates better that practice than other models). Alternatively, with Davidson, we can ask what publicly available evidence there is in support of the hypothesis that one such abstract theory of truth can do duty as an interpretive theory of meaning for a given linguistic community. Soames thinks that it is a good thing to separate in this way semantic from meta-semantic questions, and he credits Tarski with doing (more or less consciously) this.[277] His evaluation of Tarski's attitude towards semantics strikes me as largely on the right track, though, for reasons mentioned above, I do not think that Tarski adopted in the 1930s the conception of abstract model-theoretic languages that Soames propounds. Still, it seems to tally rather well with his mature model-theoretic take on semantics, and it seems to be in the spirit of recent theorizing in formal semantics, in which varying interpretation of formal language construed as mathematical objects.

7.3 Redundancy theories of truth and semantic conception of truth

So interpreted, Tarski's conception of truth definition has some affinity to deflationary conceptions of truth, whose early predecessor is the redundancy theory of truth inspired by Frege and Ramsey, and elaborated by Ayer or Strawson. In the vast literature on Tarski's work it is often said that SCT – especially the claim that T-biconditionals are partial definitions of the notion of truth - is but a sophisticated sentential version of the redundancy theory of truth, according to which there is no more to the notion of sentential truth as what is captured by particular instances of the following disquotation-schema (implicitly relativized to a language L):

---

sequences, namely those sequences that satisfy *f*. Given that all but terms paired with its free variables are garbage, we can associate with *f* the set of *n*-tuples satisfying it.

[275] Subject to the usual Tarskian conditions they are suitable for what I call metamathematics of absolute truth.

[276] Carnap (1938), (1942).

[277] This was anticipated by Carnap, who, however, would not speak of meta-semantics but rather of (very broadly conceived) pragmatics in this respect. See Carnap (1938), (1942).

(DS-schema):

'p' is true iff p

Let us see if this diagnosis is correct. The redundancy theory of truth was originally propounded for sentences such as 'It is true that snow is white', in which 'true' is a part of the sentence-forming operator 'it is true that' that behaves as the double-negation operator, turning truths into truths and falsehoods into falsehoods. The starting point is nicely captured by Frege's observation that instances of the equivalence schema:

(E-Schema):

It is true that p iff p

are platitudes that display the transparency property of the notion of truth: we can see through the prefix 'it is true that', applied to a given sentence *S*, to the content of *S*.[278] Frege and Ramsey, considered by many the grandfathers of the redundancy theory made the following observation:

> "If I assert 'It is true that sea-water is salt'. I assert the same thing as if I assert that 'sea-water is salt…the word 'true' has a sense which contributes nothing to the sense of the whole sentence within which it occurs as a predicate […]." (Frege 1978: 251).

> "[…] 'it is true that Caesar was murdered' means no more than that Caesar was murdered, and 'it is false that Caesar was murdered' means that Caesar was not murdered […]." (Ramsey 1999: 106).

Both claims suggest something stronger than E-schema, namely that the relation of content-equivalence holds between 'It is true that p' and 'p'. One who is committed to this content-equivalence is committed to E-schema, but not the other way round.[279] At any event, since the prefix 'it is true that' seems to add nothing to the content of what it is applied to, several people concluded that 'true' can be always erased (analysed away) without any loss to content, though it performs specific pragmatic functions in language not affect in the content (such as putting an emphasis on a claim, commending or endorsing it, etc.). In the classic version of the redundancy theory, this claim is wedded to another: being a content-redundant device, 'true' expresses no property, hence no robust property calling for a deep philosophical analysis:[280]

> "We conclude, then, that there is no problem of truth as it is ordinarily conceived. The traditional conception of truth as a "real quality" or a

[278] See Frege (1918/1919). But Frege was not a redundancy theorist. He held a peculiar view, on which truth is an important, indefinable notion, characterized by the transparency property. In his view, what this shows is that it is not best construed as a property corresponding to a predicate; rather, it is to be understood as that at which we aim in making judgments and assertions or forming beliefs).

[279] See Künne (2003).

[280] Ramsey (1999; 2001) is often considered the father of this approach, although he did not make the claim that 'true' does not express a property. This claim was made famous only later by Ayer (1952).

> "real relation" is due, like most philosophical mistakes, to a failure to analyse sentences correctly." (Ayer 1952: 89).

Now, Tarski criticised one version of the redundancy theory under the name 'the nihilistic approach to the theory of truth',[281] according to which 'true' makes sense only when used (syncategorematically) in the contexts 'It is (not) true that p', all other (categorematic) uses of it being treated as illegitimate (in part because they give rise to semantic paradoxes). In Tarski's opinion, the truth-nihilists are right to say that the word 'true', in its syncategorematic uses, is a sort of content-redundant device. If 'true' were used only in such contexts, one could well wonder whether it performs other than purely stylistic or ornamental functions in the discourse. Tarski's interesting and correct response to the truth-nihilists was that they propose to eliminate from the discourse precisely those predicative uses of 'true' that do not seem purely stylistic or ornamental but perform very useful expressive functions. Thus, for instance, the word 'true' is often attached to a description of a sentence that a speaker is not in a position to reproduce *verbatim* but she has reasons to accept/assert or reject/deny (the so-called 'blind ascriptions of (un)truth'):

> Though I cannot quite recall it, the first sentence in the *Tractatus* is true.

Furthemore, the word enables us to express generalizations of the following kind:

> All consequences of true sentences are true,

or

> Every alternation of a sentence and its negation is true.

The nihilistic approach to truth in particular and the redundancy theory of truth in general are hard pressed to explain such expressively useful applications of the notion of truth. Tarski's SCT also gives pride of place to the intuition that a sentence of the form "'p' is true" is equivalent to 'p' under the semantic notion of truth, but Tarski saw very clearly that many predicative uses of 'true' cannot be analysed away (eliminated) from such contexts in any automatic way.[282]

Ramsey seemed to be well aware of the fact that in the ordinary language 'true' is not easily eliminable from such contexts.[283] He considered the statement of the following type:[284]

> Russell is always right,

---

[281] Tarski (1969: 111). In (1944: 358-359) he used the label 'redundancy of semantic terms – their possible elimination'.

[282] Here Tarski was closer to modern deflationist theorists who maintain that the point of 'true' in the discourse is that is serves a certain logical or expressive function, enabling us to express such generalizations (infinite conjunctions or disjunctions), and that it is capable of performing this function because its use is governed by something like the schematic principle expressed by T-schema. See Quine (1970), Leeds (1978), Horwich (1982, 1990) or Field (1994).

[283] Ramsey (1999, 2001). But Ramsey did not make the point that modern deflationists are so proud of: namely that the point and utility of the notion of truth in language consists precisely in such uses.

[284] Ramsey (1927: 143).

or

Everything Russell says is true.

What Ramsey first proposed was the following analysis, expressed in a semi-formal English:

(For every *p*): if Russell says *p*, then *p* is true.

However, this analysis contains 'true', and for a good reason. For if we simply deleted 'true from it, we would get

(For every *p*): if Russell says *p*, then *p*.

But then we are hard pressed to explain what exactly it is that we want to express. The trouble is that we cannot understand quantification in the objectual style, because the second occurrence of '*p*' is grammatically hostile to nominals – it calls for sentences. Anticipating deflationist theories of truth a couple of decades ago, Ramsey conjectured that the function of 'true' in natural language lies exactly in the fact that it simulates the grammatical role of prosentential anaphoric reference (on analogy with pronominal anaphoric reference), since natural language does not actually contain right types of presentences:

> "The only prosentences admitted by ordinary language are 'yes' and 'no', which are regarded as by themselves expressing a complete sense, whereas 'that' and 'what' even when functioning as short for sentences always require to be supplied with a verb: this verb is often 'is true' and this peculiarity of language gives rise to artificial problems as to the nature of truth, which disappear at once when they are expressed in logical symbolism ..." (Ramsey 1990: 437)

In view of this, it would seem that truth is not a content-redundant device, since there can be hardly any definition of it that would licence its elimination form every sentential contexts in which it can meaningfully appear. Ramsey accepts this conclusion for natural languages (viz. the fact that they do not contain the right sort of presentences that would do the job of 'true') but proposes to deal with this obstacle at least for his semi-formal language. What he wants us to realize is that 'p' already contains a verb - what is supposed to come into its place is always a sentence and sentence always contains a verb. Any proposition whatever, of any logical form whatever contains a verb. Ramsey invites us to consider propositions of the (Russellian) form *aRb*. For propositions of this particular form we have:

(For every *a*, *R*, *b*): if Russell says *aRb*, then *aRb*.

which, unlike the original analysis contains an explicit verb-like element. Or so Ramsey claims. His idea was that if we could somehow gather all propositional forms of sentences that Russell could ever assert, and, for each form we could write down a similar generalization, then such generalizations would jointly capture the content of the statement "Russell is always right".

Ramsey was therefore not overly pessimistic about the prospects of giving the truth definition in the following spirit:[285]

> *x* is true iff *x* is a proposition/belief that *A* is *B* and *A* is *B*, or *x* is a proposition/belief that *aRb* and *aRb*, or *x* is a proposition/belief that ...,

which would generalize particular equivalences of the type:

> a proposition/belief that *A* is *B* is true iff *A* is *B*,

or

> a proposition/belief that *A* is *B* is true iff it is a proposition/belief that *A* is *B*, and *A* is *B*.

However, he quickly realized that the number of logical forms is going to be indefinite so that they cannot be captured in a single finite statement:

> "We cannot, in fact, assign any limit to the number of forms which may occur, and must therefore be comprehended in a definition of truth; so that if we try to make a definition to cover them all it will have to go on forever, since we must say that a belief is true, if supposing it to be a belief that *A* is *B*, *A* is *B*, or if supposing it to be a belief that *A* is not *B*, *A* is not *B*, or if supposing it to be a belief that either *A* is *B* or *C* is *D*, either *A* is *B* or *C* is *D*, and so on ad infinitum." (Ramsey 1990: 438)

For this reason he preferred to stick with the simpler definition:

> *x* is true iff $\exists p$ (*x* is a proposition/belief that *p* and *p*).

We are to realize that '*p*' already contains a verb - what comes into its place is always a sentence and sentence always contains a verb.

Note that the success of this strategy has a corollary: there can be no question that on the right side we have specified a certain property common to all true propositions/beliefs. We need not claim that it is a property that is robust in a physical or metaphysical sense, or that it can be reduced to something more fundamental. But, logically speaking, it is a property nevertheless. This is interesting, since the prominent redundancy theorists such as Ayer or Strawson maintained that, despite grammatical appearances, the truth predicate does not express a property of truth-bearers. Consider here what Ayer says:

> "… the word "truth" seems to stand for something real; and this leads the speculative philosopher to enquire what this "something" is. Naturally he fails to obtain a satisfactory answer, since For our analysis has shown that the word "truth" does not stand for anything, in the way which such a question requires." (Ayer 1936: 89)

---

[285] In (1927) he focuses on propositions, whereas in (1990) he offers some reasons to prefer beliefs.

Although Ramsey did not commit himself to this claim, it seems vital to the central redundancy claim to the effect that "That p is true" asserts no more and no less than "p".[286] If "That p is true" ascribed truth to the proposition that p, we could object that 'p' does no such thing and consequently that the two assert different things. The redundancy theorist is thus faced with the following dilemma. (1) If he does not accept non-standard quantification, he can maintain that truth is no property but the price he has to pay is that he cannot analyse away many occurrences of 'true' and he cannot even state a general truth-definition. (2) If he does accept non-standard quantification, the redundancy theorist can perhaps hope to analyse away problematic occurrences of 'true' or to state a general truth-definition, but the price he has to pay is that he can no longer seriously maintain that truth is not a property.

One may still worry whether Ramsey succeeded in making a good sense of his definitions and analyses framed in his semi-formalism. Clearly the success of his strategy depends on whether there is a working account of quantification that will make the definition intelligible, because under the objectual reading of quantifiers it does not make sense. The same goes, *mutatis mutandis*, for analyses using '*aRb*' and such like. Ramsey did not comment on the matter, but several modern authors attracted to his conception have suggested either some sort of higher-order quantification over proposition or substitutional quantification. On assumption that such interpretations of quantifiers are intelligible, there is some reason for optimism. On the other hand, one has the feeling that had Ramsey focused his attention on formalized language whose logical syntax can be defined in a neat recursive style, he would not need a devious apparatus of non-standard quantification. This is how Tarski attacked the problem, except for treating quantifiers objectually. But Ramsey was no Tarski and he could have been more concerned with natural language, for which we have reasons to doubt - as Tarski knew – if we can recursively characterize truth for it in a compositional style.

It is interesting that before giving his own truth definition, Tarski considered definitions of sentential truth closely related to redundancy-type truth definitions:

a) (For all *p*): '*p*' is true (in L) iff *p*;

b) (For all *x*): *x* is true (in L) iff $\exists p$ (*x* = '*p*' and *p*).

At first blush, (a)-definition is a straightforward generalization of SCT subsuming all T-biconditionals for L. However, Tarski rejected it on the ground that it defines 'true' only as applied to quotational names of sentences, not showing how to eliminate it from different sentential contexts (i.e. attached to a description or pronoun/variable). (b)-definition does not suffer from this problem, but Tarski complained that if "'*p*'" is taken to be the name of the letter enclosed in the quotation marks, there is nothing for the quantifier to bind, and the definition accordingly does not work (the same objection applies to (a)). One

[286] In this respect, „It is true that p" is more convenient to redundancy theorists, because 'true' does not appear there as a predicate but as a part of the prefix "It is true that ___". Indeed, the truth-nihilists called such occurrences of 'true' syncategorematic, saying that they are the only legitimate occurrence of 'true'.

may want to interpret quotational marks as a sort of function assigning expressions their quotational names. So interpreted, though, quotational marks form a non-extensional (indeed, hyper-intensional) context, and such were viewed with a great suspicion by Tarski. Finally, he worried that (a)-definition leads quickly into semantic antinomy. There is no need to go into the details of why he thought so. Suffice it to say, for the time being, that his arguments here have not been found convincing by contemporary deflationists, who argued that there is a substitutional or higher-order reading of it that is both coherent and immune paradox, provided appropriate measures are taken (in the spirit of Tarski's own restrictions).[287] Still, Tarski could have complained that the substitutional reading presupposes the very notion of truth, and the higher-order propositional reading posits propositions as values of variables, which are intensional entities of dubious clarity and ontological status (viz. his nominalism and scepticism concerning non-extensional operators and contexts). This is a serious objection, with which any decent deflationist account of truth should come to terms with.

Tarski also discussed another objection against his SCT to the effect that when we take T-biconditionals on face value what they show is that the truth-predicate can always be eliminated when attached to a quotational name of a sentence. And what is eliminable is in a way redundant or "sterile". Now, Tarskian truth definitions have the potential of eliminating truth-predicates from all contexts of the metalanguage (or metatheory) in which they are defined. However, Tarski did not regard this as a sign of the fact that SCT is a redundancy theory of truth, since, by parity of reasoning, one would have to conclude that all defined terms (in science) are useless or sterile - which he deemed absurd.

7.4 Disquotational theories of truth: Quine

Our analysis cannot be complete, unless we review modern deflationary theories, the prominent place among which has been occupied by disquotational theories of truth. Very roughly, their proponents maintain that our ordinary notion of truth is somehow captured by the disquotation-schema (or its variant framed in a different language than English), or, more precisely, by its non-paradoxical instances (as classical logical reasoning yields contradiction if self-referential sentences stating their own untruth instantiate 'p'). Sometimes it is said that our ordinary notion of truth is characterized by the fact that sentences of the form 'p' and ''p' is true' are inter-deducible. Both ways of stating disquotationalism leave much to be desired, and different authors tend to specify them in different ways. In order to see what, if anything, Tarski's theory of truth has in common with these doctrines, I shall first review the general motivation for the doctrine of disquotationalism that is commonly attributed to Quine, then briefly explaining the essentials of the best developed disquotational conception of truth due to Field.

Quine is undoubtedly the *spiritus agens* of disquotationalism, although it is a delicate question to what extent he could indeed be considered a true disquotationalist, given that his various claims sometimes pull in opposite

[287] Cf. Soames (1999), Field, 1994, or David (1994).

directions.[288] Like Tarski, he thought that if 'true' appeared only as attached to the quotational name of a sentence, then we could assert just as well the sentence itself:

> "What can justly be said is that the adjective 'true' is dispensable when attributed to sentences that are explicitly before us." (Quine 1987: 214)

At this point, the champion of propositions as primary truth-bearers might want to complain that truth depends not on language but on the world. Quine retorts to this: granted, but that is no argument in favour propositions, since, first, there are weighty arguments to the effect that they are creatures of darkness (viz. his radical translation argument that attempts to show that propositions, qua cognitive meanings preserved under translation, are simply entities without identities), and, second, the dimension of dependence of truth on the world is captured in a neat form in Tarski's biconditionals such as:

> 'Snow is white' is true iff snow is true.

The effect of quotation marks is to allow us talk about linguistic expressions, while the effect of 'true' (and other predicates belonging to what the theory of reference: '___denotes___' or '___satisfies___') undoes the effect of quotation, taking us back from the talk about expressions to the talk about the world:

> "Quotation marks make all the difference between talking about words and talking about snow. The quotation is a name of a sentence that contains a name, namely 'snow', of snow. By calling the sentence [viz. 'Snow is white'] true, we call snow white. The truth predicate is a device of disquotation." (Quine 1970: 12),

In ascribing truth to 'Snow is white' we will be understood by our fellow speakers as having taken a stand on how the world is. This is the grain of truth in correspondence conceptions of truth. Yet there is no need to go so far as to claim that truth of a sentence consists in its correspondence with a fact (state of affairs), since facts are gratuitous entities that contribute nothing 'beyond their specious support of a correspondence theory'.[289] As Ramsey observed, we may do some justice to the correspondence intuition via the following platitude:

---

[288] See especially the three Davidson's three essays on Quine's conception of truth in his (2005b). He shows that it is particularly hard to reconcile Quine's contention that the nature of truth is disquotational with his claim that semantics (meaning) is best approached in Davidson's truth-theoretic style, which formalizes Tarski-style truth definition as an axiomatic theory with primitive semantic notions (of denotation and/or satisfaction), that aims to specify in a recursive manner truth-conditions for an infinite number of sentences of a given language (see various papers in Davidson 1984 as well as his 1990). Since Dummett's classic paper (1959) the prevailing view has been that disquotationalism is incompatible with truth-conditional theory of meaning (in Davidson's style). But recently there have been attempts to argue that this is not so; Williams (1999) says that the use of the notion of truth in Davidson's theory of meaning is compatible with the disquotationalist's understanding of it as a device of generalization. Davidson (2005a) was strongly opposed to such an interpretation, pointing out that disquotationalism misses the translinguistic character of our notion of truth, whereas his theory of meaning takes a full advantage of this aspect of truth.

[289] Quine (1992: 80).

'Snow is white' is true iff it is a fact that snow is true.

However, this platitude does nothing to support a robust correspondence theory that postulates relations holding between sentences and facts, or a structural isomorphism between structured sentences and structured facts. It shows no more than saying that whenever we assert the left side we could just as well assert the right side, and *vice versa*.

Does our willingness to accept instances of the disquotation-schema means that 'true' is just the device of disquotation, its sole linguistic function being to undo the effect of quotation marks, so that in attributing truth to the sentence we always speak about the world, albeit indirectly? Not really. Quine is explicit that predicative uses of 'true', where it is attached to displayed sentences, are dispensable. Where, on the other hand, the indirect talk of the world via truth is indispensable are contexts in which the truth-predicate is not attached to a given sentence "explicitly before us", but in which it is attached to a description of a sentence not displayed or in generalizations of the sort Tarski and Ramsey mentioned.

We have noticed that in the first type of contexts the truth-predicate is needed to affirm (deny) a sentence that we cannot explicitly formulate (or we are lazy to formulate it), as when somebody asserts

Fermat's last theorem is true,

on the evidence of a reliable source but without being able to reconstruct what the conjecture actually states. In the second type of contexts, one needs the truth-predicate in order to affirm what Quine loosely calls 'some infinite lot of sentences', [290] as when one asserts

What the Pope asserted is true,

or

Every sentence of the form *p or not p* is true.

Quine's massage is that we need the notion of truth with disquotational character in order to express such generalizations, since they allow us to generalize on an infinite lot of sentences. Take, for instance, the second generalization. Confining ourselves to English, what sentences it generalizes on? Arguably on English sentences such like

Snow is white or snow is not white;
Tom is mortal or Tom is not mortal;
…

It would seem that the most straightforward way of generalizing is this:

$\forall p$ (p or not p),

[290] Quine (1970: 12). See Halbach (1999) for a thorough discussion of what the talk about 'infinite lot of sentences' amounts to.

But Quine rejects this proposal quickly on the ground that such quantification is hardly intelligible, dismissing the possibility of interpreting it either as a sort of higher-order quantification over propositions or as a substitutional quantification (with English sentences forming the substitution class). Though we shall see that not every deflationist would agree with him, he had reasons to reject this proposal. Given his two famous theses

(1) *No entity without identity*,
(2) *To be is to be a value of a bound variable in a canonical notation*,

in tandem with his radical-translation argument inveighing against propositions (putative cognitive meanings of declarative sentences) as entities without identities, there is no wonder that Quine does not deem it feasible to quantify over propositions. And he complained about substitutional quantification on the ground that we cannot in general presuppose that there is in the language an expression for every entity, in which case substitutional interpretation yields intuitively wrong predictions. Moreover, if we read such quantification in the standard manner, the notion of truth gets reintroduced, since quantified sentences are usually explained in terms of truth of all/some instances of their matrixes.

So, according to Quine, the generalization on sentences of the form *p or not p* by means of the truth-predicate helps us to express what Ramsey-style generalizations try in vain, absent an intelligible and non-circular interpretation of quantification into sentence-positions. And although he did not explicitly say so, what he arguably had in mind is the following reasoning, taking full advantage of the disquotational character of truth. We first note that each of an infinite lot of English sentences of the following type

A) Snow is white or snow is not white
Tom is mortal or Tom is not mortal;
….

is equivalent to a corresponding sentence, in which ‘true’ is attached to its quotational name:

B) ‘Snow is white or snow is not white’ is true;
‘Tom is mortal or Tom is not mortal’ is true;
….

Recall: in quoting *S* we are up to say something about *S*, but by appending ‘true’ to *S*’s name we *disquote S*, thereby saying something about the world. Granted, what we thereby say about the world could be said more directly by uttering *S*. Quine agrees with Tarski that if this was the only use of it, the truth-predicate would indeed be a redundant device having at best ornamental and pragmatically based functions in our discourse. But once we appreciate the disquotational character of ‘true’ implying the equivalence of A-type sentences with their corresponding B-type sentences, we realize that since we can quantify in the objectual style into the positions occupied by quotational names, we can frame a generalization of the following type, drawing on the fact that A-type sentence

sentences share a certain salient property (viz. all of them having the form *p or not p*):

> Every sentence of the form *p or not-p* is true,[291]

or, equivalently:

> For every *x* (if *x* is of the form *p or not-p*, then *x* is true).[292]

With such generalizations we have attained Quine's *semantic ascent*: by having said something general about linguistic items we have indirectly expressed something general about the world. In a similar vein, we could explain why the truth-predicate is needed to express 'What the Pope asserted is true", and the like. Disquotationalists are very fond of this observation of Quine, anticipated by Tarski, and, in its propositional version, by Ramsey, when he claimed that we need the truth-predicate to imitate the effect of prosentences absent from natural language. The point is neatly expressed by Leeds:

> "It is not surprising that we should have use for a predicate P with the property that '"____" is P' and '____' are always interdeducible. For we frequently find ourselves in a position to assert each sentence in a certain infinite set *z* (for example when all the members of *z* share a common form); lacking the means to formulate infinite conjunctions, we find it convenient to have a single sentence which is warranted precisely when each member of *z* is warranted. A predicate P with the property described allows us to construct such a sentence: $(x)(x \in z \rightarrow P(x))$. Truth is thus a notion that we might reasonably want to have on hand, for expressing semantic ascent and descent, infinite conjunction and disjunction. And given that we want such a notion, it is not difficult to explain how it is that we have been able to invent one: the Tarski sentences, which axiomatize the notion of truth, are by no means a complicated or recondite axiomatization; the possibility of moving from this axiomatization to the explicit truth definition was always latent in the logical structure of our language, though it took a Tarski to discover it." (Leeds 1978: 43).

For the time being, I put aside what Leeds says about Tarski's theory. It seems to me that his account renders accurately Quine's position as just described. But one thing that is new is the claim that generalization using the disquotational truth-predicate allow us to simulate infinite conjunctions or disjunctions, which, for obvious reasons, we are not in a position to assert. Thus, the foregoing generalization about all sentences of the form *p or not p* allows us to express a would-be infinite conjunction of A-type sentences

> (Snow is white or snow is not white), and (Tom is mortal or Tom is not mortal), and ...,

---

[291] Or as Quine also put it: *Every alternation of a sentence and its negation is true*.

[292] However, David (2008) argues that this seemingly innocent procedure hides many complications.

which we cannot in fact formulate. Fortunately, the sentence ‘What the Pope asserted is true’ allows us to express a would-be infinite disjunction of the sort:

> (If what the Pope asserted was “Snow is white”, then snow is white); or (if what the Pope asserted was “Grass is green”, then grass is green); or ...

Quine’s own remarks strongly suggest that the disquotational notion of truth is indispensable, so long as we want or need to express such generalizations.

> “ [...] ‘true’ is dispensable when attributed to sentences that are explicitly before us. Where it is not thus dispensable is in saying that all or some sentences of such and such specified form are or are not true, or that someone’s statement unavailable for quotation was or was not true…” (Quine 1987: 214)

The evidence for its indispensability might be that ‘true’ cannot in general be analysed away from such contexts. Indeed, if we could eliminate ‘true’ form every context without any loss whatever, we would have a good reason for the claim that it can be dispensed with. The upshot is that if we think that truth is an indispensable expressive device, we cannot hope to define it. It is nothing against this that Tarski showed how to get rid of truth through his explicit definitions. For what he showed was that if ‘true’ is restricted to a formalized language of the right type, we can define it within a stronger metatheory in terms of non-semantic notions, hence eliminate it from every context of the metatheory. Quine shares Tarski’s preference for regimented languages with 1st-order structure, but he does not seem to confine what he says about truth to such languages only. Rather, he seems to have in mind our ordinary notion of truth as applied to a natural language. But given that no general truth definition for a natural language seems possible that does justice to the disquotational character of truth and is consistent, we have no general method allowing us to eliminate ‘true’ from every context.

It is clear that we cannot formulate infinite conjunctions and disjunctions in our language. Hence the *prima facie* need for the truth-predicate. Alternatively, it is sometimes said that generalizations in terms of ‘true’ allow us to express what could be just as well expressed by using substitutional quantifiers, in the following way (I distinguish the substitutional quantifiers by using a different notation common in this context):

> Πp (p or not p)

or

> Σp (if the Pope asserted p, then p).

With substitutional quantifiers, infinite conjunction or disjunctions could be expressed. Does this mean that generalizations using the truth-predicate can be dispensed with, after all? This would follow only if substitutional quantification

could be made sense of independently of the notion of truth, and this is by no means clear, since the standard explanation of them proceeds in terms of truth of all/some substitution-instances of matrixes that such quantifiers operate on (no wonder that, so understood, substitutional quantifiers allows us to express such generalizations). If, on the other hand, we could make sense of them independently of the notion of truth, Quine's claim that the later notion is indispensable in our discourse should be rejected, since we would then have an elegant alternative how to express desired generalizations (or infinite conjunctions and disjunctions). At most, what one is justified to claim is that 'true' or substitutional quantifier is indispensable. Indeed, there are disquotationalists, most prominently Field, who have proposed to read the substitutional quantifiers - '$\prod p$', '$\sum p$' - as abbreviating infinite conjunctions and disjunctions respectively.[293] Field points out that with such quantifiers at hand, it should be easy to define 'true' (for a given L):

(For every *x*): *x* is true (in L) iff $\prod p$ (if *x* = '*p*', then *p*)

or, alternatively

(For every *x*): *x* is true (in L) iff $\sum p$ (*x* = '*p*' and *p*).

According to him, the disquotational truth and substitutional quantifier are interdefinable devices stimulating infinite conjunctions or disjunctions. Supposing L to contain the sentences 's_1', 's_2',…, 's_n', we have:

x is true (in L) iff $\prod p$ (if *x* = '*p*', then *p*) iff (if *x* = 's_1', then s_1), and (if *x* = 's_2', then s_2),….., and (if *x* = 's_n', then s_n).

or

x is true (in L) iff $\sum p$ (*x* = '*p*' and *p*) iff (*x* = 's_1' and s_1), or (*x* = 's_2' and s_2),…, or (*x* = 's_n' and s_n).[294]

Strictly speaking, this can hardly be deemed a definition of truth, so long as it is doubtful whether we have in English substitutional quantifiers of this type. At most, the two formulations suggest to us what expressive role the word 'true' is supposed to play in language according to the disquotationalist – simulating infinite conjunctions or disjunctions. Still, it is questionable whether can we take seriously explanations of quantifiers in terms of something that we cannot ever hope to entertain (infinite conjunctions or disjunctions);[295] moreover, several authors have complained that not every generalization in terms of truth can be imitated in terms of substitutional quantifiers,[296] or even that there are no clear cases of such quantifiers to be found in natural language (and how can we explain something via something not yet at our disposal?).[297] If such complaints are on the right track, Quine's indispensability thesis can be

[293] Field (1994). For more details on this proposal see David (1994).
[294] Field (1994: n. 17). The two formulations are equivalent given that there is exactly one quotational name for every sentence.
[295] See Ebbs (2009: 56-57).
[296] Soames (1999).
[297] Horwich (2010).

vindicated.

7.4.1 Disquotationalism after Quine

Quine paved the way for modern disquotationalists, who share the idea that the content of the notion of truth for a given language or idiolect L is fixed and exhausted by all non-paradoxical instances of (DS), while its point and utility is explained by its indispensable role - provided we have no other ways of simulating infinite conjunction and disjunctions are - in expressing generalizations and blind ascriptions.[298] It is then a big question how to formulate the disquotationalism in precise terms, even ignoring the problem of paradox, which is fairly pressing for the disquotationalists who aspire to capture the ordinary notion of truth, which applies also to sentences in natural language that contain that very notion. Some authors concede that all non-paradoxical instances of (DS) for L, taken together, dot not strictly speaking define the notion of truth for L, since they do not tell us how to eliminate 'true', when it is not attached to a sentence displayed within quotational marks. Still, they are ready to agree with Quine, when he says:

> "...yet [instances of DS for L] serve to endow 'true-in-L' ... with every bit as much clarity, in any particular application, as is enjoyed by the particular expressions of L to which we apply them. Attribution of truth in particular to 'Snow is white', for example, is every bit as clear to us as attribution of whiteness to snow." (Quine 1953b: 138)

> "[...] in a looser sense the disquotational account does define truth. It tells us what it is for any sentence to be true, and it tells us this in just as clear to us as the sentence in question itself [...] Evidently one who puzzles over the adjective 'true' should puzzle rather over the sentence to which he ascribes it. Truth is transparent." (Quine 1992: 82)

The idea is that instances of (DS) exhaust the content of 'true' for L, the totality of them yielding an implicit definition in axiomatic style. Quine pointed out that if we have a disquotational theory T for L in terms of 'true' and a theory T* obtained from T by replacing everywhere in it 'true' with 'true*' (so that T and T* disquotationally agree on every L-sentence), we have reason to treat the two predicates as equivalent, since

$$T \cup T^* \vdash s \text{ is true (in L) iff } s \text{ is true* (in L)},$$

holds for each L-sentence *s*. This shows that T fixes the application of 'true' with respect to every sentence of L.[299]

---

[298] Ebbs (2009) argues that disquotational truth is indispensable, since such generalizations cannot be expressed without it.

[299] Granted, this is not enough to fix uniquely the extension of 'true' w.r.t. L in the model-theoretic manner, since it does not fix the extension of 'true' in non-standard parts of non-standard models of L. Ketland (1999) used this to argue against various deflationary theories of

Having abandoned the naturalistic program of scientific semantics outlined in his Tarski-article (1972), which was reviewed and criticised in Chapter 6, Field has come to champion an influential version of disquotationalism,[300] inspired by Quine's and Leeds's observation to the effect that truth has a useful expressive role to play in language, a role that even those have to acknowledge who believe that it is a substantial notion.[301] According to Field, disquotational truth is a logical device of a sort serving us to express infinite conjunctions and disjunctions. However, in order to disable various objections – starting with the modal objection and ending with Gupta's *objection from conceptual overloading* to be reviewed in the next section – he proposes that the characteristic feature of the disquotational notion of truth is this:

(i) X can understand "'S' is true" only to the extent that he can understand 'S'.

(ii) For X, the sentence (utterance) "'S' is true" is cognitively equivalent to 'S' as X understands it.

Field therefore speaks of the conception of *pure disquotational truth* to be distinguished from the conception of *extended disquotational truth* for 'true' as applied to sentences not in X's idiolect (language comprised by sentences that X understands). Field takes the pure disquotational truth to be the basic notion, following here Quine, who argued that disquotational truth is *immanent*, because applicable only to sentences of the speaker's home language, whereas the *transcendent* notion of truth applies also to foreign sentences, and depends on the notion of interlinguistic synonymy (or translation). Clearly, in order to explain in my home language the notion of truth as applied to a foreign sentence that I do not understand, it won't do to disquote that foreign sentence. I could not understand such an explanation, because I would not understand the disquoted sentence in the first place. And, as Quine put it in the quoted passage, attribution of disquotational truth is no more but also no less clear than the sentence to which truth is attributed. That's the reason why the pure disquotational notion is the basic disquotational notion of truth. On the basis of this notion and the notion of inter-linguistic synonymy (correct translation), the extended disquotational

---

truth for L(PA) framed in ML, which is just L(PA) augmented with 'true' applicable to sentences of L(PA). Indeed, T-schema fixes the extension of 'true' only with respect to standard Gödel numbers (for standard sentences), so that its extension can be fixed in arbitrary way with respect to non-standard numbers so that the general principles fail to hold for non-standard sentences. But Bays (2009) shows that Ketland's argument is problematic, as no disquotationalist has ever wanted to claim that a disquotational theory for L(PA) fixes the extension of truth in all models, including non-standard models. Rather, what disquotational theories based on DS (w.r.t. L) aim to fix is the extension of 'true' on the intended interpretation of L (or, in the standard model, if you prefer). It is not their business to fix its extension in non-standard models, or, as Bays aptly put it: "to determine the application of T, not just to every sentence in L, but to every object that any model of PA *thinks* is a sentence in L " (Ibid: 1068).

[300] See especially Field (1994); already in his (1986) Field expressed some sympathies to deflationism.

[301] Although many thinkers who consider truth to be a substantial notion think that it admits of an informative analysis (in terms of correspondence, coherence, utility, warranted assertibility, or what not), the primitivist about truth holds that truth is a more substantial concept that the deflationist claims, without committing oneself to any informative definition or analysis of truth. Cf. Davidson (1990) or Frege (1918/19).

notion of truth can be explained as follows:

> *S* is true iff *S* is synonymous with a sentence *S** that is true in the purely disquotational sense (= there is a sentence *S** that is true in the purely disquotational sense such that *S** correctly translates *S*).

Since interlinguistic synonymy (translation) is a philosophically problematic idea – if Quine's indeterminacy arguments have some bite - Field proposes a less involved notion of *truth relative to a correlation*, requiring only so much that a foreign sentence be correlated with a sentence in the speaker's idiolect that is true in the purely disquotational sense.[302] Finally, Field thinks that it is possible to

> "...use the concept of pure disquotational truth as originally defined in connection with the foreign utterance, without relativization." (Field 1994: 79)

The idea is that a speaker of English understanding 'Der Schnee ist weiss' should also understand the sentence

> 'Der Schnee ist weiss' is true if and only if der Schnee ist weiss,

and accept it on the basis of his understanding of it.

Regarding the pure disquotational notion of truth, Field claims that instances of (DS) for X's idiolect L capture that notion for L-sentences as X understands them. So construed, the pure disquotational theory of truth cannot be finitely stated, in case L has more than finitely many sentences ((DS) is not a definition but a schema that does not state anything, though its particular instances *do* state something). Partly for this reason, Field seems to prefer a finite generalization of (DS) in terms of the universal substitutional quantifier:[303]

> (GDS) $\prod p$ ('*p*' is true iff *p*),

interpreted so as not to give rise to semantic paradoxes, and as abbreviating an infinite conjunction if its instances. Since (GDS) entails all instances of the (DS), it seems to capture all that is essential to the pure disquotational notion of truth, and we can thus say that it axiomatizes this notion. Or so Field claims.

7.6 Problems for disquotationalism

Field's last proposal inherits potential objections against substitutional quantification. In view of this, one may rather follow Leeds in saying that the totality of instances of (DS) for L axiomatizes the pure disquotational notion of truth for L. Be that as it may, disquotationalism remains a controversial doctrine. First, our familiar modal objection can be levelled against it: disquotational biconditionals are intuitively contingent but the deflationists have to treat them as necessary, because definitional, axiomatic or analytic of truth. The objector might point out that owing to the strong (cognitive) equivalence between 'S' and "'S' is true" the disquotationalist conception has the unacceptable consequence

[302] Field (1994: 78).
[303] Ibid: 69.

that the following two sentences are equivalent

> (1) Had 'Snow is white' been used to mean that snow is black, the sentence 'Snow is white' would not have been true,
>
> (2) Had 'Snow is white' been used to mean that snow is black, snow would not have been white.

While (1) seems to state something true, (2) seems to be plain false, because whiteness of snow does not depend on linguistic matters (recall the arguments of Pap, Etchemendy and Putnam).

However, according to the disquotationalist of Field's calibre, (1) is false, if (2) is false. But Field is happy to embrace this consequence, because it is this feature that makes the disquotational notion of truth an interesting and useful expressive device that it is, its application being use-independent in roughly Quine's sense: to attribute truth, say, to 'Snow is white', is for me just to attribute whiteness to snow, irrespectively of how people could have used 'snow' or 'white' in counterfactual situations.[304] Truth, as a device of semantic ascent, is characterized by this feature; by using the disquotational notion of truth with this property we are able to affirm or reject an infinite conjunction (as when we want to affirm or deny that all axioms of a given theory are true). Thus, for instance, suppose we wanted to say that axioms of Euclidian geometry are contingent, that they might have been false:

> "Surely what we wanted to say wasn't simply that speakers might have used their words in such a way that the axioms weren't true, it is that space itself might have differed so as to make the axioms as we understand them not true. A use-independent notion of truth is precisely what we require." (Field 1994: 71)

Granted, it sounds odd to us that (1) should be false. But the disquotationalist has resources to explain the data. He could say – following van McGee' proposal - that 'true', as it occurs in (1) is ambiguous. In order to disambiguate it, we have to realize that 'true' needs to be relativized to some language or other, for reasons that Tarski already spelled out:

> "It makes no sense to ask, simply and in isolation, whether a sentence is true or to what a term refers, because the same sentence and the same term can occur in many different languages. Before answering the question "Is the sentence true?" one needs to ask "True in what language?" When one is able intelligibly to ask simply "Is the sentence true?" one is able to do so because the context has tacitly established some particular language as the relevant one. By default, if there is no other language in view, we ordinarily take "true" and "refers" to denote truth and reference in the speaker's own language." (McGee 1993: 118)

[304] Field (1986), 1994).

Following McGee's suggestion, one might propose that (1) can mean either something false, under the following reading:

> (1*) Had 'Snow is white' been used to mean that snow is black, then the sentence 'Snow is white' would not have been true as I now use (understand) it (i.e. in my current idiolect),

or something true, under the following reading:

> (1**) Had 'Snow is white' been used to mean that snow is black, then the sentence 'Snow is white' would not have been true as I would then use (understand) it (i.e. in my counterfactual idiolect).

We see that under reading (1*), (1) is equivalent to (2), as disquotationalism predicts; but under reading (1**), (1) is not equivalent to (2). So the modal objection can be rebutted. And, as Field argued, the deflationist has good independent reasons to deny that it touches the pure disquotational notion of truth.

There are more serious objections. Since the content of the notion of truth is assumed to be exhausted by instances of (DS), it would seem that one cannot fully grasp it unless one understands all such instances, and one cannot understand all of them, unless one understands each of infinitely many L-sentences. But it is hardly acceptable to claim that one does not understand the notion for truth for L when one does not understand a single L-word, hence L-sentences containing it. This objection was levelled by Gupta,[305] who argued that if disquotationalism (so conceived) is correct, one's understanding of the notion of truth (for L) would require "massive conceptual resources" on one's part. As it seems that one can (because one does) understand that notion without such massive conceptual resources, he concluded that disquotationalism (of this sort) cannot be correct.

This objection makes for a problem when public languages are concerned, but Field's (or McGee's) idiolectic disquotationalism is not touched by it, as it deals with the notion of truth that applies only to sentences that a speaker understands. And the axiomatization using (GTS) does not suffer from this problem either, being pleasingly finite. One may further object that since the minimal disquotational theory of truth is axiomatized by all instances of (DS) for X's idiolect L, it cannot be finitely axiomatized, if L has an infinite number of sentences. That is something the disquotationalist can live with, as he can well say that his theory can be finitely stated, though not axiomatized.

The second objection levelled by Gupta against disquotationalism is more worrying. Note first that (GDS) is not only pleasingly finite but, as Field points out, allows us to derive all instances of (DS), along with truth involving generalizations such as ('*A*' ranges over L-sentences):

(i) (For every *A*): [$\neg A$ is true iff *A* is not true]
(ii) (For all *A* and *B*): [$A \wedge B$ is true iff *A* is true and *B* is true]
(iii) (For all *A* and *B*): [$A \vee B$ is true iff *A* is true or *B* is true]

[305] Gupta (1993).

(iv) (For every A): [$A \vee \neg A$ is true].[306]

The trouble is that such generalizations are not deductive consequences of the minimal disquotational theory consisting of the collection of all instances of (DS) for L. The reason is simple: a generalization entails each particular instance, but it is not the case that the totality of instances entails the generalization – not if the underlying logic is 1st order (or compact, in general).

If the disquotational theory of truth for L is axiomatized by all instances of (DS) for L, then generalizations involving truth such as (i),...,(iii) are not consequences of it (in classical logic).[307] Gupta's generalization problem is particularly embarrassing for the conception of truth based on the claim that the point of the disquotational truth-predicate – whose content is allegedly exhausted by the totality of instances of (DS) - is that it allows expressing such generalizations, often claimed by the disquotationalists to be nothing but convenient abbreviations of infinite conjunctions. But how could they be mere abbreviations of such infinite conjunctions, if these do not even entail them? The disquotationalist who wants to face up this objection has several choices. They may grant that the minimal disquotational theory for L does not entail truth-involving generalizations, while maintaining that they are still expressible in it. Why want more? On the other hand, those who want more may want to revise their theory of truth so as to entail such generalizations via supplementing inference rules. Thus, Horwich proposes a version of the ω-rule that would allow us to derive truth-involving generalizations from the collection of their instances (however, his theory is a non-disquotationalist theory of truth focused on propositional truth).[308] More to the point, Field suggests to enrich the language by sentential variables and schemata using them - such as (DS) - and adopt two rules:[309]

- a rule allowing replacement of all instances of a schematic letter by a sentence;

- a rule allowing inference of $(\forall x)(\text{Sentence}(x) \rightarrow A(x))$ from the schema $A(\text{“}p\text{”})$, in which all occurrences of the schematic letter p are surrounded by quotes.

In effect, though, the second rule amounts to a version of the ω-rule. So, to overcome the generalization problem, both leading deflationists suggest adopting non-effective inference rules.[310]

This strategy is problematic not only because of the non-effective character of those rules but because it seems to be an *ad hoc* response to Gupta's

---

[306] Assuming the additional axiom: For every S (if S is a sentence of L, then $\sum p$ (x = 'p')).

[307] Gupta says that, following Quine, deflationists (disquotationalist or minimalists) have not clearly distinguished between *affirming a generalization* and *affirming a lot of sentences*, each of a given (finite or infinite) collection (or conjunction thereof). Of course, the collection of all instances plus the claim that they are all its instances entails the generalization, but the latter claim is itself a generalization.

[308] Horwich (1998: 22, 137).

[309] Field (1994: 63).

[310] Shapiro (2002) notes that this proposal amounts to a free-variable 2nd order logic with non-effective consequence-relation. Much the same can be said of Field's approach via substitutional quantification considered in previous sections.

generalization problem. In view of this, one may prefer to add to the set of axioms governing truth. A natural option here might be to add a couple of compositional axioms *à la* Tarski. Still another alternative would be to adopt the generalized schema (GDS), or something of the sort. But none of these alternatives is without problems from the deflationary point of view. The second alternative is problematic because the substitutional approach *vis-à-vis* truth is problematic. The first alternative runs into trouble if the deflationist wants to capture the notion of truth for a natural language, since it is by no means clear whether we can provide a compositional style truth theory for a substantive part of natural language (indeed, both Field and Horwich find this to be a serious shortcoming of Tarski's approach).

Combined with other difficulties that surround any attempt to explain the application of the notion of truth to sentences that the speaker does not have in his conceptual repertoire, these problems might detract significantly from the initial appeal of disquotational theories of truth.[311] First, any two speakers X and Y who differ in the set of sentences they understand have different (pure) disquotational notions of truth. Indeed, X has different disquotational notions of truth at different stages of his/her life. Second, it seems that we can meaningfully employ the notion of truth with respect to particular sentence that we do not understand, say, with respect to a sentence that somebody asserts on whose epistemic authority and sincerity we rely, even though what he/she asserted was in a language that we do not understand. And it seems that we meaningfully make blind ascriptions and generalizations with respect to sentences not in our conceptual repertoire. How could this be if our only notion of truth is the idiolectic?

The worry is not so much that the pure disquotational notion of truth is not intelligible but that it is by no means clear what it has to do with our ordinary notion of truth that has such uses.[312]

7.7 Comparing Tarskian and disquotationalist theories of truth

Tarski preferred to have truth defined for L(T) in essentially stronger MT (e.g. T expanded by higher order variables) through the recursive definition of satisfaction, in a way that reveals the contribution of logico-syntactic structure to truth conditions of sentences and allows derivation of truth-involving generalizations. With such generalizations at hand, he was in a position to decide Gödel-type sentences (such as Con_T) belonging to L(T), undecidable in T. The upshot is that Tarski's preferred theory of truth $MT \cup D_{TR}$ – where D_{TR} encapsulates the explicit definition of '*Tr*' - is not conservative over T, because it proves L(T)-sentences not provable in T. Now, $MT \cup D_{TR}$ yields a considerably more powerful theory than T∪TRUE, where TRUE is the complete set of T-biconditionals for L(T). In spite of the fact that T∪TRUE is materially adequate

[311] Cf. David (1994), Gupta (1993), Künne (2003), Shapiro (1998).

[312] For more objections see David (1994) and Künne (2003). The disquotationalists have come up with various responses and proposal. Perhaps Field's extended disquotational notion of truth could help them to explain some *prima facie* recalcitrant data.[312] Such attempts, though, also face serious problems. Moreover, Field recognizes himself that the extended disquotational truth relies on the notion of interlinguistic sameness of meaning, which may be hard to explain without having recourse to truth.

and conservative over T (so consistent, provided T is), Tarski did not take it as a satisfactory theory of truth for L(T), on the ground that it is not possible to deduce from it generalizations involving truth – viz. the principles (I) - (IX).[313] It thus appears that although the material adequacy criterion spelled out in Convention T is the heart of Tarski's semantic conception of truth, it is by no means all that he expected from a satisfactory theory of truth, or T∪TRUE would have to be completely satisfactory by his lights. Consequently, in so far as the minimal disquotational theory of truth for L(T) is T∪TRUE, formulated in L(T)∪{'*Tr*'}, it is clear that Tarski did not consider it satisfactory.

On the other hand, Tarski-style theory of truth based on the explicit definition of '*Tr*' for L(T) is not so attractive option for the disquotationalists, because of its substantial ontological commitments and its demand that '*Tr*' belong to an essentially stronger language. Since, by Tarski's standards, no language is essentially stronger than natural language, his truth-definition cannot be given for natural language. Still, one may hope to specify its properties. The disquotationalists do not want to limit their inquiries to logico-mathematical languages; rather, it is their contention that the notion of sentential truth applicable to a natural language is best thought of as disquotational. Even in those cases where the direct definition of truth is possible, they may prefer treating truth as a primitive notion to be axiomatized, instead of defining it by means of higher-order (set-theoretically stronger) machinery with its substantial ontological commitments, on the ground that truth is conceptually more fundamental than those higher-order (set-theoretical) means needed to define it directly. Also, it is utterly implausible to maintain that a higher-order (or set-theoretic) formula needed to define truth could conceivably *fix the meaning* of the truth-predicate for L(T) (not just its extension), whereas the disquotationalist might hope to capture its meaning - *implicitly define* it - via carefully selecting axioms laying down the basic properties of the truth-predicate. If the disquotationalist wants to characterize adequately this notion of truth, he has to think of alternative axiomatizations, which do not require essentially stronger resources.[314]

Although the disquotationalist do not confine their investigations to languages of logic or mathematics, some hotly discussed topics concern precisely such languages - typically, L(PA). I cannot enter the debate in detail that it deserves, doing justice to all technical aspects. I shall confine myself to a few clarificatory comments in order to show its connections to Tarski's conception of truth. The question is what axiomatic theories of truth over the base theory PA are available to the disquotationalist and which is the most attractive.

---

[313] Tarski considered the more general case of TRUE being added to MT not essentially stronger than T.

[314] However, by investigating possible axiomatization of truth for logico-mathematical languages one can learn a lot about what is needed in order to explicitly define truth for such languages, since it turns out that certain interesting axiomatizations are proof-theoretically equivalent to certain higher-order theories. See Halbach (2009) for a useful survey of axiomatic theories of truth. As he points out, one can interpret proofs of such equivalences so that they amount to interesting ontological reductions, because substantial ontological commitments of higher order theories are absent from axiomatic theories of truth that lay down only the properties of the truth-predicate.

Indeed, what criteria are the disquotationalists to use to settle this question? Should they expect from a reasonable axiomatic theory of truth that it proves a certain set of truth-involving generalizations, and/or should it be conservative over PA? It has become common to consider the following basic axiomatic theories of truth over the base theory PA, all of which are formulated in L_T = L(PA)∪{'*Tr*'} and are typed (as opposed to type-free), '*Tr*' being restricted to sentences of L(PA) none of which contains '*Tr*'.

- **T(PA)**: the minimal disquotational theory consisting of PA plus every T-biconditional

  $Tr(\langle\underline{\varphi}\rangle) \leftrightarrow \varphi$, for $\varphi \in$ L(PA),

  plus the induction schema

  $[\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(Sx))] \rightarrow \forall x\phi(x)$

  restricted to L(PA) – not allowing ϕ to contain '*Tr*'.

- **T(PA)*:** PA plus all T-biconditionals for L(PA) plus the induction schema allowing ϕ to contain '*Tr*'.

- **T_T(PA)**: PA plus the induction schema restricted to L(PA) - not allowing ϕ to contain '*Tr*' - plus the following Tarski-style axioms for L(PA):

  (i) For all atomic sentences $A \in$ L(PA):

  $Tr(A) \leftrightarrow A$

  (ii) For all sentences $A \in$ L(PA):

  $Tr(\neg A) \leftrightarrow \neg Tr(A)$;

  (iii) For all sentences A and $B \in$ L(PA):

  $Tr(A \wedge B) \leftrightarrow Tr(A) \wedge Tr(B)$;

  (iv) For all sentences A and $B \in$ L(PA):

  $Tr(A \vee B) \leftrightarrow Tr(A) \vee Tr(B)$;

  (v) For all formulas $A(v) \in$ L(PA) with exactly one free variable *v:*

  $Tr(\forall v\, A(v)) \leftrightarrow Tr(A(\underline{n}))$, for every *n*.[315]

---

[315] To be precise, we should formalize the clauses (i),…(v) within T_T(PA), where 'Ats', 'Sent', 'Var', 'Form', 'Neg', 'Con', 'Dis', 'Sub', UQuant' are number-theoretic analogues of (that is, represent in T_T(PA)) corresponding syntactic properties, relations and operations (viz. being an atomic sentence of L(PA), being a sentence of L(PA), etc.):

(i) $\forall x\,(\text{Ats}(x) \rightarrow (Tr(x) \leftrightarrow x))$;
(ii) $\forall x \forall y\,(\text{Sent}(x) \wedge \text{Sent}(y) \wedge \text{Neg}(y, x) \rightarrow (Tr(y) \leftrightarrow \neg Tr(x)))$;
(iii) $\forall x \forall y \forall z\,(\text{Sent}(x) \wedge \text{Sent}(y) \wedge \text{Con}(z, y, x) \rightarrow (Tr(z) \leftrightarrow Tr(x) \wedge Tr(y)))$;
etc.

- **T_T(PA)***: PA plus Tarski-style axioms (i),..., (v) plus the induction schema allowing ϕ to contain '*Tr*'.

Except of being materially adequate with respect to L(PA), the truth-theories T(PA), T(PA)* and T_T(PA) are all conservative over PA - none proves any truth-free sentence of L(PA) not provable already in PA.[316] But whereas the first two theories are deductively weak as regards truth-involving generalizations, the third is considerably better in this respect.[317]

Though T_T(PA) proves the number-theoretic counterpart of the claim that all PA-axioms are true and all inference rules of PA are truth-preserving, it does not prove soundness of T (the variable '*x*' ranging over (Gödel codes of) sentences of L(PA)):

$$True_{PA}\colon \forall x\,(Pr_{PA}(x) \rightarrow Tr(x))).$$

To prove $True_{PA}$, T_T(PA) would need the induction axioms with φ involving '*Tr*', but it does not have such resources. T_T(PA)*, on the other hand, can prove even $True_T$, as its induction axioms contain '*Tr*'. For this reason, however, it is not a conservative extension of PA, since it proves consistency of PA

$$Con_{PA}\colon \neg Pr_{PA}(\underline{\langle 0 = 1\rangle}),$$

a purely number-theoretic (truth-free) sentence of L(PA) not provable in PA.[318] What follows is an adaptation of the proof introduced in Chapter 4 for Tarski-style truth theory for L(T) – T embedding PA – based on the explicit definition of '*Tr*' in the higher-order MT. We start assuming that T_T(PA)* is materially adequate with respect to L(PA) and proves $True_{PA}$:

(1) $T_T(PA)^* \vdash \varphi \leftrightarrow Tr(\underline{\langle\varphi\rangle})$, for any sentence φ of L(PA).

(2) $T_T(PA)^* \vdash \forall x\,(Pr_{PA}(x) \rightarrow Tr(x))$.

But from (2) it follows:

(3) $T_T(PA)^* \vdash Pr_{PA}(\underline{\langle 0 = 1\rangle}) \rightarrow Tr(\underline{\langle 0 = 1\rangle})$.

Since, by (1), we can "disquote" the consequent in (3), we have

(4) $T_T(PA)^* \vdash Pr_{PA}(\underline{\langle 0 = 1\rangle}) \rightarrow 0 = 1$,

But given that T_T(PA)* contains PA, we also have

(5) $T_T(PA)^* \vdash \neg(0 = 1)$.

---

[316] See Shapiro (2002), Halbach & Horsten (2002b).

[317] While it is easy to prove this for T(PA) and T*(PA), it is a non-trivial result that this holds also for T_T(PA). The model-theoretic demonstration of this is due to Kotlarski, Krajewski and Lachlan (1981); the proof-theoretic demonstration is due to Halbach (1999).

Applying *modus tollens* to (4) and (5) we finally get:

(6) $T_T(PA)^* \vdash \neg Pr_{PA}(\underline{\langle 0 = 1 \rangle})$ QED.

We have said that when it came to providing a satisfactory theory of truth for a logico-mathematical language (L(T), Tarski required more than material adequacy. He was thus ready to sacrifice conservativeness in favour of deductive capacity. In view of this, $T_T(PA)^*$ comes closest to what he would have deemed a satisfactory axiomatic theory of truth, in spite of the fact that he did not explicitly mention it in CTFL. After all, it amounts to a formalization of his recursive-style truth-definition for L(PA) – without the detour through satisfaction – which he preferred to formalize as a direct definition of '*Tr*' in a higher-order metatheory. To be sure, the conservativeness of T(PA), T(PA)* and $T_T(PA)$ over PA implies that they are consistent provided PA is consistent. But the fact that $T_T(PA)^*$ is not conservative over PA does not mean that we can have no assurance at all of its consistency. Tarski's familiar observation applies: relative consistency of $T_T(PA)^*$ can be established on the basis of a higher order theory.

Indeed, Feferman showed $T_T(PA)^*$ to be equivalent to the subsystem ACA of the 2nd-order arithmetic. ACA has the comprehension axiom for arithmetical sets:[319]

$$\exists X \, \forall y \, (y \in X \leftrightarrow \varphi(y)),$$

where φ does not contain 'X' or any 2nd-order bound 2nd/order variable, plus the full 2nd-order induction:

$$\forall X \, [(X(0) \wedge \forall x \, (X(x) \rightarrow X(Sx))) \rightarrow \forall x \, X(x)],$$

This is quite an interesting observation, because it shows what resources "essentially stronger" than those available in PA are needed to define '*Tr*' satisfying the axioms of $T_T(PA)^*$. Thus, ACA embodies the set-theoretical assumptions required to define that predicate.

On this basis, we could conjecture that Tarski would have been happy with the axiomatic theory $T_T(PA)^*$ but not with T(PA), T(PA)* and $T_T(PA)$, since these are meta-theoretically too weak to count as satisfactory theories of truth. Shapiro and Ketland have argued, on very similar grounds, that the minimal disquotationalist theory of truth T(PA) or any conservative extension of PA cannot be an adequate theory of truth, since no such theory can prove $True_{PA}$ and Con_{PA}. But should this bother the disquotationalist? Can he not choose $T_T(PA)^*$ that is strong enough to prove them?

If we are to believe Shapiro and Ketland, the disquotationalist is committed to conservativeness of his axiomatic theory of truth over the base theory (PA), because the disquotational truth-predicate is supposed to be free of substantive content, serving only as a convenient device of disquotation,

[319] Feferman (1991). It can be shown that ACA's quantification over arithmetical sets can be defined in $T_T(PA)^*$ as quantification over 1-place formulas and membership of n in the arithmetical set as truth of a formula applied to n. In fact, the membership predicate for ACA and the truth-predicate '*Tr*' of $T_T(PA)^*$ for L(PA) are interdefinable.

generalization or blind assertion. A disquotational theory of truth over PA (be it minimal or not) should not prove any substantial (number-theoretic) claims that PA does not already prove. In short: insubstantial theory should not have substantial consequences. Since $T_T(PA)^*$ obviously proves substantial number-theoretic claims such as Con_{PA} (or provably equivalent Gödel-type sentences) that are undecidable in PA, it follows that it cannot be as insubstantial as the minimal disquotationalist theory of truth or any conservative extension of PA. Now, as the disquotational axioms over PA do not suffice to prove $True_{PA}$ and Con_{PA}, but recursive-axioms over PA plus induction for L_T suffice to prove them, $T_T(PA)^*$ seems to put more content into '*Tr*' than is acceptable to the disquotationalist. Shapiro sums up: if the disquotationalist wants to preserve conservativeness he has basically two choices; either to stick to too weak a theory of truth or to adopt a non-effective (non-compact) logic ('*Tr*' would preserve its "thinness" dear to the deflationists, but consequence relation would be "thick", possibly to the point of intractability).

Let's distinguish two questions here. The first question is whether any reasonable theory truth for L(PA) should prove - and not just express - truth-involving generalizations including $True_{PA}$. The second question is whether the deflationist should expect from his preferred axiomatic theory of truth for L(PA) that it be conservative over PA. If the deflationist is committed to the positive answers to both questions, then Shapiro-Ketland arguments show that he wants to have his cake and it eat. Consequently, if he wants to avoid their trap, he has to answer negatively at least one question.

Some deflationists have expressed their commitment to the minimal disquotational theory T(PA), conservative over T but too weak to prove desired generalizations.[320] Others, however, have preferred a theory of truth that proves such generalizations. For the latter, the question arises what specific axioms are available to them and whether they should be conservative (and over what base theory). Field, who expects a good theory of truth to prove such truth-involving generalizations, responds to Shapiro-Ketland argument in the following way:

> "[...] it is quite uncontroversial that the notion of truth can be used to make generalizations that cannot be made without it, and that these generalizations can be important in giving rise to commitments not involving the notion of truth." (Field 1999: 536)

Field's response is that Tarski-style recursive axioms do not put more content into '*Tr*' than the disquotational schema $Tr(\langle\varphi\rangle) \leftrightarrow \varphi$, as witnessed by the fact that $T_T(PA)$ is conservative over PA. So, the disquotationalist may well agree with Shapiro that these purely truth-theoretic axioms are conservative over PA:

> "[...] there is no need to disagree with Shapiro when he says 'conservativeness is essential to deflationism' ." (Ibid: 536)

Granted, $T_T(PA)$ does not prove soundness and consistency of PA, whereas $T_T(PA)^*$, which proves both, is not conservative over PA. Does this fact place the more powerful axiomatic theory beyond the reach of the deflationist? This would follow, according to Field, only if the *purely truth-theoretic* axioms

[320] Cf. Tennant (2001).

were responsible for the substantial (number-theoretic) consequences of $T_T(PA)^*$. Given that the only difference between conservative $T_T(PA)$ and non-conservative $T_T(PA)^*$ is the fact that induction axioms of the latter contain '*Tr*', it rather seems that it is these induction axioms that are responsible for the remarkable increase in the deductive power of $T_T(PA)^*$ compared to $T_T(PA)$. And these induction axioms depend also "if not solely" on the nature of natural numbers, hence they are not purely truth-theoretic. However, no sane deflationist has ever claimed that truth-involving generalizations cannot yield substantial consequences, if such generalization depend also "if not solely" on other matters such as the nature of natural numbers or, perhaps, the behaviour of the provability predicate. Consequently, if the axiomatic theory of *arithmetical truth* such as $T_T(PA)^*$ is not conservative over PA, this might be due to the axioms not essential to truth.

Field seems to have a point here: as a powerful device of generalization, it should not come as a surprise to us that the disquotational notion of truth might help us in establishing "substantial" consequences non involving truth, if combined with other (powerful) mathematical principles. The problem is that Field seems to think that there are purely truth-theoretic principles "essential to truth", which do not depend on other matters such as, say, the nature of natural numbers. To this Halbach objects that not even recursive axioms are purely truth-theoretic in Field's sense, since they "depend" to some extent on the nature of numbers.[321] Moreover, he showed that not even the minimal disquotational theory of truth is not entirely free of ontological commitments, because it can be proved that it is not conservative over 1st order logic (empty 1st order theories). From the minimal disquotational theory of truth it can be deduced that *m* and *n* coding formulas φ and $\neg\varphi$ respectively are distinct numbers; consequently, *there are at least two objects*. And this is a non-logical claim, on the prevailing view.

The question now is whether this should worry the deflationist. Already Tarski taught us that a formal theory of truth comes always with certain commitments (viz. the metatheory containing the syntactic theory of the object-language). If so, Halbach remarks, it should not come as a surprise to us when the theory of truth unfolds its commitments. Given that even the minimal disquotational theory of truth is trivially non-conservative over pure logic, it is not charitable to saddle the disquotationalist with such a commitment. Should the theory of truth be at least conservative over the syntactic theory of type-sentences (over PA - the two being interpretable in each other)? As conservative axiomatic theories (over PA) are metatheoretically inadequate, Shapiro assumes that anybody wanting a reasonable theory of truth, the deflationist included, is committed to a non-conservative theory along the lines of $T_T(PA)^*$ (or something of its sort), provided that one works with effective consequence relation. It follows according to Shapiro that when the notion of truth is axiomatized in this manner, the resulting theory is substantial, having substantial consequences going well beyond PA (viz. the deductive power of ACA). Now, Field and Halbach seem to grant the first claim, but they do not accept the conclusion that Shapiro deduces from it. There is no need for the deflationists to claim that the notion of truth should not be any use in proving interesting "truth-free" claims.

---

[321] Halbach (2001: 179, 187).

All parties to the dispute – starting with Tarski – agree that truth might be a proof-theoretically robust notion, if it is added to a base theory along with recursive and induction axioms containing it:

> "Although deflationist truth may be 'only' a device for generalizing, it is not innocent in its arithmetical consequences. The purpose assigned to truth by the deflationist is quite simple: it is 'only' generalization, not the expression of a correspondence relation, nothing deeply entrenched in causal relations, and so on. But there is a lot to this simple purpose." (Halbach 2002: 187)

But the debate generally suffers from vagueness of the *substantial-insubstantial* distinction. As Halbach concedes:

> "But it should be added that there are similar notions – like membership in arithmetical sets – sharing this feature with truth that are usually not described as 'thin' and 'unsubstantial." (Ibid: 187)

Does this discredit the main tenet of deflationism, which is that truth is a logical or logico-mathematical notion with important expressive uses? Well, it should be noted that disquotational and Tarski style theories of truth do not explicate truth in terms of robust properties or relations: causal, physical, etc. In this sense, then, we can well say that they explicate truth as a metaphysically *thin* property. But that does not mean that truth is a *thin* property in the sense of being logico-mathematically sterile, since Tarski and many (though not all) disquotationalists prefer a theory of truth over a logico-mathematical base theory that allows us to establish truth-free claims belonging to the base theory that the base theory does not prove. Now, taking this seriously, T-biconditionals cannot be all that there is to the notion of truth (or, at least, they cannot be all that there is to the theory of truth). Some claims of Tarski and the disquotationalists are highly misleading in this respect.

7.8 Is Tarski's conception of truth deflationary ?

Is Tarski's conception of truth deflationary? Well, that depends on what one understands under "deflationism". Quine said that to ascribe truth to 'Snow is white' amounts to attributing whiteness to snow. Redundancy theorists made similar claims, though they usually directed them at the notion of truth as applied to propositional contents (or beliefs). Keeping this in view, note the striking analogy with the following claim of Tarski (the italics is mine):

> "Consider a sentence in English whose meaning does not raise any doubts, say the sentence snow is white. For brevity we denote this sentence by 'S', so that 'S' becomes the name of the sentence. We ask ourselves the question: What do we mean by saying that S is true or that it is false? The answer to this question is simple: in the spirit of Aristotelian explanation, *by saying that S is true we mean simply that snow is white*, and *by saying that S is false we mean that snow is not white*. By eliminating the symbol S we arrive at the following formulations:
>
> (1) "Snow is white" is true if and only if snow is white.

(1') "Snow is white" is false if and only if snow is not white.

> Thus (1) and (1') provide satisfactory explanations of the meaning of the terms true and false when these terms are referred to the sentence "snow is white"." (Tarski 1969: 103-104).

This and similar passages lend some credence to the interpretation of Tarski's conception of truth as a disquotationalist theory of truth. In particular, the special role played in it by T-schema and its instances, as well as its alleged philosophical neutrality, have been influential among the deflationists. If Tarski's theory is a brand of deflationism, then it is a sentential variety. So much should be clear, though it is possible provide similar truth definitions for other kinds of truth-bearers (given that such-and-such conditions are satisfied). What next? Next comes Tarski's suggestion that T-biconditionals fix the meaning of 'true' – and not just its extension - for L, the truth definition for L being materially adequate if and only if it subsumes all T-biconditionals for L as its deductive consequences. He called particular T-biconditionals partial definitions of 'true' for particular sentences (of L), each explaining the meaning of 'true' with respect to one particular L-sentence (in as clear terms as are used in the L-sentence itself). The above quoted passage even suggests that the sentence mentioned on the left side of

'Snow is white' is true iff snow is white,

says the same thing as the sentence used on the right side, but that the whole sentence on the left side says the same as the sentence on the right side. This is a rather strong claim. One could object that the two sentences cannot mean the same, on the ground that one might understand the left-side sentence without understanding the right-side sentence (say, if one knows some English, 'true' included', but does not know what 'snow' or 'white' means in English). Alternatively, one could deny their synonymy, by invoking a variant of Church-Langford translation test. In fact, except for the above quoted passage, there is little evidence that Tarski was committed to such a strong claim.

Still, Tarski was committed to the claim that T-biconditionals partially define (axiomatize) the notion of truth, which might seem enough to render his conception of truth deflationary in the disquotationalist way. However, there are three main problems with this quick conclusion.

The first problem is that Tarski did not mean to restrict his theory of truth to disquotational (immanent) notions of truth. A large part of CTFL is concerned to show how to define the notion of truth for a properly formalized L in a different and logically stronger meta-L. Indeed, his general paradigm is the schema

X is true (in L) iff p,

> where 'X' stands in for a syntactically perspicuous meta-L designator of an L-sentence and 'p' for an meta-L translation of that sentence.

So understood, the truth-schema may not be to the disquotationalist liking,

because he or she is typically uneasy about inter-linguistic sameness of meaning (translation). Consequently, it is misleading to call Tarski's conception of truth "disquotational".

The second problem is that Tarski's method of truth definition is seriously limited, being applicable only to languages with the right type of extensional structure. But the deflationists typically want to have a theory of truth for natural languages, which (a) contain many constructions for which no plausible compositional methods have yet been found (maybe none will be found), (b) contain their own truth-predicate, and (c) for which there are no "essentially richer" metalanguages. Consequently, if one wants a deflationary theory (or definition) of truth for such languages, one has to look elsewhere.

The third problem is that Tarski noted that the minimal disquotational theory of truth for L(T) is not adequate as a theory of truth for a reasonably rich L(T), since it does not prove recursive clauses, and hence other truth-involving generalizations and important metatheorems. Despite his notorious claim to the effect that a formally correct and materially adequate theory of truth for L(T) is one that has all T-biconditionals for L(T) among its deductive consequences, we have seen that Tarski would not have considered the minimal theory satisfactory, since he advocated logico-mathematically "robust" theories of truth. It continues to be the subject of ongoing controversy whether such a theory can be deemed deflationary - and in what sense - given that it proves "truth-free" sentences of L(T) not provable in T. It may appear inflationary to those truth-deflationists, who claim that only minimal-conservative theories of truth are, strictly speaking, deflationary, but it may well appear deflationary in spirit to those deflationists, who take truth (and related semantic properties) to be a metaphysically thin but logico-mathematically thick property.[322]

Tarski's theory of truth might not the best choice for the deflationist, for the reasons spelled out above. But it can well be regarded a deflationary theory, because it does not explicate truth as a metaphysically thick property, though it sometimes employs a heavy logico-mathematical machinery. I have attempted to show that Tarski's theory was not designed to answer big foundational (meta-semantic) questions, which Field, Putnam and others would expect a theory of truth to answer. However, this can be seen as its laudable feature: abstracting from linguistic practices and assuming meaning properties of L to be fixed and known, we can define truth for L according to Tarski's routine. In other words, we can define restricted semantic predicates with provably right extensions in a mathematical manner, without having to bother about how truth and related semantic properties depend or supervene on speaker's linguistic practices. This should not come as a surprise to us, given that Tarski was primarily interested to provide a mathematically precise theory of truth (and truth-theoretic semantics in general), which project surely does not require that the foundational questions be answered. Under this deflationary reading of his theory of truth, all questions about the metaphysical or epistemological status of meaning, content and semantic properties are to be attacked at a different level. Along with Carnap, Tarski sharply separated formal-semantic from metasemantic questions.

---

[322] Conservative over a non-semantic background theory (such as PA or some physical theory), not over logic.

[8]

Conclusion

My main aim in this work has been to give a systematic, careful and critical examination of its nature and significance, based on the thorough exposition of its historical, conceptual and technical underpinnings. Having explained the conceptual background of Tarski's conception of truth and his method of truth definition for increasingly more complex formalized languages (Chapters 2-3), I argued (Chapter 4) that its logico-mathematical import consists mainly in his systematic method of formalization, indeed, mathematization of informal metamathematical ideas of the semantic variety. Its fruit was a greater precision in metamathematics: namely, precise definitions of fundamental metalogical notions and exact formulations and proofs of fundamental metalogical results couched in terms of such notions. In Chapter 5, I dealt with the question to what extent Tarski's CTFL (and related articles from the 1930s) anticipates the modern model-theoretic approach, and what elements might be missing from it. The main conclusions of my discussion were as follows: (1) in the 1930s, Tarski did not yet have the full-blooded model-theoretic notion of truth in a structure, since he still held Frege-Peano view of language as a meaningful formalism and subscribed to the doctrine of absolute truth (as a property applying to fully interpreted sentences), and accordingly did not have the modern notion of uninterpreted non-logical constant. (2) Partly for this reason and partly because he held the fixed-domain conception of models, his account of logical consequence in (1936a) is not to be identified with the modern model-theoretic account of consequence (although it seems that this does not create as many problems as some critics – e.g. Etchemendy (1988) – suspected). Already in Chapters 4 and 5 I hinted that Tarski's method has a "deflationary" character in that it is, in the first place, a logico-mathematical theory designed to serve logico-mathematical needs, and not to answer so-called foundational (metasemantic) questions.

In Chapter 6 I reviewed a number of objections and arguments that purport to show that Tarski's method of truth definition fails as an explanation (explication) of our common notion of truth, and, in particular, that it is a confusion to think that Tarski's truth definitions have semantic interest. I argued that the critics are right to say that particular truth definitions in Tarski-style do not explain our common notion of truth, but it does not follow that we cannot think of Tarski's method of truth definition as giving us a valuable insight of a different sort: a workable model of how truth conditions of sentences of a properly formalized language depend on semantic properties of their significant

parts and syntactic structure. Indeed, this is how Tarski's method has been viewed by those theorists who see in it the foundation of formal semantics (though requiring further modifications, to be sure). What the critics do not appreciate is that there is more to Tarski's conception of truth than the particular formal definitions for particular languages. Its heart is the material adequacy criterion stated in Convention T, which assures that a truth definition that meets it captures all and only the true sentences of a given language (the extension of truth with respect to that language), where a good intuitive grasp of the ordinary notion of truth is presupposed in the form of the semantic conception of truth. The standard formal technique – for reasonably rich languages - is recursion on a generalized semantic relation between expressions and objects, in terms of which truth for the language is defined in such a way that materially correct statements of truth-conditions can be delivered for each of the indefinite number of its sentences. In tandem, the two moments indicate to us where to look for the semantic import of Tarski's method, in which its philosophical value largely consists. I attempted to show that Tarski's method of truth definition has logico-mathematical as well as philosophical aspects, trying to persuade the reader that once we understand and distinguish these aspects, its contribution to semantics dwells in the fact that a recursive truth-definition for a reasonably complex L is equivalent to a compositional axiomatic truth-theory for L – with semantic terms construed as its primitives – which illuminates the compositional semantic structure of L.

On the other hand, in Chapter 7 I wanted to substantiate the announced claim that Tarski's method of truth does not give us satisfying answers to foundational or metasemantic questions such as:

> What facts about usage (if any) determine L's semantics (intended interpretation)?
>
> On the basis of what evidence can we tell that a truth theory (or semantics) for L is correct?
>
> ....

Field famously argued that Tarski's truth definition is only a partial success, on the ground that it does not provide a *genuine reductive explanation* of primitive denotation (nominal denotation, predicative application and functional fulfilment) in terms of scientifically respectable notions. I agreed: it does not. Mere lists - here base-clauses for predicates or terms - do not provide genuine explanations. However, I found myself in agreement with Soames in that it is a laudable feature of Tarski's method of truth definition that it sharply separates metasemantic from formal-semantic issues, allowing us to deal with formal-semantic issues in a mathematically precise manner. Tarski's conversion to the model theoretic approach tallies well with this approach to semantics, which I therefore call "deflationary". It should be clear that a formal interpretation of a theory of truth for L(T) in set theory - via explicit definitions of semantic notions in terms of primitive notions of set theory – can neither answer deep foundational questions of philosophers nor explain the meaning of our common semantic notions. I agree that some claims of Tarski are misleading in this respect. But then they are aberrations on his part that, in my opinion, do not reveal his considered philosophical position. Finally, having explained the basic

ideas animating modern deflationism (in particular, disquotationalism), I compared it with Tarski's conception of truth, of which some deflationists even claimed that it is a paradigmatic deflationary theory of truth. My conclusion was that it is problematic to take Tarski's theory of truth to be deflationary in the disquotationalist sense. Still, it can well be regarded a deflationary theory in my preferred sense, because it arguably abstracts from the so-called meta-semantical issues concerning the metaphysical or epistemological basis or status of semantic properties.

By way of conclusion, I should say that in spite of the fact that Tarski's method of truth definition has the deflationary flavour it has turned out that its formal methods can be interpreted in several different ways, some of them deflationary, others more substantive. Davidson, for instance, has long tried to persuade us that the heart of Tarski's method is recursion (not elimination), and, accordingly, that we can look at the clauses of the recursive definitions as axioms with the primitive notion of truth (or formalize them as axioms) having an empirically confirmable content (via his theory of interpretation). This may well be at odds with Tarski's "deflationary" claim that T-biconditionals are definitional of truth, but it is in my opinion a legitimate way of using Tarski's formal structure. Field's early naturalistic program in semantics may be an alternative way of interpreting and using the same formal structure, supplemented by explanatory reductions of primitive denotation. While Davidson's holistic framework puts stronger emphasis on the role of recursive structure (the notions of reference and satisfaction are primitive but instrumental with respect to the notion of truth), Field's atomistic framework puts stronger emphasis on base clauses for terms and predicates, "reducing" truth to denotation and application (a physicalist correspondence theory of truth without facts). A serious challenge to both conceptions may come from the deflationists, who claim that no substantial empirical content is to be read into the clauses of Tarski's formal structure, because these clauses are definitional of truth (related semantic notions), playing no genuine explanatory role there. On the other hand, the deflationist theories face serious difficulties, some of which were reviewed in Chapter 7. Moreover, the proponents of Davidson's or Field's approach to semantics may argue that semantic notions are not as insubstantial as the deflationists believe, if they play the explanatory role in the semantic theory that they reserve for them. Fortunately, it was not my goal in this work to decide the extremely difficult question as to which party is more right about truth and semantic notions in general.

Appendix

1 Tarski's truth definition for the language of calculus of classes (LCC)

In Part III of CTFL, Tarski shows how to define the predicate '*Tr*' – or the sentential function '$x \in Tr$' – whose extension contains all and only the true sentences of LCC. The calculus of classes - the deductive theory built-in LCC - is a rather weak fragment of the system of simple (finite) theory of types assumed in CTFL, with variables interpreted as ranging over *classes of elements* of the universal domain of the type-theoretic system and one primitive 2-place predicate for class inclusion between such classes. LCC is syntactically easy to handle, containing a few constants and operations (syntactic constructions) for forming complex expressions:

a) a (countable) sequence of variables 'x''', 'x'''', 'x''''',…., each variable being formed by appending n strokes to 'x' (for $1 \leq n$) - the variable with n-strokes is referred to as 'x_n'.[323]

b) the logical constants '*N*', '*A*', and '*Π*' (throughout CTFL Tarski uses the Polish notation due to Lukasiewicz) and the 2-place predicate '*I*' of class-inclusion;[324]

The metalanguage in which the metatheory for LCC is framed contains the following signs:

a) signs translating the constants of LCC: 'not' ('it is not the case that'), 'or', 'for all', 'is included in';

b) signs for the usual set-theoretic notions: '∈', 'individual', ' is identical' (=), 'class', 'cardinal number', 'domain', 'ordered n-tuple', 'infinite sequence', 'relation', etc.

c) signs by which the structural-descriptive names of LCC-expressions are formed: '*ng*' (for *negation*); '*sm*' (for *disjunction*], '*un*' (for *universal quantification*), 'v_k' (for *the k-th variable*), '$x \hat{} y$' (for: *the expression consisting of the expression* 'x' *followed by the expression* 'y'), etc.

d) conventions for abbreviation, based on (c):

[323] For typographical convenience I use strokes as superscripts and not as subscripts (as Tarski does in CTFL). The same applies to two other signs to be introduced: the metalinguistic sign for concatenation and for the metalinguistic sign for negation operation.

(i)	$x = \iota_{k,l}$	iff	$x = (in\hat{}v_k)\hat{}v_l$,
(ii)	$x = \underline{y}$	iff	$x = ng\hat{}y$;
(iii)	$x = y + z$	iff	$x = (sm\hat{}y)\hat{}z)$;
(iv)	$x = \cap_k y$	iff	$x = (un\hat{}v_k)\hat{}y$.[325]

Except of general logical axioms and axioms translating axioms specific to the calculus of classes, the meta-language contains also axioms that form the syntactic theory of LCC. In particular, Tarski defines the set *E* of expressions of LCC such that (a) *E* contains the distinct signs '*sm*', '$\cap$', 'ι', 'v_k', (b) *E* contains 'v_k', if *k* is a positive integer distinct from 0, (c) $x\hat{}y$ belongs to *E*, if *x* and *y* both belong to *E*, and (d) nothing belongs to E except what belong to E by (a),...,(c).[326] Having thereby laid down the first rigorous axiomatization of concatenation theory (also called "a theory of strings"), Tarski provides the inductive definition of sentential functions of LCC:

> ***f* is a sentential function (of LCC)** iff one the following conditions is satisfied:
>
> (a) $f = \iota_{i,j}$ [i.e. $(in\hat{}v_i)\hat{}v_j$] for some positive integers *i* and *j*;
>
> (b) $f = \underline{y}$ [i.e. $ng\hat{}y$], for some sentential function *y*;
>
> (c) $f = y + z$ [i.e. $(sm\hat{}y)\hat{}z)$], for some sentential functions *y* and *z*;
>
> (d) $f = \cap_k y$, [i.e. $(un\hat{}v_k)\hat{}y$], for some positive integer *k* and sentential function *y*.

Tarski then points out that, like most inductive definitions given in that section, this definition can be converted into an explicit definition of the smallest set *X* such that (a) *X* contains every simple sentential function of the form $\iota_{i,j}$ (for some positive integers *i* and *j*) and (b) *X* is closed under the operations of negation, conjunction and universal quantification with respect to the *i*-th variable (*X* contains $\underline{y}$, $y + z$, and $\cap_k y$, whenever *X* contains *y* and *z*). Having defined what it takes for a variable to have a free occurrence in a sentential function:

> **The variable v_i is free in the sentential function *f*** iff *i* is a positive integer and one of the following conditions is satisfied:
>
> (a) $f = \iota_{i,j}$ or $f = \iota_{j,i}$, for some positive integer *j*;

---

[326] He also formulates the following fundamental law of concatenation (viz. Axiom 4, Tarski 1983: 173):
If *x*, *y*, *z* and *t* are expressions, then we have $x\hat{}y = z\hat{}t$ iff one of the following conditions is satisfied:

(a) $x = z$ and $y = t$;
(b) $x = z\hat{}u$ and $t = u\hat{}y$, for some expression *u*;
(c) $z = x\hat{}u$ and $y = u\hat{}t$, for some expression *u*.

(b) $f = \underline{y}$, for some sentential function y, and v_i occurs free in y;

(c) $f = y + z$, for some sentential functions y and z, and v_i occurs free in y or v_i occurs free in z;

(d) $f = \cap_k y$ for some positive integer k and sentential function y, and v_i occurs free in y and $i \neq k$,

sentences (closed sentential functions) of LCC are defined as those with no variable free:

***s* ∈ *S* (the set of sentences of LCC) iff** *s* is a sentential function of LCC and *s* contains no free variables.

Satisfaction-relation is then inductively defined: it is first specified under what conditions an arbitrary sequence *s* of classes satisfies simple open-sentences of the form $\iota_{j,i}$, and then the conditions are specified under which *s* satisfies complex open-sentences of the form $\underline{y}$, $y + z$, and $\cap_k y$ in terms of the conditions under which *s* satisfies (or does not) their component sub-sentences.

The infinite sequence of classes *p* satisfies the sentential function *f* (of LCC) iff one of the following conditions is satisfied:

(a) $f = \iota_{i,j}$, for some positive integers i and j, and p_i is included in p_j;

(b) $f = \underline{y}$, for some sentential function y and it is not the case that p satisfies y;

(c) $f = y + z$, for some sentential functions y and z, and p satisfies y or p satisfies z;

(d) $f = \cap_k y$, for some positive integer k and sentential function y, and every infinite sequence of classes p^* satisfies y, which differs from p at most at the k-th place.

Once again, the recursive definition can be turned to an explicit definition, according to which

The infinite sequence of classes *p* satisfies the sentential function *f* (of LCC) iff $\langle p, f \rangle$ belongs to every set S such that, for every r and q, $\langle r, q \rangle$ belongs to S iff q is a sentential function q and r is an infinite sequence of objects, and one of the following conditions are satisfied for S:

(a) $q = \iota_{i,j}$, for some positive integers i and j, and r_i is included in r_j;

(b) $q = \underline{y}$, for some sentential function y, and it is not the case that $\langle r, y \rangle$ belongs to S;

(c) $q = y + z$, for some sentential functions y and z, and $\langle r, y \rangle$ belongs

to S or $\langle r, z\rangle$ belongs to S;

(d) $q = \cap_k y$, for some positive integer k and sentential function y, and $\langle r^*, y\rangle$ belongs to S, for every infinite sequence of classes r^* which differs from r at most at the k-th place.

Finally, truth for sentences of LCC is defined directly as a limiting case of satisfaction by all sequences of objects:

***x* ∈ *Tr* (the set of true sentences of LCC)** iff x is a sentence of L and $\forall y$ (sequence $(y) \rightarrow y$ satisfies x).

This definition can be turned into non-semantical definition, if we replace 'satisfies' in it in accordance with the previous definition.

2 Material adequacy

In order to "empirically' confirm material adequacy of the truth definition for LCC, Tarski shows how the following T-biconditional follows from it:

(*) $\cap_1 \cup_2 \iota_{1,2} \in Tr$ iff for every class a there is a class b such that $a \subseteq b$.

Now, '$\cap_1 \cup_2 \iota_{1,2}$' is designed to serve as a revealing (i.e. structural-descriptive) name in ML of the following sentence

$$\prod x' N \prod x'' N I x' x''$$

of LCC, based on the conventions introduced above. As for this LCC-sentence, '$\prod x''$' reads *for every class* x_1, 'N' reads *not* (or: *it is not the case that*) and '$I x' x''$' reads x_1 *is included in* x_2, and the whole sentence reads:

For every class x_1, not every class x_2 is such that x_1 is not included in x_2.

Since, now, the operation of existential quantification with respect to the i-th variable is equivalent to (hence can be introduced in terms of) '$N \prod N x'^{\cdots}$' (i-1 strokes following the first stroke), the sentence can be given an equivalent but more natural reading:

For every class x_1, there is a class x_2 such that x_1 is included in x_2,

According to Tarski's conventions, '$\cup_2$' describes '$N \prod N x'''$', so the operation of existential quantification with respect to the i-th variable, whereas the operation of universal quantification with respect to the i-th variable is referred to by '$\cap_i$'). Given that '$x \in Tr$' reads *x belongs to the set of true sentences* (in this context: of LCC), or, more simply, *x is a true sentence*, the left side of the equivalence is to be read as follows: *the expression consisting of the sign for universal quantification followed by the sign for the 1st variable, followed by the sign for existential quantification, followed by ...., is a true sentence (of LCC).* Here, then, is Tarski's informal justification of (*):[327]

"According to the Def. 22 the sentential function $\iota_{1,2}$ is satisfied by those and only those sequences f such that $f_1 \subseteq f_2$. So the negation

[327] For typographical convenience, I have replaced overlined by underlined sentential functions – viz '$\underline{y}$'.

> $\underline{\iota_{1,2}}$ is satisfied by exactly those sequences f such that $\underline{f_1 \subseteq f_2}$. Consequently a sequence f satisfies the [sentential]function $\cap_2\underline{\iota_{1,2}}$ if every sequence g which differs from f in at most the 2nd place satisfies the function $\underline{\iota_{1,2}}$ and thus verifies the formula $\underline{g_1 \subseteq g_2}$. Since $g_1 = f_1$ and the class g_2 may be quite arbitrary, only those sequences f satisfy the function $\cap_2\underline{\iota_{1,2}}$ which are such that $\underline{f_1 \subseteq b}$ for any class b. If we proceed in an analogous way, we reach the result that the sequence f satisfies the function $\cup_2\iota_{1,2}$, i.e. the negation of the function $\cap_2\iota_{1,2}$, only if there is a class b for which $f_1 \subseteq b$ holds. Moreover the sentence $\cap_1\cup_2 \iota_{1,2}$ is only satisfied (by an arbitrary sequence f) if there is for an arbitrary class a, a class b for which $a \subseteq b$. Finally by applying Def. 23 we at once obtain one of the theorems which were described in the condition (α) of the convention T:
>
> > $\cap_1\cup_2 \iota_{1,2} \in Tr$ iff for every class a there is a class b such that $a \subseteq b$." (Tarski 1983: 196).

At a crucial point in this justification, Tarski makes use of the same idea that we have seen at work in our informal proof in Chapter 3: quantification over all k-variants of a sequence f does the same job as quantifying over arbitrary objects over which the quantifiers range:

> "Since $g_1 = f_1$ and the class g_2 may be quite arbitrary, only those sequences f satisfy the function $\cap_2\iota_{1,2}$ which are such that $f_1 \subseteq b$ for any class b."

Strictly speaking, in order to deduce from the metatheory augmented with the truth-definition T-biconditionals for sentences of the form $\forall v_i A$ that do not mention infinite sequences, we need to prove in the metatheory an instance of the following schema for each given sentence in question:

$\forall s^*[\forall k\,(k \neq i \rightarrow s(k) = s^*(i)) \rightarrow A]$ iff $\forall x_i A$.

3 Satisfaction and correctness in an individual domain

It is now easy to provide the rigorous definitions of satisfaction and correctness in an individual domain a, where a is a subclass of the universal domain containing all (arbitrary) individuals. In fact, we should only to properly relativize the clauses in the recursive definition of satisfaction *simpliciter* (as it were, satisfaction of a sentential function in the universal domain):

> **The infinite sequence of classes p satisfies the sentential function f (of LCC) in the individual domain a** iff a is class of individuals, p an infinite sequence of subclasses of a and f a sentential function such that the following conditions are satisfied:
>
> (a) $f = \iota_{i,j}$, for some positive integers i and j, and p_i is included in p_j;
>
> (b) $f = \underline{y}$, for some sentential function y and it is not the case that

p satisfies *y* in *a*;

(c) $f = y + z$, for some sentential functions *y* and *z*, and *p* satisfies *y* in *a* or *p* satisfies *z* in *a*;

(d) $f = \cap_k y$, for some positive integer *k* and sentential function *y*, and every infinite sequence of classes *p** satisfies *y* in *a*, which differs from *p* at most at the *k*-th place.

The explicit definition is easy to give, except that now we have to say that the infinite sequence *p* of subclasses of *a* satisfies the sentential function *f* in *a* iff the ordered triple $\langle p, f, a\rangle$ belong to every set *S* satisfying certain obvious conditions, which reflect the role of the additional parameter – i.e. *a*. Relative correctness (or, relative truth, as Tarski sometimes says:) of sentences of LCC in *a* is defined directly as satisfaction of the sentence by all sequences of objects in *a*:

> ***x* is a correct sentence in the individual domain *a*** iff *x* is a sentence (of LCC) and every infinite sequence of sub-classes of *a* satisfies *x* in *a*.

Consequently, (1) a sentence (of LCC) is correct in an individual domain with *k* elements iff if it is correct in some individual domain *a* such that *a* has *k* elements; (2) a sentence (of LCC) is correct in every individual domain (is universally valid) iff it is correct in *a*, for every individual domain *a*; and (3) a sentence (of LCC) is true (*simpliciter*) iff it is correct in *a* such that *a* is the universal domain of all individuals.

4 Tarskian truth definition for the language of set theory

Finally, let us see what happens when we attempt to define in Tarski-style truth for the standard 1st-order language L of set-theory (ZF), whose signature $\{\in\}$ contains just one sign for the set-theoretical relation of *elementhood*, and does not otherwise differ in its logical basis from our simple 1st-order language L_2 introduced in Chapter 3, except that it contains the sign for identity ('='). I shall not bother to lay down the syntactic definitions for this language, as this poses no special difficulties. We just take the recursive definition of *sentential function* that we gave for L_2, we replace the base clauses (a), (b) and (c) as follows

> ***f* is a sentential function (of L)** iff one of the following conditions is satisfied:
>
> a*) *f* is $v_i \in v_k$, for some positive integers *i* and *k*;
>
> b*) *f* is $v_i = v_k$, for some positive integers *i* and *k*;
>
> ……

The recursive clauses are the same as in the definition for L_2 (as well as the definitions of free variable in a sentential function and of a sentence). Now, the

recursive definition of satisfaction for L is as follows:

> **The infinite sequence of sets *p* satisfies the sentential function *f* of L** iff one of the following conditions is satisfied:
>
> (a) *f* is $v_i \in v_k$, and p_i is an element of p_k;
>
> (b) *f* is $v_i = v_k$, and p_i is the same as p_k;
>
> (c) *f* is $\neg A$ and *p* does not satisfy *A;*
>
> (d) *f* is $A \wedge B$ and *p* satisfies *A* and *p* satisfies *B*;
>
> (e) *f* is $A \vee B$ and *p* satisfies *A* or *p* satisfies *B*;
>
> (f) *f* is $\forall v_k A$ and every infinite sequence of sets *p** satisfies *A* that differs from *p* at most at the *k*-th place.

By the free variable lemma, sentences (of L) are true iff they are satisfied by all/some sequences of sets.

The question now arises whether we can turn the recursive definition to an explicit, set-theoretical definition of truth for L, based on the explicit set-theoretical definition of satisfaction. Recall that Frege-Dedekind procedure can be employed with respect to (D5) (for L_2) or with respect to the recursive truth definition for L(PA) in Chapter 5 relative to the standard/intended interpretation, since the domains of their quantifiers are restricted (form a set) and we assume that a set theory in which we carry out the procedure is essentially logically richer than L(PA) or L_2 in that it allows us to quantify over arbitrary subsets of their respective quantifier-domains. If, now, we could effectively use the procedure with respect to L, we would have a truth definition for L within L! But Tarski's indefinability of truth theorem tells us that this is impossible, on pain of inconsistency. Assuming, then, that the standard set theory (ZF) is consistent, there must be a problem to be identified. Considering the following attempt to construct an explicit definition of set-theoretical truth within set-theory (strictly speaking, we employ an informal set-theoretical language):

> **The infinite sequence of sets *p* satisfies the sentential function *f* of L** iff $\langle p, f \rangle$ belongs to every set S such that, for every *r* and *q*, $\langle r, q \rangle$ belongs to S iff *q* is a sentential function *q* and *r* is an infinite sequence of sets, and one of the following conditions are satisfied for S:
>
> (a) *q* is $v_i \in v_k$, and r_i is an element of r_k;
>
> (b) *q* is $v_i = v_k$, and r_i is the same as r_k;
>
> (c) *q* is $\neg A$ and $A \notin S$*;*
>
> (d) *q* is $A \wedge B$ and both $<r, A> \in S$ and $<r, B> \in S$;

(e) q is $A \vee B$ and $\langle r, A\rangle \in$ S or $\langle r, B\rangle \in$ S;

(f) q is $\forall v_k A$ and $\langle r^*, A\rangle \in$ S, for any infinite sequence of sets r^* that differs from r at most at the k-th place.

we observe that no set S satisfies these conditions: if there were a set satisfying the conditions, S would contain $\langle r, q\rangle$, for any sequence of sets r whatever, including sequences of sets that have the same rank as S (indeed, since the definition places no restriction at all on r, there would be a sequence of sets r among whose terms is S itself; but such r cannot have a lower rank than S). But this is absurd, since S cannot have the same rank as r if $\langle r, q\rangle \in$ S. So there is no S satisfying the conditions. And if there is no such set, then every sentential function of L is satisfied by every sequence of sets, hence every sentence of L is true - which is absurd. It is well known that we could alternatively use a would-be truth definition that requires the existence of a set S satisfying the conditions spelled out in the above definition. But then it would follow that since there is no set satisfying the conditions, no sentence of L is true - which is an equally embarrassing result. The moral is that the satisfaction relation on L is not a set but a proper class. We have seen in (4.2) that Tarski showed (Theorem II) that for any given natural number k, we can define within the (simple type theoretic) general calculus of classes satisfaction and truth for any sub-language of LGC that contains only sentences with variables whose order is at most equal to k. Something more general also holds goods: for any definite ordinal k, if the quantifiers of the set-theoretical language L are restricted to range only over sets of a rank $R<k$, then satisfaction and truth for L are explicitly definable within the standard 1st order set theory (ZF). We can ask whether the satisfaction relation for L(ZF) with unrestricted quantifiers (its domain being a proper class) is definable in a stronger system. Arguably, it can be defined within 2nd-order set theory, when 2nd-order variables interpreted as ranging over classes or as devices of plural quantification, or in Morse-Kelley 1st order set theory that allows quantification over proper classes in the comprehension axiom (in contradistinction to Gödel-Bernays-von Neumann se theory that does not allow quantifiers to range over proper classes in the comprehension axiom). But then the question arises whether we can define truth for such stronger systems. Recall here the discussion in section (4.2).

Bibliography

AJDUKIEWICZ, K.(1935): “Die syntaktische konnexität.” *Studia Philosophica* (1): 1–27. English translation “Syntactic Connection” published in McCALL (1967), pp. 207-231.

ALMONG, J., PERRY, J & WETTSTEIN, H., eds. (1989): *Themes from Kaplan*. Oxford University Press, Oxford.

ARISTOTLE (1923): *Metaphysics*. Clarendon Press, Oxford. Transl. by R. Kirwan.

— (1963): *Categoriae et Liber de interpretation*. Oxford University Press, Oxford. Transl. by J. L.Ackrill.

ARMOUR-GARB, B. P. & JC BEALL, eds. (2005): *Deflationary Truth*. Open Court, Chicago and La Salle, Illinois.

AUXIER, R. E. & HAHN, L. E., eds. (2006): *The Philosophy of Jaakko Hintikka*. *Library of Living Philosophers*. Vol 30. Open Court, Chicago.

AWODEY, S & A.W. CARUS (2007): “Carnap’s dream: Gödel, Wittgenstein, and Logical Syntax.” *Synthese* 159 (1): 23-45.

AWODEY, S (2007): “Carnap’s Quest for Analyticity: the *Studies in Semantics*.” In FRIEDMAN & CREATH (2007), pp. 226-247.

AYER, A. J. (1935): “The Criterion of Truth.” *Analysis* (3): 28–32.

— (1952): *Language, Truth and Logic* (reprint of 2nd edn). Dover Publications, New York.

BARWISE, J., ed. (1977): *Handbook of Mathematical Logic*. North Holland, Amsterdam.

BAYS, T. (2001): “On Tarski on models.” *Journal of Symbolic Logic* (66): 1701–1726.

— (2009): “Beth’s Theorem and Deflationism.” *Mind* 118 (472): 1061-1073.

BEALL, JC, ed. (2005): *Deflationism and Paradox*. Clarendon Press, Oxford.

BELNAP, N. & A. GUPTA (1993): *The Revision Theory of Truth*. MIT Press, Cambridge.

BELNAP, N. (1999): “On Rigorous Definitions.” *Philosophical Studies* (72): 115–146.

BETH, E. W. (1953): “On Padoa’s method in the theory of definition.” *Indagationes Mathematicae* (15): 330–339.

BLACK, M (1948): “The semantic definition of truth.” *Analysis* (8): 49-63.

BLACKBURN, S. & K. SIMMONS, eds. (1999): *Truth*. Oxford University

Press, Oxford.

BLACKBURN, S. (1984) *Spreading the Word*. Oxford University Press, Oxford.

BOLZANO, B. (1837): *Wissenschaftslehre: Versuch einer ausführlichen und grösstentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter*. Seidel, Sulzbach. Partial English translation 'Theory of Science', ed. Rolf George, University of California Press, Berkeley, 1972.

BOOLOS, G. (1998): *Logic, Logic, and Logic*. Harvard University Press, Cambridge MA. Ed. by R. Jeffrey.

BOOLOS, G, R. JEFFREY & BURGESS, J. (2002): *Computability and Logic* (4th edn). Cambridge University Press, Cambridge.

BURGESS, J. P. (2008a): *Mathematics, Models, and Modality: Selected Philosophical Essays*. Cambridge University Press, Cambridge.

— (2008b): "Tarski's Tort." In BURGESS (2008a), pp. 149-168.

BUTLER, R. J., ed. (1962): *Analytical Philosophy*. Blackwell, Oxford.

CANDLISH, S. & N. DEMNJANOVIC (2007): "A Brief History of Truth." In JACQUETTE (2007), pp. 273–369.

CARNAP (1934): *Logische Syntax der Sprache*. Springer, Vienna. English translation of an enlarged version: *Logical Syntax of Language*, Routledge and Kegan Paul, London, 1937. Page references are to the latter.

— (1936): "Wahrheit und Bewährung." *Actes du Congres International de Philosophie Scientifique*, Vol. 4., Hermann, Paris, pp. 18–23.

— (1939): "Foundations of Logic and Mathematics." *International Encyclopedia of Unified Science*, Vol.1, no. 3, University of Chicago Press, Chicago.

— (1949): "Truth and Confirmation." In FEIGL and SELLARS (1949), pp. 119-27. English version of CARNAP (1936).

— (1952): "Meaning Postulates." *Philosophical Studies* 3 (5). Reprinted in CARNAP (1956), pp. 222-229.

— (1955): "Meaning and Synonymy in Natural Languages." *Philosophical Studies* 6 (3). Reprinted in CARNAP (1956), pp. 233-247.

— (1956): *Meaning and Necessity*, University of Chicago, Chicago.

— (1963): "Intellectual autobiography:" In SCHILPP (1963), pp. 1-84.

CARWTRIGHT, R. (1962): "Propositions." In BUTLER (1962), pp. 81-103

CHIHARA, CH. (1979): "The Semantic Paradoxes: A Diagnostic Investigation," *Philosophical Review* (88): 590-618.

CHANG, C. & H. KEISLER (1990): *Model Theory* (3rd edn). North Holland.

CHILDERS, T., P. KOLÁŘ & V. SVOBODA eds. (1997): *Logica' 96:*

Proceedings of the 10th International Symposium. Filosofia, Prague.

CHURCH, A. (1956): *Introduction to Mathematical Logic*. Vol. I. Princeton University Press, Princeton.

COFFA, J. A. (1991): *The Semantic Tradition from Kant to Carnap*. Cambridge University Press, Cambridge.

DAVID, M. (1994): *Correspondence and Disquotation.* Oxford University Press, Oxford.

— (1996): "Analyticity, Carnap, Quine and Truth." *Philosophical Perspectives* (10): 281-296.

— (2004): "Theories of Truth." In NIINILUOTO et al (2004), pp. 331–413.

— (2008a): "Tarski's Convention T and the Concept of Truth." In PATTERSON (2008a), pp. 133-156.

— (2008b): "Quine's Ladder: Two and a Half Pages from the Philosophy of Logic." In FRENCH et al (2008), pp. 274-312.

DAVIES, M. (1981): *Meaning, Quantification and Necessity*. Routlege, London.

DAVIDSON, D. (1967): "Truth and meaning." *Synthese* (17): 304–23. Reprinted in DAVIDSON (1984), pp. 17–36.

— (1969): "True to the facts." Reprinted in DAVIDSON (1984), pp. 37- 54.

— (1973a): "Radical interpretation." *Dialectica* (27): 313–28. Reprinted in DAVIDSON (1984), pp. 125–140.

— (1973b): "In defence of convention T." Reprinted in DAVIDSON (1984), pp. 65-77.

— (1977): "Reality without reference." *Dialectica* (31): 247–53. Reprinted in DAVIDSON (1984), pp. 215–226.

—(1984): *Inquiries into the Truth and Interpretation*. Oxford University Press, Oxford.

— (1990): "The Structure and Content of Truth." *Journal of Philosophy* (87): 279–328.

— (2005a): *Truth and Predication*. The Belknap Press of Oxford University Press, Oxford.

— (2005b): *Truth, Language, and History*. Clarendon Press, Oxford.

DAWSON, J. W. (1989): "The reception of Gödel's incompleteness theorems." In SHANKER (1989), pp. 74–95.

DEMOPOULOS, W. (1994): "Frege, Hilbert, and the conceptual structure of model theory." *History and Philosophy of Logic* (15): 211–225.

DREBEN, B. & J. van HEIJENOORT (1986):"Introductory note to Gödel 1929, 1930, 1930a." In GÖDEL (1986), pp. 44–59.

DUMMETT, M. (1973). *Frege: The Philosophy of Language*. Duckworth, London.

— (1978a): *Truth and Other Enigmas*. Duckworth, London.

— (1978b): “Truth.” In DUMMETT (1978a), pp. 1-24.

EDWARDS, J. (2003): “Reduction and Tarski’s definition of logical consequence”. *Notre Dame Journal of Formal Logic* (44): 49–62.

ENDERTON, H. (2001): *A Mathematical Introduction to Logic* (2nd edn). Academic Press, San Diego.

ETCHEMENDY, J. (1988): “Tarski on Truth and Logical Consequence.” *Journal of Symbolic Logic* (53): 51–79.

— (1990) *The Concept of Logical Consequence*. Harvard University Press, Cambridge MA.

EWALD, W. (1996): *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*. Oxford University Press, NewYork.

FEFERMAN, A. & S. FEFERMAN (2004): *Alfred Tarski: Life and Logic*. Cambridge University Press, Cambridge.

FEFERMAN, S. (1991): “Reflecting on Incompleteness”, *Journal of Symbolic Logic*, 56: 1–49.

— (1998a): *In the Light of Logic*. Oxford University Press, New York and Oxford.

— (1998b): “Kurt Gödel: Conviction and Caution.” In FEFERMAN, S. (1998a): 150-164.

— (1999): “Tarski and Gödel: Between the Lines.’’ In WOLENSKI & KOHLER (1999), pp. 53-63.

— (2008a): “Tarski’s Conceptual Analysis of Semantical Notions.” In PATTERSON (2008a), pp. 72-93.

— (2008b): “Harmonious logic: Craig’s Interpolation Theorem and its Descendants.” *Synthese* (164):341–357.

FEIGL, H. & W. SELLARS, eds. (1949): *Readings in Philosophical Analysis*. Ridgeview Publishing Company, Atascadero CA.

FERNANDEZ MORENO, L. (1992): *Wahrheit und Korrespondenz bei Tarski. Eine Untersuchung der Wahrheitstheorie Tarskis als Korrespondenztheorie der Wahrheit.* Königshausen & Neumann.

— (1992): “Putnam, Tarski, Carnap und die Wahrheit.” *Gräzer philosophische Studien* (43): 33-44.

— (1997): “Truth in Pure Semantics: A Reply to Putnam.” *Sorites* (8): 15-23.

— (2001): “Tarskian Truth and the Correspondence Theory.” *Synthese* (126): 123-147.

FERREIROS, J. & J. GRAYAND, eds. (2006): *The Architecture of Modern Mathematics.* Oxford University Press, Oxford.

FIELD, H. (1972): “Tarski’s Theory of Truth.” *The Journal of Philosophy* (69): 347–75. Reprinted (with postscript) in FIELD (2001), pp. 3-29.

— (1986): “The Deflationary Conception of Truth.” In MacDONALD & WRIGHT (1986), pp. 55-117.

— (1994): “Deflationist Views of Meaning and Content.” *Mind* 103 (411): 249–84. Reprinted (with postscript) in ARMOUR-GARB & BEAL (2005), pp. 50-110. Page references are to the reprint.

— (1999): “Deflating the Conservativeness Argument.” *Journal of Philosophy* (96): 533–40.

— (2001): *Truth and the Absence of Fact*. Oxford University Press, Oxford.

— (2008): *Saving Truth from Paradox*. Oxford University Press, Oxford.

FLOYD, J. & S. SHIEH, eds. (2001): *Future Pasts: the Analytic Tradition in Twentieth-Century Philosophy*. Oxford University Press, Oxford.

FREGE, G. (1879): *Begriffschrift*. Nebert, Halle.

— (1884): *Die Grundlagen der Arithmetik. Eine logisch mathematische Untersuchung über den Begriff der Zahl*. Wilhelm Koebner, Breslau.

— (1893): *Grundgesetze der Arithmetik, begriffsschriftlich abgeleite.* Vol.1. Pohle, Jena.

— (1918/1919): “Der Gedanke. Eine logische Untersuchung.” *Beiträge zur Philosophie des deutschen Idealismus* (1): 58-77.

FRENCH, P. A., T. E. UEHLING, & H. K. WETTSTEIN, eds. (2008): *Midwest Studies in Philosophy* (32). Wiley-Blackwell.

FRIEDMAN, M. & R. CREATH, eds. (2007): *The Cambridge Companion to Carnap*. Cambridge University Press, Cambridge.

FROST-ARNOLD, G. (2004): “Was Tarski's Theory of Truth Motivated by Physicalism?” *History and Philosophy of Logic* (25): 265–80.

— (2006): *Carnap, Tarski, and Quine’s Year Together: Conversations on Logic, Science, and Mathematics.* Unpublished dissertation thesis. Available online at http://faculty.unlv.edu/frostarn/. To appear in *Full circle Series* of Open Court Press, La Salle.

GABBAY, D. & J. WOODS, eds. (2009): *Handbook of the History of Logic. Vol. V. Logic from Russell to Church*. Elsevier, Amsterdam.

GARCÍA-CARPINTERO, M. (1996): “What Is a Tarskian Definition of Truth?” *Philosophical Studies* (82): 113–44.

GENTZEN, G. (1935): “Die Widerspruchsfreiheit der reinen Zahlentheorie.” Published as “Der erste Widerspruchsfreiheitsbeweis für die klassische Zahlentheorie” in Archiv für mathematische Logik und Grundlagenforschung (16): 97–118. English translation in GENTZEN (1969), pp. 132–213.

— (1936): “Die Widerspruchsfreiheit der reinen Zahlentheorie. Mathematische Annalen (112): 493–565. English translation in GENTZEN (1969), pp. 132–213.

— (1969): *The Collected Papers of Gerhard Gentzen*. North-Holland, Amsterdam.

GÖDEL, K. (1930a): “Die Vollständigkeit der Axiome des logischen Funktionenkalküls.” *Monatshefte für Mathematik und Physik* (37): 349–

360. Reprinted and translated in GÖDEL (1986), pp. 102–123.

— (1931): "Über formal unentscheidbare Sätze der Principia Mathematica undverwandter Systeme I." *Monatshefte für Mathematik und Physik* (38): 173–198. Reprinted and translated in GÖDEL (1986), pp. 144–195. Page references are to the reprint.

— (1931?): "Über formal unentscheidbare Sätze." In Gödel (1995), pp. 30-35.

— (1934): "On Undecidable Propositions of Formal Mathematical Systems." Mimeographed lecture notes, taken by S.C. Kleene and J. B. Rosser in Princeton. Reprinted in GÖDEL (1986): 346–371.

— (1986): *Collected Works.* Vol. 1. Oxford University Press, Oxford.

— (1995): *Collected Works*. Vol. 3. Oxford University Press, Oxford.

— (2002a): *Collected Works*. Vol. 4. Oxford University Press, Oxford.

— (2002b): *Collected Works*. Vol. 4. Oxford University Press, Oxford.

GOLDFARB, W. D. (1979): "Logic in the twenties: the nature of the quantifier." *Journal of Symbolic Logic* (44): 351–368.

— (2001): "Frege's conception of logic." In FLOYD & SHIEH (2001), pp. 25–41.

GÓMES-TORRENTE, M. (1996): "Tarski on Logical Consequence." *Notre Dame Journal of Formal Logic* (37): 125–51.

GROVER, D., CAMP, J., & BELNAP, N., (1975): "A Prosentential Theory of Truth." *Philosophical Studies* (27): 73–125.

GUPTA, A. (1993): "A Critique of Deflationism." *Philosophical Topics* (21): 57–81.

HAAPARANTA, L. & J. HINTIKKA, eds. (1986): *Frege Synthesized*. Reidel, Dordrecht.

HAAPARANTA, L., ed. (2009): *The Development of Logic*. Oxford University Press, New York.

HALBACH, V. & L. HORSTEN, eds. (2002a): *Principles of Truth*. Dr. Hänsel-Hohenhausen, Frankfurt am Main.

HALBACH, V. & L. HORSTEN, eds. (2002b): "Introduction." In HALBACH, V. & L. HORSTEN, eds. (2002a), pp. 11-35.

HALBACH, V. (1999): "Disquotationalism and Infinite Conjunctions." *Mind* (108): 1–22.

— (2000): "How Innocent is Deflationism?" *Synthese* (126): 167–94.

— (2001): *Semantics and Deflationism*. Unpublished manuscript.

— (2009): "Axiomatic Theories of Truth." In *Stanford Encyclopedia of Philosophy*, ed. by E. Zalta, http://plato.stanford.edu/entries/truth-axiomatic/.

HECK, R. (1997): "Tarski, Truth and Semantics." *The Philosophical Review*

(106): 533-54.

— (2004): “Truth and Disquotation.” *Synthese* (142): 317–52.

— (2010): “Frege and semantics.” In POTTER & RICKETTS (2010), pp. 342 - 378.

HENKIN, L (1949): “The Completeness of the First-Order Functional Calculus.” *Journal of Symbolic Logic* 14 (3): 159-166.

— (1967): “Formal Systems and Their Models.” *The Encyclopedia of Philosophy*, Vol .VIII, MacMillian and Free Press, New York, pp. 61-74. Ed. by P. Edwards.

HENKIN, L. et al., eds. (1974): *Proceedings of the Tarski Symposium*. In: Proceedings of Symposia in Pure Mathematics, vol. XXV, RI: American Mathematical Society, Providence

HILBERT, D. (1899): *Grundlagen der Geometrie*. In *Festschrift zur Feier der Enthüllung des Gauss-Weber-Denkmals in Göttingen*, Teubner, 1st ed., Leipzig, pp. 1–92.

— (1900): “Mathematische Probleme.” *Nachrichten von der Königlichen Gesellschaf der Wissenschaften zu Göttingen*, Math.-Phys. Klasse 253–297. English translation in Ewald (1996), pp. 1096–1105. Page references are to the latter.

HILBERT, D. & W. ACKERMANN (1928): *Grundzüge der theoretischen Logik*. Springer, Berlin.

HILBERT, D. & BERNAYS, P. (1939): *Grundlagen der Mathematik* Vol. 1. Springer, Berlin.

— (1939): *Grundlagen der Mathematik* Vol. 2. Springer, Berlin.

HINTIKKA, J. (1988): “On the development of the model-theoretic viewpoint in logical theory.” *Synthese* (77): 1–36.

— (1996a): *Lingua Universalis vs. Calculus Ratiocinator: An ultimate presupposition of twentieth-century philosophy*. Kluwer, Dordrecht.

— (1996b): *The Principles of Mathematics Revisited*. Cambridge University Press, New York.

HODGES, W. (1985/86): “Truth in a Structure:” *Proceedings of the Aristotelian Society*, new series (86): 131–51.

— (1997): *A Shorter Model Theory*. Cambridge University Press, Cambridge.

— (20??): “Model Theory.” Draft. Available on-line at http://wilfridhodges.co.uk/.

— (2008): “Tarski’s Theory of Definition.” In PATTERSON (2008a), pp. 94-132.

HORWICH, P. (1982): “Three Forms of Realism.” *Synthese* (51): 181-201

— (1990): *Truth*. Blackwell, Cambridge.

— (1998): *Truth* (2nd rev. edn). Blackwell, Oxford.

— (2005): “A Minimalist Critique of Tarski on Truth.” In BEAL (2005), pp. 75-84.

HUGHES, R.I.G. (1993): *A Philosophical Companion to First-order Logic*. Hackett Publishing Company, Indianapolis.

HUNTINGTON, E. V. (1904): “Sets of independent postulates for the algebra of logic.” *Transactions of the American Mathematical Society* (5): 288–309.

— (1905): “A set of postulates for real algebra, comprising postulates for a one dimensional continuum and for the theory of groups.” *Transactions of the American Mathematical Society* (6): 17–41.

JACQUETTE, D., ed. (2007): *Handbook of the Philosophy of Science. Volume 5: Philosophy of Logic*, Elsevier BV, Netherlands, pp. 273–369. Handbook Ed. by Dov Gabbay, Paul Thagard and John Woods.

JANÉ, I. (2006): “What Is Tarski's Common Concept of Consequence?” *Bulletin of Symbolic Logic* (12): 1–42.

KAPLAN, D. (1989): “Demonstratives.” In ALMONG et al (1989), pp. 565-614.

KEMENY, J. (1948): “Models of Logical Systems.” *Journal of Symbolic Logic* 13 (1):16-30.

KLEENE, S.C. (1939): “Review: Rudolf Carnap, *The Logical Syntax of Language*.” 4(2): 82-87.

(1953): *Introduction to Metamathematics*. Van Nostrand Reinhold, Amsterdam and New York.

— (1976): “The Work of Kurt Gödel.” *Journal of Symbolic Logic* (41): 761-778.

— (1986): “Introductory notes.” In GÖDEL (1986), pp. 126-139.

— (1987): “Kurt Gödel.” *Biographical Memoirs*, vol. 56, National Academy of Sciences of the United States of America, National Academy Press, Washington, D.C, pp. 134-179.

KOREN, L. (2010a): “Tarski’s Method of Truth Definition: its Nature and Significance.” In *Miscellanea Logica VIII: Foundations of Logic*, (ed.) J. Peregrin, ACTA UNIVERSITATIS CAROLINAE PHILOSOPHICA ET HISTORICA 2/2007, Karolinum, Prague, pp. 71-112.

(2010b): “In What Sense is Tarski’s Semantic Conception of Truth Semantic?” In *An Anthology of Philosophical Studies*, (ed.) P. Hanna, Atiner, Athens, pp. 153 -166.

KOTARBINSKI, T. (1929): *Elementy teorii poznania, logiki formalnej i metodologii nauk*. Ossolineum, Lwów. English translation *Gnosiology. The Scientific Approach to the Theory of Knowledge*, Pergamon Press, Oxford, 1960.

KOTLARSKI, H., S. KRAJEWSKI & A. LACHLAN(1981): “Construction of Satisfaction Classes for Nonstandard Models.” *Canadian Mathematical Bulletin* (24): 283–293.

KRIPKE, S. (1963): “Semantical Considerations on Modal Logic.” *Acta Philosophica Fennica* (16): 83-94.

— (1975): “Outline of a Theory of Truth.” *Journal of Philosophy* 72 (19): 690-716.

KÜNNE, W. (2003): *Conceptions of Truth*. Oxford University Press, Oxford.

LARSON, R.K. & G. SEGAL (1995): *Knowledge of Meaning*. MIT Press, Cambridge MA.

LEEDS, S. (1978): “Theories of Reference and Truth.” *Erkentniss* (13): 111-130. Reprinted in ARMOUR-GARB & BEAL (2005), pp. 33-49. Page references are to the reprint.

LePORE, E. (1983): “What model theoretic semantics cannot do?” *Synthese* 54 (2): 167-187.

LePORE, E. & K. LUDWIG (2005): *Donald Davidson: Meaning, Truth, Language and Reality.* Oxford University Press, Oxford.

LESNIEWSKI, S (1929): “Grundzüge eines neuen Systems der Grundlagen der Mathematik.” *Fundamenta Mathematicae* (14): 1–81.

LEWIS, D (1969): *Convention*. Harvard University Press, Cambridge MA.

— (1970): “General semantics.” *Synthese* (22): 18–67. Reprinted in LEWIS (1983), pp. 189–229.

— (1974): “Radical interpretation.” *Synthese* (23): 331–44. Reprinted in LEWIS (1983), pp. 108–18.

— (1975): “Language and languages.” In *Minnesota Studies in the Philosophy of Science*, volume VII, University of Minnesota Press, pp. 3–35. Reprinted in LEWIS (1983), pp. 163-88.

— (1983): *Philosophical Papers* I. Oxford University Press, Oxford.

LEWY, C. (1947): “Truth and Significance.” *Analysis* (8): 24-27.

LÖWENHEIM, L. (1915): “Über Mölichkeiten im Relativkalkül.” *Mathematische Annalen* 447–470. Translated in van HEIJENOORT (1967a), pp. 228–251.

LYNCH, M. P., ed. (2001): *The Nature of Truth*. MIT Press, Cambridge MA.

MacDONALD, C & C. Wright, eds. (1986): *Fact, Science and Morality*, Blackwell, Oxford.

MacLANE, S. (1938): “Carnap on logical syntax.” *Bulletin of the American Mathematical Society* (44): 171-176.

MANCOSU, P. (2006): “Tarski on models and logical consequence.” In FERREIROS (2006), pp. 209-37.

— (2008): “Tarski, Neurath, and Kokoszýnska on the Semantic Conception of Truth.” In PATTERSON (2008a), pp. 192-224.

MANCOSU, P., C. BADESA & R. ZACH (2009): “The Development of Mathematical Logic from Russell toTarski, 1900–1935.” In HAAPARANTA (2009), pp. 318-470.

McCALL, S., ed. (1967): *Polish Logic1920-1939.* Oxford University Press, Oxford.

McDOWELL, J. (1978): “Physicalism and primitive denotation: Field on Tarski.” *Erkenntnis* (13): 131–52.

McGEE, V. (1991): *Truth, Vagueness, and Paradox*. Hackett, Indianapolis.

— (1993): “A Semantic Conception of Truth?” *Philosophical Topics* (21): 83–111. Reprinted (with the postscript) in ARMOUR-GARB & BEALL (2005), pp. 111-152. Page references are to the reprint.

— (2005): “Maximal Consistent Sets of Instances of Tarski’s Schema (T).” *Journal of Philosophical Logic* (21): 235–41.

MENDELLSON, E. (1997): *Introduction to Mathematical Logic* (4th edn). Chapman and Hall, Boca Raton, Florida.

MILNE, P. (1997): “Tarski on Truth and Its Definition.” In CHILDERS et al (1997), pp. 189–210.

— (1999): “Tarski, Truth and Model Theory.” *Proceedings of the Aristotelian Society* (99): 141–67.

MONTAGUE, R. (1974): *Formal Philosophy. Selected Papers of Richard Montague*. Yale University Press, New Haven. Ed. by R. Thomason.

MOORE, G. E. (1953): *Some Main Problems of Philosophy*. George Allen and Unwin London.

MURAWSKI, R. (1998): “Undefinability of truth. The problem of priority: Tarski vs Gödel.” *History and Philosophy of Logic* (19): 153–160.

MURAWSKI. R. & WOLENSKI, J. (2008): "Tarski and his Polish Predecessors on Truth.” In PATTERSON (2008a), pp. 21-43.

NEURATH, O. (1983a): *Philosophical Papers (1913-1946)*. D. Reidel, Dordrecht.

— (1983b): “Physicalism.” In NEURATH (1983a), pp. 52-57.

NIINILUOUTO, I., M. SINTONEN & J. WOLLÉNSKI, eds. (2004): *The Handbook of Epistemology*. Kluwer, Dordrecht.

LYNCH, M. P. ed. (2001): *The Nature of Truth*. MIT Press, Cambridge MA.

PADOA, A. (1901): “Essai d’une théorie algébrique des nombres entiers, précédé d’une introduction logique à une théorie déductive quelconque.” In *Bibliothèque du Congrès international de philosophie*, Paris, 1900, Vol. III, Armand Colin, Paris, pp. 309–324. Partial English translation published in van HEIJENOORT (1967a), pp. 118–23.

PAP, A. (1954): “Propositions, Sentences, and the Semantic Definition of Truth.” *Theoria* (XX): 23–35.

PARSONS, CH (1974): “Informal Axiomatization, Formalization and the Concept of Truth.” *Synthese* (27): 27-47.

PARTEE, B (1977): “Possible World Semantics and Linguistic Theory.” *Monist* (60): 303-26.

PATTERSON, D., ed. (2008): *New Essays on Tarski*. Oxford University Press.

— (2008a). “Tarski’s Conception of Meaning.” In PATTERSON (2008a), pp.

157-191.

PEANO, G. (1889): *I Principii di Geometria Logicamente Esposti*. Fratelli Bocca, Torino.

PEREGRIN, J., ed. (1999): *Truth and its Nature (if any)*. Kluwer, Dordrecht.

— (2006): "Consequence and Inference." In *Miscellanea Logica VI: From Truth to Proof*, (ed.) V. Kolman, 2006, pp. 1-18.

PLATO (1997): *Plato Complete Works*. Hackett Publishing Co. Ed. by J. M. Cooper.

POPPER, K. (1972a): *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford.

— (1972b): "Philosophical Comments on Tarski's Theory of Truth." In POPPER (1972a), pp. 319-40.

POTTER, M. & RICKETTS, T., eds. (2010): *The Cambridge Companion to Frege*. Cambridge University Press, Cambridge.

PROCHÁZKA, K. (2006): "Consequence and Semantics in Carnap's System." In *Miscellanea Logica* VI: *From Truth to Proof*, (ed.) V. Kolman, 2006, pp. 79-113.

— (2010): *Truth between Syntax and Semantics*. Unpublished dissertation thesis (Charles University in Prague.)

PUTNAM, H. (1985): *Representation and Reality*. MIT Press, Cambridge.

— (1994a): *Words and Life*. Harvard University Press, Harvard.

— (1994b): "On Truth." In PUTNAM (1994a), pp. 315–29

— (1994c): "Comparison of Something with Something Else." In PUTNAM (1994a), pp. 330–50.

QUINE, W. V. O. (1953a): *From a Logical Point of View*. Harvard University Press, Cambridge MA.

— (1953b): "Notes on Theory of Reference." In QUINE (1953a), pp. 130-138.

— (1953c): "Two Dogmas of Empiricism." In QUINE (1953a), pp. 20-46.

— (1960): *Word and Object*. Cambridge University Press, Cambridge.

— (1970): *Philosophy of Logic*. Harvard University Press, Cambridge MA.

— (1987): *Quiddities: An Intermittently Philosophical Dictionary*. Cambridge, Mass. Harvard University Press.

— (1992): *Pursuit of Truth* (2nd rev. edn). Harvard University Press, Cambridge, MA.

RAMSEY, F. P. (1927): "Facts and Propositions." *Proceedings of the Aristotelian Society* (7) (Supplementary): 153–170. Reprinted in RAMSEY (1990), pp. 34-51. Page references are to the reprint.

— (1990): *Philosophical Papers*. Cambridge University Press, New York. Ed. by D. H. Mellor.

— (2001): “The Nature of Truth.” Reprinted in LYNCH (2001), pp. 433-445. Page references are to the reprint.

REICHENBACH, H. (1938): *Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge*. University of Chicago Press, Chicago.

RICKETTS, T. (1986): “Objectivity and objecthood: Frege’s metaphysics of judgement.” In HAAPARANTA & HINTIKKA (1986), pp. 65–95.

— (1996): “Logic and truth in Frege.” *Proceedings of the Aristotelian Society*, sup. vol. (70): 121–140.

ROBINSON, R. (1952): “An essentially undecidable axiom system.” In *Proceedings of the International Congress of Mathematicians, Cambridge Mass. 1950, Vol.1*, Providence, R.I., pp. 729–730.

RODRIGEZ-CONSUEGRA (2005): “Tarski on Sets.” In SICA (2005), pp. 228-269.

ROSSER, J. B. (1936): “Extensions of some theorems of Gödel and Church.” *Journal of Symbolic Logic* (1): 87–91.

de ROUILHAN, P. & S. Bozon (2006): “The Truth of IF: has Hintikka Really Exorcised Tarski’s Curse.” In AUXIER & HAHN (2006), pp. 683–705.

de ROUILHAN, P. (2009): “Carnap on logical consequence for Languages I and II.” In WAGNER (2009), pp. 121- 146.

RUSSELL, B. (1908): “Mathematical Logic as Based on the Theory of Types.” *American Journal of Mathematics* (30): 222–262. Reprinted in van HEIJENOORT (1967a), pp. 150–182.

SAGÜILLO, J. M. (1997): “Logical Consequence Revisited,” *The Bulletin of Symbolic Logic* (3): 216–241.

SCHILPP, P. A., ed. (1963): *The Philosophy of Rudolf Carnap*. *The Library of Living Philosophers.* Vol. XI. Open Court, LaSalle, Illinois.

SELLARS, W. (1962): “Truth and ‘Correspondence’.” *Joural of Philosophy* LXI (2): 29-56.

SHANKER, S. G., ed. (1989): *Gödel’s Theorem in Focus*. Routledge, London.

SHAPIRO, S (1998): “Truth and Proof: Through Thick and Thin.” *Journal of Philosophy* (10): 493–521.

— (2002): “Deflation and Conservation”, In HALBACH & HORSTEN (2000a), pp. 103-128

SHER, G. (1999): “What Is Tarski's Theory of Truth?” *Topoi* (59): 149–66.

SICA, G. ed. (2005): *Essays on the Foundations of Mathematics and Logic*. Polimetrica International Scientific Publisher, Monza, Italy.

SIEG, W. (2009): “Hilbert's Proof Theory.“ In GABBAY & WOODS (2009), pp. 321-384.

SIMMONS, K. (2008): Tarski’s Logic.” In GABBAY & WOODS (2009), pp. 511-616.

SINACEUR, H. (2001): “Alfred Tarski: Semantic Shift, Heuristic Shift in Metamathematics. *Synthese* (126): 49–65.

SKOLEM, T. (1920): “Logisch-kombinatorische Untersuchungen über die Erfüllbarkeit oder Beweisbarkeit mathematischer Sätze nebst einem Theoreme über dichte Mengen. *Videnskasselskapets skrifter, I. Matematisk-naturvidenskabelig klasse 4*. Partial English translation published in van HEIJENOORT (1967a), pp. 252–263.

— (1922): “Einige Bemerkungen zur axiomatischen Begründung der Mengenlehre. In *Matematiker kongressen I Helsingfors*, Akademiska Bokhandeln, Helsinki, pp. 217–232. English translation published in van HEIJENOORT (1967a), pp. 290–301.

(1933): “Über die Unmöglichkeit einer vollständigen Charakterieserung der Zahlenreihe mittels eines endlichen Axiomen-systems.” *Norsk Matematisk Forenings Skrifter* 2 (10): 73–82.

SMITH, P. (2007): *An Introduction to Gödel’s theorems*. Cambridge University Press, New York.

SMORÝNSKI, C. (1977): “The incompleteness theorems.” In BARWISE (1977), pp. 821–865.

SOAMES, S. (1984): “What is a Theory of Truth?” *Journal of Philosophy* (81): 411–29.

— (1999): *Understanding Truth*. Oxford University Press, Oxford.

STRAWSON, P. F. (1949): “Truth.” *Analysis* 9 (6): 83-97.

STANLEY, J. (1996): “Truth and metatheory in Frege.” *Pacific Philosophical Quarterly* (77): 45–70.

TAPPENDEN, J. (1997): “Metatheory and mathematical practice in Frege.” *Philosophical Topics* (25): 213–64.

TARSKI, A. (1930): “Fundamentale Begriffe der Methodologie der deduktiven Wissenschaften. I.” *Monatshefte für Mathematik und Physik* (37): 361–404. English translation by J. H. Woodger in TARSKI (1983), pp. 60-109.

— (1931): “Sur les ensembles définissables de nombres réels. I. *Fundamenta Mathematicae* (17): 210–239. English translation by J. H. Woodger published in TARSKI (1983), pp. 110–142. Page references are to the latter.

— (1933a): *Pojęcie prawdy w językach nauk dedukcyjnych*. Nakładem Towarzystwa Naukowego Warszawskiego, Warszawa.

— (1933b): “Einige Betrachtungen über die Begriffe ω-Widerspruchsfreiheit und der ω-Vollständigkeit.“ *Monatshefte für Mathematik und Physik* (40): pp. 97–112. English translation by J. H. Woodger in TARSKI (1983), pp. 279–295.

— (1934–35): “Einige methodologische Untersuchungen über die Definierbarkeit der Begriffe.” *Erkenntnis* (5): 80–100. English

translation by J. H. Woodger published in TARSKI (1983), pp. 296–319.

— (1935): “Der Wahrheitsbegriff in den formalisierten Sprachen.” *Studia Philosophica* (1): 261–405. Enlarged and revised version of TARSKI (1933a). English translation by J. H. Woodger published in TARSKI (1983), pp. 152–278. Page references are to the translation.

— (1936a): “O pojeciu wynikania logicznego.” *Przeglad Filozoczny* (39): 58-68.German version “Über den Begriff der logischen Folgerung” in *Actes du Congrès International de Philosophie Scientifique* 7, Actualités Scientifiques et Industrielles, Hermann et Cie, Paris, pp. 1–11. English translation by J. H. Woodger published in TARSKI (1983), pp. 409–20. Page references are to the translation.

— (1936b): “O ungruntowaniu naukowej semantyki. *Przeglad Filozoczny* (39): 50-57. German version “Grundlagen der Wissenschaftlichen Semantik” in *Actes du Congres Internationale de Philosophie Scientifiques*, vol. 3, Hermann and Cie, Paris. English translation by J. H. Woodger published in TARSKI (1983), pp. 409–20. Page references are to the translation.

— (1937): *Einführung in die mathematische Logik und in die Methodologie der Mathematik*. Springer, Vienna.

— (1941): *Introduction to Logic and to the Methodology of Deductive Science*. Oxford University Press, New York. English translation, with additions, of TARSKI (1937).

— (1944): “The Semantic Conception of Truth and the Foundations of Semantics.” *Philosophy and Phenomenological Research* (4): 341–376.

— (1948a): “A Problem Concerning the Notion of Definability.” *Journal of Symbolic Logic* 13(2): 107-111

— (1948b): *A decision method for elementary algebra and geometry.* (prepared with the assistance of J. C. C. McKinsey), RAND Corp. (Santa Monica); University of California Press, Berkeley.

— (1954): “Contributions to the theory of models I.“ *Indagationes Mathematicae* (16): 572–81.

— (1969): “Truth and Proof.” *Scientific American* (220): 63–77. Reprinted in HUGHES (1993), pp. 101-125. Page references are to the reprint.

— (1983): *Logic, Semantics, Metamathematics* (2nd rev. edn). Hackett, Indianapolis. Ed. by J. Corcoran.

— (1986): “What Are Logical Notions?” *History and Philosophy of Logic* (7): 143–54. Ed. by J. Corcoran.

TARSKI, A. MOSTOWSKI & R. ROBINSON (1953): *Undecidable Theories.*

North-Holland Publishing Co, Amsterdam.

TARSKI, A. & R. L. VAUGHT (1956): “Arithmetical extensions of relational systems.” *Compositio Mathematica* (13): 81–102.

TENNANT, N. (2002): “Deflationism and the Gödel-Phenomena”, *Mind* (111): 551-582.

THOMASON, R. (1974): “Introduction.” In MONTAGUE (1974), pp. 1–69.

Van HEIJENOORT, J., ed. (1967a): *From Frege to Gödel. A Source Book in Mathematical Logic*, 1897–1931. Harvard University Press. Cambridge MA.

— (1967b): “Logic as calculus and logic as language.” *Boston Studies in the Philosophy of Science* (3): 440–446.

VAUGHT, R. L. (1974): “Model theory before 1945.” In HENKIN et al (1974), pp. 153–172.

Von NEUMANN, J. (1925): “Eine Axiomatisierung der Mengenlehre.” *Journal für die reine und angewandte Mathematik* (154): 219–240. English translation in van HEIJENOORT (1967a), pp. 393-413.

— (1927): “Zur Hilbertschen Beweistheorie.” *Mathematische Zeitschrift* (26): 1–46.

— (1966): *Theory of Self-Repriducing Automata*. University of Illinois Press, Urbana. Ed. by E. W. Burks.

WAGNER, P., ed. (2009): *Carnap’s Logical Syntax of Language*. Palgrave Macmillan, Basingstoke, Hampshire.

WANG, H. (1974): *From Mathematics to Philosophy*. Routledge and Kegan Paul, London.

— (1981): “Some Facts about K.Gödel.” *Journal of Symbolic Logic* (46): 653–659.

— (1987): *Reflections on Kurt Gödel*. MIT Press, Cambridge.

— (1996): *A Logical Journey: Form Gödel to Philosphy*. MIT Press, Cambridge MA.

WEYL, H. (1910): “Über die Definitionen der mathematischen Grundbegriffe.” *Mathematisch-naturwissenschaftliche Blätter* (7): 93–95, 109–113.

— (1949): *Philosophy of Mathematics and natural Science*. Princeton University Press, Princeton.

WHITEHEAD, A. N. & B. RUSSELL (1910-13): *Principia Mathematica* (Vol. 1-3). Cambridge University Press, Cambridge.

WILLIAMS, M., (1999): “Meaning and Deflationary truth.” *Journal of Philosophy* (96): 545–64.

WITTGENSTEIN, L. (1922): *Tractatus Logico-Philosophicus*. Routledge, London.

WOLÉNSKI, J. & E. KÖHLER (1998): *Alfred Tarski and the Vienna Circle*. Kluwer, Dordrecht.

WOLENSKI, J. (2005): “Gödel, Tarski and Truth.” *Revue international de philosophie* (4): 459-490.

ZACH, R. (2003): “Hilbert’s Program.” In *Stanford Encyclopedia of Philosophy*,

ed. By E. Zalta, http://plato.stanford.edu/entries/hilbert-program/.

— (2006): “Hilbert’s Program Then and Now.” In JACQUETTE (2006), pp. 411-447.

ZERMELO, E. (1908): “Untersuchungen über die Grundlagen der Mengenlehre I.” *Mathematische Annalen* (65): 261–281. English translation published in van HEIJENOORT (1967a), pp. 199-215.