An Observation about Truth

(with Implications for Meaning and Language)

Thesis submitted for the degree of

"Doctor of Philosophy"

By

David Kashtan

Submitted to the Senate of the Hebrew University of Jerusalem

November 2017

This work was carried out under the supervision of:

Prof. Carl Posy

Racheli Kasztan-Czerwonogórze

i Avigail Kasztan-Czerwonogórze

i ich matce

Acknowledgements

I am the proximal cause of this dissertation. It has many distal causes, of which I will give only a partial list.

During the several years in which this work was in preparation I was lucky to be supported by several sources. For almost the whole duration of my doctoral studies I was a funded doctoral fellow at the Language, Logic and Cognition Center. The membership in the LLCC has been of tremendous significance to my scientific education and to the final shape and content of this dissertation. I thank especially Danny Fox, from whom I learnt how linguistics is done (though, due to my stubbornness, not how to do it) and all the other members of this important place, past and present, and in particular Lital Myers who makes it all come together. In the years 2011-2013 I was a funded participant in an inter-university research and study program about Kantian philosophy. The program was organized by Ido Geiger and Yakir Levin from Ben-Gurion University, and by Eli Friedlander, Ofra Rechter and Yaron Senderowicz from Tel-Aviv University. The atmosphere in the program was one of intimacy and devotion to philosophy; I predict we will soon witness blossoms, the seeds of which were planted there.

In the winter of 2012 I visited the Munich Center for Mathematical Philosophy for a one-month funded doctoral fellowship. I thank Hannes Leitgeb and the fellows present there for a warm and enlightening stay. The spring semester of 2016 I spent in Germany again, this time around Berlin, and was hosted as a visiting fellow at the Zentrum für allgemeine Sprachwissenschaft (ZAS). An important part of the dissertation was written in that time. I thank Uli Sauerland for having me, Marie-Christine Meyer for making the connection, and the other fellows in ZAS for an interesting and enjoyable visit.

The Hebrew University has a reputation for being a cold and impersonal place, whatever its academic merits. But it is not cold, just misunderstood. Apart from having been taught, I have been helped, encouraged, inspired and, most importantly, educated, by many different people in the philosophy department (and outside of it). I thank David Enoch, Michael Roubach, Oron Shagrir, Nali Thaler, and all the other members of the department, and especially Limor Eilon, without whom the department would not be the unified organism that it is. I had the privilege to be awarded the departmental prize in memory of Yael Cohen. I thank the family for keeping up this institution throughout all these years. It is an important event in the yearly cycle of the department.

This dissertation developed out of my MA thesis, entitled Kant, Tarski and Quine: The Logical Form of Transcendental Philosophy, written (in Hebrew) under the supervision of Yemima Ben-Menahem. I thank Yemima for her guidance, tolerance and support in the face of my stubbornness. The dissertation also owes a lot to its two committee members, Gila Sher of the University of San Diego, and Eli Dresner of Tel-Aviv University. It was in listening to a mini-course Gila gave for the LLCC that many of my

ideas about Tarskian truth started to crystallize. Eli's presence on the committee had an important anti-stress effect, and his philosophical modesty and tolerance are virtues that I hope in the future to succeed better in adopting.

I will only be able to give a partial account of my debt to Carl Posy, the advisor of the present thesis. Carl has not been a particularly easy supervisor, which is the first thing I am grateful for. Philosophy is (or should be conceived) first and foremost (as) a profession. This does not, as some people might worry, diminish its loftiness. It just implies that there is a certain standard to which a piece of philosophical argumentation or exposition must adhere. Learning to do philosophy is learning to identify, in particular in one's own work, when philosophical content is present and when, on the contrary, philosophical ideas are at most grazed, maybe caressed, but not really handled. Prior to education, one may possess philosophical vision, sensitivity, even genius; but the professional standard is acquired knowledge. This dissertation, I hope, is the fruit of the first step in my acquisition of this standard, and I thank Carl for his central role in this step.

The growth of philosophers depends on the soil in which they are planted, and the peers of a graduate student are the constituents of that soil. Since I have spent so many years on this project, the list of peers who own a piece of it is long. A special thanks goes to Gil Sagi, now of Haifa University. Gil's energy and initiative in organizing reading groups, conferences, workshops and invited talks, during the years of her doctoral studies as well as after they were over, have both been a benefit and an inspiration. Other friends and colleagues from whom I've learnt and with whom time passed quicker than it would have otherwise are Ron Aboodi, Roy Amir, Dustin Atlas, Dan Baras, Moysh Bar-Lev, Itai Bassi, Henry Brice, Nora and Will Danielson-Lanier, Eran Fish, Moran Godess-Riccitelli, Rea Golan, Aviv Keren, Nati Kupfer, Ran Lanzet, Daniel Margulis, Danny November, Etye Steinberg, Shlomit Wygoda, and Ynon Wygoda. I apologize to the people I've forgotten; writing a philosophy dissertation decreases your cognitive competence considerably.

This dissertation took a long time to write. During this time, I got married three times (to the same woman, it's a long story) and had two daughters (two different ones). A PhD dissertation, especially one that seems never to end, puts a great strain on the writer's surroundings. My surroundings have been supporting beyond any reasonable expectation, a fact it has been easy to take for granted. I thank my parents for incredible amounts of patience and tolerance, and for similar amounts of technical, financial and especially moral support, and apologize for having taken them so often for granted. At some age it is no longer excusable to take your parents for granted, and I am well past that age.

Undoubtedly the star, as far as this dissertation and the life of its composer are concerned, is my wife Katarzyna Czerwonogóra. The final phase of writing coincided with the birth of our second daughter, and the strain on our household was heavy. Kasia was patient and accommodating to the point of heroism, and we are both very glad that this period is behind us. The biggest thanks goes to her. I

couldn’t have wished myself a better partner and teacher in the spiritual, intellectual, ethical and emotional journeys that make up an adult life.

Lastly, I would like to thank the first cause of this essay.

Abstract

This dissertation is a philosophical analysis of the concept of truth. It is a development and defense of the "stratified" or "language-level" conception of truth, first advanced in Alfred Tarski's 1933 monograph The Concept of Truth in Formalized Languages. Although Tarski's paper had seminal influence both in philosophy and in more technical disciplines, its central philosophical claim has not been generally accepted. This work has two central goals: (a) to give a detailed and analytic presentation of Tarski's theory and the problems it faces; (b) to offer a solution to these problems and assess the philosophical significance of this solution.

The essay is divided in two parts. Part One contains a detailed and analytic presentation and interpretation of the stratified conception of truth. The analysis contains several steps: (a) Crucial basic assumptions, such as the limitation to formalized languages and the requirement of explicit definitions, are stated explicitly, motivated, and their philosophical significance discussed. (b) The main negative result of the stratified conception, the impossibility of semantic closure and of a universal language, is given in detail and interpreted. (c) Tarski's criterion for adequate truth definitions, known as Convention T, is stated and motivated. (d) The deep structure of Tarski-style truth definitions and the necessary conditions for their availability are analyzed. In particular, the philosophical significance of Tarski's notion of "essential richness" is discussed. (e) Finally, several problems are raised for the stratified conception, chief among them *the unity objection*, according to which the stratified conception is not a viable analysis of the concept of truth, since (by (a) above) an analysis should take the form of a definition, and on the stratified conception different languages have different definitions. There is therefore no one analysis of the concept.

Part Two is a development of answers to the problems raised at the end of Part One. The crux of the answer to the unity objection is that Convention T, the adequacy criterion, connects the many definitions of truth into a single concept. However, in order to fulfill that role Convention T must apply universally, and a universal language was shown to be impossible ((c) above). The task of Part Two is therefore to develop a mode of expression that allows the universal applicability of Convention T without commitment to a universal metalanguage. The procedure is as follows. (a) Convention T is formalized in order to isolate the place in which universal applicability is required. (b) A new expressive resource of "abstract generality" is developed. To this purpose a digression into the semantics of natural language indexicals is undertaken. David Kaplan's thesis of the direct reference of indexicals is analyzed and a new formal system is proposed that embodies it. It is shown that this formal system expresses abstract generality. (c) The notion of abstract generality is adapted to languages without indexicals and it is

shown that Convention T can be expressed without assuming a universal language. (d) A reconstrual of the task of concept analysis is proposed, which is a generalization of the answer to the unity objection.

It is often complained against Tarski's stratified conception of truth that it is of limited philosophical significance. In this work I show that, on the contrary, the problems it faces and the solutions that can be advanced to answer these problems have substantive philosophical consequences. The notion of abstract generality gives rise to a distinction between two fundamentally different modes of discourse: a universal but merely abstract *methodological* discourse on the one hand, and a concrete but inevitably restricted *theoretical* discourse on the other. This distinction has many important implications for our understanding of the concepts of truth, meaning and language.

Contents

Introduction

This essay is a defense of what I shall call *the stratified conception of truth*. The basic idea of the stratified conception is that truth, as a concept that applies to judgments, applies to them not absolutely but as belonging to a system of representation (more concretely, to sentences in a language); and that the truth concept for judgments of a particular system cannot be represented in that very system, but only in a different one which is in some sense stronger. The concept of truth thus induces a stratification of our representational capacities: there is no *semantically closed* representational system.

The source of the stratified conception is Alfred Tarski's seminal monograph *The Concept of Truth in Formalized Languages* (henceforth *CTFL*) from 1931.[1] Tarski's analysis of truth in *CTFL* was highly influential both philosophically and technically. Today, however, it is generally rejected. Philosophically its immediate effect was to dispel the skepticism concerning the viability of semantic concepts such as truth and denotation. This skepticism was especially prevalent among the philosophers of the Vienna Circle, stemming to a certain extent from the influence on them of Wittgenstein's *Tractatus*. Tarski's *CTFL* is sometimes described as contributing to the liberation, especially of Carnap, from the *Tractatus*' inhibiting influence.[2] Technically, *CTFL* and the developments that followed it served as the basis for the by now well-established fields of model theory in mathematical logic and formal semantics in linguistics. Model theory was established in its present form in the 1950s through further work of Tarski and associates.[3] Modern linguistic formal semantics is a complex framework with roots especially in the works of Davidson and Tarski's student Montague, from the 1960s.[4]

In the last half-century the stratified conception has become somewhat of an abused orthodoxy, an obviously inadequate foil against which to develop better theories. The main reason it came to be rejected, at least as a philosophical account of truth, is that it applies to partial, formalized, languages rather than to ordinary language, and therefore that it is not an analysis of "our ordinary concept" of truth. For ordinary language is semantically closed and unstratified, a fact we learn, for example, from the intuitive acceptability of unstratified truth ascriptions. The analysis of ordinary language and of our

[1] The monograph was first published in Polish in (1933). A translation into German by Leopold Blaustein, authorized by Tarski, was published in (1935). An English translation by Joseph Woodger appear in the (1956) collection of early papers by Tarski. Page numbers for *CTFL* will refer to (1956). Another important source is Tarski's (1944) originally English non-technical summary directed at philosophers.

[2] See, e.g., Popper (1979, p.319), Mancosu (2008), Tarski (1992) (quoted in Patterson (2012, p.216f)), Awodey and Carus (2007).

[3] Though it embodies a substantial modification of the conception of truth of *CTFL*, see Etchemendy (1988, §2), Hodges (1984).

[4] See Glanzberg (2014, §1) for an account of the way in which modern semantics combines Davidson's and Montague's approaches.

ordinary concept of truth has, since the 1970s, been taken to be the central philosophical task of semantics. And the first thesis of *CTFL* was that the concept of truth does not apply to ordinary language. In this essay I will largely ignore this objection from ordinary language. My intention is to defend the stratified conception on its own terms. In a coda to the essay I offer some preliminary considerations against the objection.[5]

This dissertation has two parts. Part One contains a detailed and analytic presentation and interpretation of the stratified conception of truth, more or less as we find it in Tarski's *CTFL*. Chapter 1 states some preliminary assumptions regarding the form that the inquiry will take. In particular, it is set down that an analysis of the concept of truth should take the form of an explicit definition in a regimented language. Sentences in regimented languages are also taken to be the truth-bearers, and a detailed presentation of the regimentation scheme is provided. Chapter 2 presents the naïve non-stratified conception of truth, according to which truth can and should be defined for a single universal language. A refutation of the naïve conception is given along the lines of §1 of *CTFL*. Chapter 3 has a detailed exposition of the philosophical core of Tarski's stratified conception of truth. Since there is no universal language, truth has to be defined anew for every partial language. A criterion is provided by which the success of a definition can be decided, Tarski's famous Convention T. The possibility of a successful definition is demonstrated and conclusions about the conditions for such a definition are drawn. The main condition is that the language in which truth is defined (the metalanguage) should be distinct and expressively richer, in a sense made precise, than the language for which truth is defined (the object-language). Finally, some problems for the stratified conception are raised, in particular the fact that the stratified conception doesn't fulfill the task as it was set up in Chapter 1, that of giving a definition of the concept of truth. This is because on the stratified conception each language has a different truth definition, and the plurality of languages cannot be avoided. This is called *the unity objection*.

Part Two develops answers to the unity objection and to the other problems raised at the end of Part One, along the following lines. It is observed that the stratified conception doesn't consist only of particular definitions, but also of more general results, in particular Convention T, the criterion for correctness of definitions. The main claim of Part Two is that Convention T is that which confers unity on the many definitions, making them analyses of a single concept. The problem of Part Two is to make sense of this idea without contradicting the main negative result of Part One, that no universal metalanguage exists. The strategy is to find the language in which Convention T can be formulated and show that it doesn't have to be a universal metalanguage. Chapter 4 contains a formalization of

[5] See for example Strawson (1949), Priest (1984), Putnam (1985) for statements of something like the objection from ordinary language. Others are mentioned in the coda.

Convention T, with the intention of laying bare the expressive resources required for its formulation. It is shown that the language of Convention T cannot belong to the regimentation scheme in and for which definitions of truth are given, but requires some new expressive device, which I dub *abstract generality*. Chapter 5 contains the theory of this expressive device. It involves a digression into the field of pragmatics, in particular into David Kaplan's theory of indexicality. Briefly, I argue that the standard Kaplanian formal semantics for indexicals is not adequate to Kaplan's own philosophical thesis that indexicals refer directly. The direct reference thesis, I argue, requires a new expressive device. Taking up the cues from Kaplan's thesis, I analyze the concepts of context, linguistic agency and object, and come up with new semantic notions of *abstract reference* and *abstract objects*. Together they allow a precise understanding of the device of abstract generality, for the case of indexicals. I give a formal description of a new kind of language, an abstract pragmatic language, which has a device of abstract generality.

Chapter 6 contains an adaptation of these notions from the case of indexicals to the case of regimented languages. In the new abstract metalanguage we can formulate Convention T in a way that doesn't imply a universal metalanguage. The concluding Chapter 7 gives explicit answers to the objections of Part Two, revises the approach to concept analysis that was adopted in Chapter 1, and offers a brief discussion of some broader philosophical implications of the stratified approach.

This essay is not meant as a historical or scholarly work on Tarski, but nor is it entirely free from exegetical presumption. As will become clear, especially in Part Two but also in many points in Part One, my stratified conception of truth contains many elements that could not plausibly be ascribed to Tarski. However, in general the development of these elements is inspired by close attention, first, to Tarski's formulations of the philosophical aspects of his work, and second, to their historical context. This connection works both ways: a philosophically viable stratified conception of truth, even if not historically Tarski's, faces problems that Tarski's historical position may have faced, and can accordingly help us improve our understanding of Tarski's choice of words.[6]

The defining feature of Tarski's stratified conception is the fact that on it truth for a language can only be defined in a distinct metalanguage. This is the feature under which the conception is usually considered and objected to. There are, however, other aspects of it that are not always considered essential to it and are therefore often ignored in the literature on truth. It is my contention that at least

[6] There is no dearth, fortunately, of Tarski scholarship. See Feferman and Feferman (2004) for a philosophically and mathematically informed biography. Patterson (2012) gives a historical account of the development of Tarski's analyses of the concepts of truth and logical consequence. Patterson (2008a) is a collection of essays about various aspects of Tarski's work. See Woleński (2012) for a comprehensive account of the philosophical school Tarski was part of, the Lvov-Warsaw school of logic and philosophy. Other works are referred to in the body of the text.

some of these aspects are philosophically significant. I mention three here. First, the restriction to regimented languages is often understood as an artificial device with no philosophical motivation, at best tolerated as a simplification on the way to a full account. On the contrary, I take the explication of truth for regimented languages to be philosophically primary. Second, writers on truth who are not exclusively interested in natural language are usually interested in mathematical truth. Such writers almost invariably make use of arithmetization of syntax rather than, with Tarski, a direct syntactic theory. I will develop the theory in terms of strings of letters, not of arithmetized syntax. This is more a shift of emphasis than a substantial philosophical difference, but it allows insight into some issues that are less visible on the arithmetized approach. The third and most important point on which I will stay closer to Tarski is the emphatic demand to provide an explicit definition of truth. Many theorists either underemphasize the role of the explicit definition, or forsake it altogether. Although this tendency has led to many valuable technical results, I will contend that spelling out the deep structure of the explicit definition holds the key to the philosophical significance of the concept of truth.

PART ONE

DEFINING TRUTH

1 Preliminary Methodological Issues

1.1. Concepts

The purpose of this work is to offer an analysis of the concept of truth that will shed light on the philosophical issues involved. There is no general agreement on what the nature of concepts is and what it means to analyze them. Regarding their nature, I will make the minimal assumption that says that a concept is something that applies to objects. Correspondingly, I take the analysis of a concept to reveal the necessary and sufficient conditions for its application to an arbitrary object. I also assume that these conditions are expressible in language, thereby excluding concepts with ineffable application conditions, if such there be. An analysis may consequently be identified with a statement of application conditions. Such a statement is called a *definition*.

A theory or a definition should be judged by its import, by what it commits its proponent to. More concretely, it should be judged by the collection of sentences that it logically entails. If a definition implies an absurdity, even when hard to detect, then it can't function as an analysis. The problem is that language, as it stands, does not allow us to state a definition in such a way that its import will be determinate, since there is no well-defined relation of logical entailment between sentences in language. The issues are well known. Sentences can be ambiguous with respect to the logical form of the content expressed: e.g., the sentence

(1) A boy danced with every girl,

is ambiguous with respect to the relative scope of its two quantifiers (must there have been one boy dancing with all the girls?). In addition, different concepts can be expressed by the same string of letters (e.g. "bank" as the side of a river, or as something that's too big to fall?). Such ambiguities are a serious problem, since due to them there is no fact of the matter in general whether one sentence logically entails another. But then we cannot use definitions as our concept analyses.

In order to avoid or overcome these problems, we can adopt in advance a set of norms to which the discourse must conform, called a *regimentation scheme*, or a *logic*. A regimentation scheme specifies the available sentence forms and the entailment relations between them. For example, we can decide to avoid using sentences of the form of (1) in favor of special forms that are unambiguous with respect to quantifier scope relation. The regimentation scheme that I will adopt in this essay is first-order logic. It consists only of predicates (expressing concepts), singular terms (referring to objects), truth-functional

connectives and quantifiers. First-order regimentation is conceptually very simple and unobjectionable, and its logical properties are well understood. I will have occasion to discuss the expressive shortcomings it might have later in the essay. Adopting a regimentation scheme solves the problem of structural ambiguity.

To handle lexical ambiguity, we need to make a precise and unambiguous list of all the basic concepts that a definition uses. If a certain piece of discourse contains concepts that are usually expressed by the same word, such as "bank", we artificially distinguish the words by using a synonym or a typographical addition (or by using artificial symbols to begin with). Another kind of ambiguity, that was not yet mentioned, is ambiguity with respect to the universe of discourse. A sentence such as (1) will not usually be used to speak about every girl whatsoever, but only about every girl in some context, some party, say. This information is not expressed in (1), a fact which again makes it impossible to determine logical relations on the basis of the sentence alone. In order to avoid this ambiguity, we can say explicitly in advance which objects some stretch of discourse will be limited to. A specification of the list of basic concepts and of the universe of discourse is a *regimented language.* Following Quine, we say that the basic concepts (the predicates) in a regimented language form its *ideology* and that the objects to which it refers (the domain of quantification) are its *ontology*.[7]

The scheme I will adopt corresponds to first-order extensional logic with identity. The available sentence forms are predication (including identity statements), truth-functional connection between formulas, and quantification of formulas with respect to individual variables. The logical consequence relation I take to be classical, which means in particular that the law of excluded middle holds (and of course, the law of non-contradiction).

The analysis of a concept is given by a definition, which is a sentence of the following form:

(2) $x\ is\ P\ if\ and\ only\ if\ \phi$.[8]

Here "P" is to be replaced by the (unambiguous) name of the concept to be defined (the *definiendum*) and "ϕ" by a first-order formula (the *definiens*) with at most "x" as free variable, which states the conditions under which P applies to x. (An implicit "for all x" is understood to be present and have top scope.)

[7] The term "regimentation" is from Quine (1960, ch.5). Tarski speaks of "formalized languages" (1956, p.165f), but the idea is the same. For "ontology" and "ideology", see Quine (1951; 1953, p.131). Notice that Quine is speaking of theories, not of languages. This is a complication I won't go into (though see a comment in §7.5).
[8] This is the form for unary concepts. Generalizing to relation concepts is straightforward.

A definition is successful as a concept analysis if it is, in Tarski's words, *formally correct* and *materially adequate*. Material adequacy will be the topic of Chapters 2 and 3. Formal correctness is a condition on definitions independent of the particular concept to be defined. A definition is formally correct, first of all, only if it is stated in a regimented language, the ideology and ontology of which are clear and unambiguous.[9] For a concept *P*, a subclass of a definition's ontology that deserves special mention is the class of *P*-bearers, so to speak. These are objects to which the concept *P* stands to be applied, or in other words, the objects from which it is significant to withhold the concept. For example, although the concept *daisy* fails to apply to, say, chairs, this is not usually a fact of any interest. It is (say) only with respect to flowers that it is it possible to apply it and interesting to deny it. Flowers are in this case the bearers of the concept *daisy*. In the general case, for a definition of *P* to be formally correct it is important that the domain of *P*-bearers be given clearly, since it is in terms of the properties of the *P*-bearers that it is determined whether *P* applies to them or not. If the language in which the definition is given doesn't express the properties relevant to being a *P* (in the case of *daisy*, either morphological or genetic properties), then a definition of *P* will not be possible.

Another important condition of formal correctness is that the definition not be circular. This condition needs to be fulfilled both in ideology and in ontology. On the side of ideology the condition amounts to saying that the definiens must not mention the definiendum, nor any concept understood in terms of the definiendum, nor any concept understood in terms of such a concept, etc. When we give a definition we are effectively assuming that the concepts in the definiens are understood and unproblematic. On the side of the ontology, it is required that the specification of the domain doesn't involve the definiendum, i.e. that the objects over which we quantify are given determinately to begin with, and don't depend for their properties on the concept to be defined. This kind of circularity is called *impredicativity*.

The formal correctness of a definition guarantees that it is determinate for every object of the domain whether it falls under the concept or not, and that the definition doesn't lead to absurdities.[10]

---

[9] Tarski (1956): "The question how a certain concept is to be defined is correctly formulated only if a list is given of the terms by means of which the required definition is to be constructed. If the definition is to fulfil its proper task, the sense of the terms in this list must admit of no doubt." (p.152**)**

[10] These conditions might not be exactly what Tarski means by "formally correct", but they come close. See Hodges (2008), pp. 115-117 for some information. Notice that, as Hodges remarks, it is the definition and not the definiens that must be consistent. A definition with an inconsistent definiens is a formally correct definition of the empty concept.
There are various approaches that weaken the ban on circularity. See Gupta and Belnap (1993) for such an approach to the concept of truth. See Priest (2006) for an approach that purports to be tolerant to absurdities (or at least to contradictions). I don't engage with these approaches in this work.

1.2. Truth-bearers

What can we say about the truth-bearers? Loosely speaking, truth is a concept that applies to judgment-type representational objects. By judgment-type, still speaking loosely, I mean representations of how things are (e.g. that snow is white) rather than of things (e.g. snow). In the philosophical literature one finds various such objects (or would-be objects): sentences, statements, thoughts, propositions, beliefs, etc. They differ, sometimes to a great degree, with respect to their properties and constitution. For instance, a sentence is made up of words, a belief is held by a cognitive subject, a statement is made at a time, etc. Since a definition of a concept is stated in terms of the constitution of the concept-bearers, a definition that covers all of the various truth-bearers would be a pretty heterogeneous thing. However, since the truth-bearers are related in systematic ways, it makes sense to choose one of them as the *basic truth-bearer* and find a basic definition of truth in terms of that kind's constitution. If we wish, we can then define truth derivatively for the others. For example, one can treat propositions as the basic truth bearers and define a true belief as one with true propositional content, or a true sentence as one which expresses a true proposition, etc.

Following Tarski, my basic truth-bearer will be the sentence.[11] The reason is that I can give a pretty good account of the constitution of sentences. However, in many cases sentences are not good truth-bearers because they do not express judgments, or they do not express them adequately. For example, the sentence

(3) It is raining,

is not a truth-bearer as it stands, but only when in fact used. This suggests that not sentences, but acts of their use, or *utterances*, should be truth-bearers. Second, even when a sentence does express a judgment, the sounds or marks of which it is composed are not essentially connected to that judgment. For example, the same judgment will often be expressed by several different sentences, say in different languages. Worse, it is sometimes the case that different judgments can be expressed by the same string of marks, as we saw in sentence (1). The inessentiality of the marks is especially visible when we want to speak about the judgment itself, for example to say that somebody holds it. I may wish to say that the Greeks knew all the true sentences of geometry. But surely they didn't know the *sentence* "the sum of the angles of a triangle is equal to two right angles", since that sentence is in English. These considerations, and related ones, lead many theorists to adopt *propositions* as truth-bearers – the contents of sentences shelled of their inessential sensible husk.

There are, however, serious problems with taking utterances and propositions as basic truth-bearers. The deeper philosophical problems, especially with respect to utterances, will be central to Chapter 5.

[11] The declarative sentence, of course. See Rojszczak (2005) for the history of Tarski's choice.

For now let me just say that since the constitution of utterances and propositions is less than clear, making them the basic truth-bearers compromises formal correctness. With utterances, the problem is that the notion of context is not clear enough, and there is, to the best of my knowledge, no theory that presumes to give a complete and reliable account of what contexts are. With propositions the problem is the reverse: there are a great many theories of propositions, and as many objections to each. Moreover, many of these theories rely on certain objects in themselves not clearly enough defined, such as possible worlds, or possible situations, etc. Since a definition of truth needs to be stated in terms of the constitution of the truth-bearers, the problem is that we don't exactly know what the constitution of utterances and propositions is. It therefore seems safest to develop a theory of truth for sentences first, and then apply it, if desired, to utterances and propositions.

Still, we must answer the objections raised against sentences. The main objections were context-dependence and possible mismatch between a sentence's content and its outward form. The mismatch is problematic especially in two cases: when it creates ambiguity (as in (1)), and for propositional attitudes and intensional contexts. The context-dependence problem and the ambiguity problem can be handled by taking as truth-bearers, not naturally occurring sentences, but sentences in regimented languages, of the kind that we use in order to express our philosophical definitions. In Part One I will ignore the issue of intensionality.[12] Regimented languages thus play a double role for us: they are both the medium in which we state definitions, and they furnish us with the objects to which the definition stands to be applied.

1.3. Formalized languages

We chose sentences because their constitution is relatively clear and unproblematic. It is incumbent upon us to give a clear and unproblematic definition of them. The concept of a sentence is relative to the concept of language, so really what we need to define is the concept of (regimented) language. I will limit the definition to first-order extensional languages.[13] My definition won't diverge much from the ones ordinarily found in logic textbooks, except that in logic, the relevant notion is that of truth in a model, and our topic is what is sometimes called *absolute truth*, or *truth simpliciter*.[14] Also, I aim to be

[12] Extending the stratified conception to intensional contexts lies beyond the scope of this essay.
[13] Tarski (*CTFL*) actually uses a simple type-theoretic framework, but the transition to first-order semantics is already present in the (1935) postscript.
[14] See *CTFL*, p. 199 for this use of "absolute truth" (not in contrast with "truth in a model" but with "truth in a domain"). Gupta and Belnap (1993) are, I think, mistaken to identify absolute truth with "truth in the unique model that represents the actual world" (p. 22f). This is because absolute truth doesn't require the existence of a model. For example, "snow is white" is absolutely true because snow is white, not because there exists a model in which the sentence is true and that model represents the actual world.
Carnap (1942) uses the term "absolute truth" to denote truth for propositions, in contrast with "semantic truth" which means truth for sentences in a language. So Tarski's "absolute truth" is the contrary of Carnap's "absolute truth". Twardowski (1900) argues that truth is absolute in the sense that it properly applies to judgments with no

more explicit than customary about the philosophical or ontological grounding of the notions I use. I take a language to consist of a *phonology*, a *lexicon*, a *syntax* and a *semantics*.

1.3.1. Phonology

Phonology is the sensory aspect of language. Language can affect the senses in several ways, or sensory modalities, the main ones being audial, in speech, and visual, in written or typed text. I will present phonology in terms of typed text, but the idea is to abstract from any particular choice of sensory modality. The phonology of a language consists of an *alphabet* together with the concatenation operation. An alphabet is a finite set of symbols called *phonemes*. A *string* is the result of the successive concatenation of finitely many phoneme tokens. The precise nature of this operation depends on the sensory modality; in typed text it means placing one phoneme after the other on a line.

The sharp-eyed reader will have noticed that the concept of string has not been properly defined. The straightforward definition would be inductive: a string is any phoneme, or the result of concatenating a phoneme to a previous string. However, inductive definitions are not formally correct by our standards.[15] Rather, here the definition is *genetic* in something like Kleene's sense in the case of numbers, i.e. a definition by which the objects are "generated or constructed in a certain orderly manner".[16] This talk of "generation" must be mere metaphor, since it is not the case that the numbers are brought into existence at any time. It is not clear, however, how to state the matter without metaphor. In any case, my intention is not the same as Kleene's, since Kleene identifies his intention with Brouwer's and Weyl's, for whom the important thing is that there not be reference to the entire set of inductively "constructed" items.[17] My own view is something of a compromise. I take strings to be a privileged concept, the one case in which an inductive definition suffices to fix a domain of objects. But it's not that the definition generates the strings; their existence is grounded in our sensory capacities, for example our conception of typed text. That typed text has the desired properties is an assumption, and if you are reading this then I am guessing that you know enough about typed text to agree with me. Unlike Brouwer and Weyl, I allow the collection of strings to be an object, where needed.

I am not pretending that phonology is free of philosophical problems. First, it is not straightforward to define a phoneme. The usual definition involves appeal to the semantics of a language and in any case a definition will have to take into account the entire alphabet.[18] Another issue is that the type-token

---

indexical elements, what Quine (1960, p.193) calls "eternal sentences", and what Kaplan (1989a, p.503) calls "perfect sentences".

[15] More on this in §3.2.

[16] Kleene (1952), p. 26.

[17] Kleene (1952), p. 48, quoting Weyl (1946).

[18] The problem is to find a suitable equivalence relation to determine when two things instantiate the same phoneme (are *allophones*). The most dominant definition relies on semantics (through the device of a minimal

distinction, silently appealed to above, is not that simple to draw.[19] And there are many other problems.[20] Despite all of these issues, I take phonology as here conceived to be a firmer basis than, say, possible worlds and contexts. [21]

Informally I will refer to phonemes and strings using quotations, e.g. "$abba$". When I need to explicitly express concatenation I do it with the symbol "·" (centered dot), for example:

(4) If $s = "ab"$, then $s \cdot "ba" = "abba"$.

I use Quine's corner quotes when quantifying into strings.[22] Then a different alphabet should be used for the variables – usually a Greek character will range over Latin characters, and "n" will range over numerals. Examples:

(5) $\ulcorner a\varphi \urcorner$ for $\varphi \in \{a, b\}$,

(6) $\ulcorner x_n \urcorner$ for $n \geq 0$.[23]

Here (5) refers to the strings "aa" and "ab", and (6) to the strings "x_0", "x_1", "x_2", … In general I will consider sub- and superscripted symbols to be a single phoneme.[24] Corner quotes will also be used in another way starting from §2.1. I will allow myself to be careless with quote marks when I judge that

---

pair, see e.g. Odden (2005, p. 44)). In any case, it is clear that the equivalence of two sounds or figures is relative to the entire alphabet. For example, in Ancient Latin the forms "v" and "u" were allophones (allographs), while in modern English they are not.

[19] The letter "a" occurs twice (is the type of two tokens) in the string "aa". The string "aa" occurs twice in the string "aabaa", which in turn occurs twice in this footnote. This text, I hope, is going to be printed, perhaps even in vast quantities. Then each token of "aabaa" in this footnote will be a type for an occurrence in a printed note. You can see how complicated things are going to get. See Bromberger and Halle (2000) for a discussion by working phonologists.

[20] E.g. the nature of the senses; the linearity of the concatenation relation; and many others, I'm sure, will have occurred to the reader.

[21] Two comments. (a) The connections between strings and numbers are deep and probably not accidental. First, we can define the natural numbers and the arithmetical operations in terms of strings and concatenation (see Quine (1946), Corcoran et al. (1974)). Quine and Corcoran et al. show the equivalence between phonological and arithmetical *theories* (first- and second-order, respectively). My topic has not been the theories, but the domains themselves. Second, we can define strings in terms of natural numbers. So if you're not convinced of the philosophical safety of phonology, we can take the natural numbers as our ontologically privileged concept (made by God, say), and use a coding scheme (made by Gödel) to define the strings. This is what most of the formal literature on truth does anyway.

(b) In accepting types I am rejecting a certain kind of philosophical nominalism associated with Tarski. In *CTFL* Tarski uses types, though somewhat apologetically (see p.156fn). Tarski's nominalism is addressed in Frost-Arnold (2008), which studies the protocols of the Harvard 1940 conversations between Carnap, Quine and Tarski. Frost-Arnold mentions two conditions, required by Tarski, on the acceptability of a regimented language: that the language be first-order and that its domain consist of physical objects. Phoneme types are not physical objects, so it seems they should be excluded. In this sense my framework is more liberal than Tarski, since I don't make the second requirement.

[22] See Quine (1940), §6.

[23] Notice the easy transition between numeral and number.

[24] This is only for convenience. The alphabet is still obviously finite.

no confusion is likely to result. I assume that there is a single finite universal alphabet, of which the alphabet of every language forms a subset.

1.3.2. Lexicon

The *lexicon* of a language is a finite set of strings over its alphabet. A member of the lexicon is called a *lexeme*. Lexemes will be the basic meaningful units of a language. The lexicon, being finite, can be defined by enumeration.[25] (I include infinitely many numerals $\ulcorner n \urcorner$ and the infinitely many symbols $\ulcorner x_n \urcorner$ in the lexicon without thereby ceasing to think of it as finite.[26]) We distinguish the logical part of a lexicon from the non-logical part. Since we keep the regimentation scheme fixed (first-order logic), we can keep the logical part constant across languages. It consists of the symbols: $\forall, =, \downarrow$, $\ulcorner x_n \urcorner \; for \; every \; n, (,$ and).[27]

If the lexemes are the basic meaningful units of the language, why begin with phonology at all? Phonology by itself is representationally inert, in the sense that phonemes do not have anything to do with the meaning of expressions. Surely we could replace every occurrence of "sh" in our English sentences by an occurrence of "ź" and get a language, call it Engliź, that has exactly the same representational power as English. Using arithmetical coding as a basis is no improvement, since the particular number used to represent an expression is just as irrelevant for its truth as the phonological string, as witnessed by the plurality of possible coding schemes. Since language begins to be meaningful only at the level of lexemes, it makes sense to ask whether it would not be better to drop phonology altogether from our concept of language. Let's call this alternative notion *the intelligible language conception*, in contrast with the *phonological* one, which I'm adopting. On this view the alphabet and the lexicon coincide. I won't spell out the intelligible conception in detail, but we can mimic it to a certain extent within the phonological conception by demanding that there be no multi-phoneme lexemes:

(A) **The symbolization stipulation**: all lexemes are single-phoneme strings.

This is the common procedure in symbolized languages. In §2.2, and later in §4, we will see that postulating a representationally inert medium is crucial to the stratified conception.[28]

---

[25] E.g., $Lexeme(x) \; if \; and \; only \; if \; x = "P" \; or \; x = "Q", etc.$

[26] See footnote 24.

[27] We can view these symbols as abbreviations for longer strings: "for all", "equals", etc. The symbol "↓" is the Sheffer stroke, which will be a binary connective meaning "neither… nor …". Using it makes the definitions of the concepts of the language simpler (see Mendelson (1997, p. 29)). When formulating arbitrary sentences, I will make use of the other connectives and the existential quantifier. They are to be seen as abbreviations.

[28] I can't ascribe the intelligible conception of language to anyone with any certainty, but something along these lines might underlie the Language of Thought Hypothesis (see Crane (1990)); maybe also Chomsky's Y-model conception of language, see for example Chomsky (1993, p.168f). If propositions are to be regarded as

On most treatments, the lexicon of first-order languages consists of predicates and function symbols (the limit case of the latter being the individual constant).[29] However, we can without loss of generality assume our lexicon to contain only predicate symbols. If we want to regiment a given n-place function f, we can do it using an $(n+1)$-place predicate P_f such that $P_f(x_1, \dots, x_n, y)$ holds whenever $f(x_1, \dots, x_n) = y$. This has several minor advantages. First, in using function symbols we are presupposing that the value of the function exists and is unique for any sequence of arguments; but if we regiment everything using predicates, this presupposition needs to be stated explicitly. This is an advantage, since it makes the existence assumption part of the discourse and not a presupposition. If we like, we can reintroduce function symbols by contextual definition:

(7) $\ulcorner \phi\big(f(x_1, \dots, x_n)\big) = \exists y(P_f(x_1, \dots, x_n, y) \wedge \phi(y))\urcorner$.[30]

In practice I will use function symbols and individual constants without notice. The second advantage of limiting our lexicon to predicates is that our definitions become simpler. This will be felt presently. The third advantage is in philosophical interpretation: the non-logical resources of a language now fall neatly into two categories – objects (in the domain) and concepts (predicates). There is no need to find a philosophical gloss for functions or individual constants.

1.3.3. Syntax

The languages that we are interested in are *usable languages*. Being the finite beings that we are, we can't comprehend an infinitely long string. First-order languages allow the expression of infinitely many sentences, but if they are to be used, this infinite collection has to be specified using some single finite sentence. We therefore have:

(B) **The finitude constraint:** A language is given in a finite statement.[31]

---

representational content free of representationally inert matter, then considerations against the intelligible conception of language might count against propositions as well. I won't elaborate on this here.

[29] See for example Mendelson (1997, p.57).

[30] A contextual definition is not a definition in the strict sense, but as mere abbreviations of strings. Abbreviations such as (7) introduce ambiguity into our language, since they don't, e.g., distinguish between negations of abbreviations and abbreviations of negations. There are well-known ways to handle this, and I will not pursue these issues further (though see §5.2.4). Russell's theory of descriptions is the classic place to look for a discussion of this, see Chapter III of the Introduction of *Principia Mathematica*.

[31] In Frost-Arnold's (2008) report on the Tarski-Carnap-Quine protocols of 1940 (see footnote 21), besides Tarski's nominalism, it is also mentioned that acceptable languages have to be finitistic, in the sense that the domain of a language not be infinite, and that there be only finitely many predicates. Again, my finitistic requirement is a liberalization of Tarski.

A similar constraint is central to Davidson's concept of language[32] and to Chomsky's notion of a generative grammar.[33] Of course, Davidson and Chomsky are interested in unregimented language (though they have, perhaps, different conceptions of it).[34] I'm not saying there is not an interesting concept of unusable languages, for example one with infinitely long sentences, but for our purposes these are best treated as mathematical abstractions, derivative from the primary object of investigation. In a sense, the distinction between usable and unusable languages plays for me the role that the distinction between natural and artificial languages plays for many philosophers.

Technically, we can treat syntax as a single predicate $wff(x)$ which holds of strings which are well-formed formulas. This predicate can be defined properly (i.e. with formal correctness), but its content is displayed better by an inductive definition:

(C) Inductive definition of $\boldsymbol{wff}$:

a. If α_1 is a variable, then $wff(\ulcorner W(\alpha_1)\urcorner)$,
b. If α_1 and α_2 are variables, then $wff(\ulcorner(\alpha_1 = \alpha_2)\urcorner)$,
c. If $wff(\phi)$ and $wff(\psi)$ then $wff(\ulcorner(\phi \downarrow \psi)\urcorner)$,
d. If $wff(\phi)$ and α is a variable, then $wff(\ulcorner\forall\alpha\phi\urcorner)$.[35]

This definition has two base clauses specifying the atomic formulas. Clause (a) treats of the non-logical predicates, and it changes from language to language based on the lexicon. In the example I assumed only a single unary predicate "W". The other clauses are common for all first-order languages. The inductive definition is the most transparent way to carve out the infinite subset of formulas from the infinite set of strings, since it traces the "generation" of each complex expression from the finitely many lexemes. The problem is, as in the case of strings, that it is not a formally correct definition. In this case, however, we wouldn't want to make it a genetic definition, and we don't need to. With any wff we can associate a sequence of shorter formulas that were steps in the imaginary generation procedure according to the inductive definition. We can represent such a sequence using a string, and this allows

[32] Davidson (1965).
[33] See Hauser, Chomsky and Fitch (2002). Chomsky uses the term "discrete infinity". I'm not sure what this means beyond countable infinity, or an infinity produced from combinations over a finite set. See Pullum and Scholz (2010) for a critique of the view that (natural) language has infinitely many sentences.
[34] Note the difference between how the infinite number of sentences is "generated" on Chomsky's view and on the present one. For Chomsky it is the syntax that merges lexemes and complex phrase structures together into more complex phrase structures "recursively", so that the number of well-formed sentences is infinite. On my view an infinity of strings is provided by the phonology, and the task of the syntax is to carve out the well-formed subset of them (also, in another sense, recursively; see below). This is why I am tempted to ascribe to Chomsky the intelligible language conception from above, see footnote 28.
[35] The parentheses in wffs will be used as usual and omitted according to the usual conventions.

us to define wff explicitly, by quantification over strings.[36] It is straightforward to define a function which takes a wff to the number of its free variables. Henceforth I will write $wff^n(x)$ to say that x is a well-formed formula with n free variables. A sentence is a string of category wff^0.

Since both the lexicon and the syntax are defined explicitly in terms of phonology alone, phonology is so far the sole basis of our conception of language.

1.3.4. Semantics

But the syntactical or phonological notion of sentence is not the one relevant for truth. Truth is a concept that applies to sentences as interpreted, not as mere strings. In general, the *interpretation*, or *semantic theory*, of a language L is a specification of what each expression of L means. We can represent it as a function $\lambda x.\,|x|^L$ on strings, such that for s a string, $|s|^L$ is the meaning of s in L.[37] For reasons that will be explored in depth, the letter L here cannot be a variable – to every language there corresponds a different interpretation function.

The question of the nature of meaning is a highly vexed one in philosophy, and I prefer to prejudge it as little as possible and stick to a minimal view of meaning. We can make two assumptions. The first is that meaning is compositional:

(D) **Compositionality of meaning:** the semantic value of a complex expression is determined by the semantic values of its lexemes and its syntactic composition.

This assumption stems from the fact that our languages need to be usable and it actually a little stronger than we need. Strictly, it is enough that there be a finitely formulated specification of the meaning of every expression (or at least every sentence) of the language. Since strings are finite and there are finitely many lexemes, (D) guarantees this.

The second assumption is that semantic theory suffices for the truth of sentences, in the sense that from the meaning of a sentence its truth conditions are derivable:

(E) **Truth-sufficiency of meaning**: A semantic theory for a language entails a truth theory for it.

This assumption is meant to exclude semantic theories not based on truth, such as translational or "internalistic" accounts of meaning, according to which meanings are given in terms of further

[36] See Quine (1946) for an example of an explicit definition based on an inductive one. It is easy to generalize the example into a method that will work for any inductive definition, provided that the generation sequence can be represented in a string.

[37] I will omit quote marks within the scope of $|x|$, so that, e.g., $|P|$ is the same as $|"P"|$.

representational entities (semantic markers, or cognitive representations, and the like).[38] The assumption is compatible with most of the dominant paradigms in linguistic semantics practiced, which have their origin in the works of Montague and Davidson.[39] Without this assumption sentences cannot be truth bearers.

The non-logical lexicon contains only predicates, and the logical lexicon is the same for all languages. We can therefore identify a language with the interpretation of its lexemes plus its domain D^L – effectively its ideology and ontology. For convenience, we can say that the domain is the interpretation of the symbol "∀",[40] and identify a language L with its semantic function $\lambda x.\,|x|^L$. The form of semantic theory thus comes down to the kind of meaning that predicates have. We are interested in extensional languages, which means that whatever meaning is, for our purposes it is enough to look at its extensional aspect. A straightforward approach is to assign to each lexical predicate its extension. It is then routine to show how the extensions of complex predicates is calculated from the extensions of their components depending on the mode of composition, in a way that satisfies (D). One problem with this is that sentences don't have extensions. In this case we take two arbitrary objects, say 1 and 0, and put $|s| = 1$ if s is true and $|s| = 0$ otherwise.[41] This is of course only a figure of speech. To say that $|s| = 1$ is simply to say that $True(s)$, but in a way that allows for notational uniformity of the semantics. This is the *model-theoretic* style of semantic theory.

The model-theoretic style relies on the assumption that the extension of every predicate is an object (usually a set, but we can be more liberal). This is too restrictive, for we know of cases in which there is no object that corresponds to a certain predicate (for example, in the case of proper classes in ZF set theory). In such cases we cannot think of the semantic theory as a function in the proper sense, but we can still specify the semantics of predicates by stating their application conditions. This is called the *truth-conditional* style. Thus if in the model-theoretic style, the extension of a predicate P would be a collection, say:

---

[38] See Lewis (1970). See Pietroski (2005) and Glanzberg (2014) for recent gestures towards internalistic interpretations of semantics.

[39] See, e.g. Chierchia and McConnell-Ginet (1990), p.61ff; Larson and Segal (1995) p.25ff; Heim and Kratzer (1998), p.1; Jacobson (2014), pp.27ff.

[40] This is Enderton's (1972) procedure (p. 81).

[41] This device is, of course, Frege's. Frege used it to make his semantic system uniform in assigning reference to every well-formed expression. This is formally elegant, but it opens the door to a fallacy. If interpretation is a mathematical function, and if sentences refer either to 1 or to 0, then we are tempted to entertain the possibility that they might refer to something else as well, or to nothing at all. But the definition went: 1 if true, 0 otherwise. Can a sentence fail to be either true or otherwise? The fallacy is in thinking that 0 models something over and above absence of truth in a sentence, maybe some independently conceivable notion of falsity. This problem exists already in truth-table semantics for the sentential connectives, where we a lot one line in the table for truth and another for "otherwise". We are then tempted to add another line, which is neither truth nor otherwise. I call this *the algebraic fallacy*. We should therefore keep in mind that the "$|s| = 1$" is just a figure of speech, and really what we are saying is "$True(s)$".

(8) $|P| = \{x: x\ is\ a\ philosopher\}$,

on the truth-conditional style we can only say what the conditions are for a predication of *P* to be true:

(9) $|Px| = 1\ if\ and\ only\ if\ x\ is\ a\ philosopher.$

On the truth-conditional style we don't assign reference to subsentential expressions. But if sentential reference is only a figure of speech (see above), then an interpretation function is no more than a truth predicate or, to be more precise, a predicate expressing the satisfaction relation between objects (or sequences thereof) and formulas. On the truth-conditional style, we therefore identify a language with the extension of its satisfaction relation.

It may seem to you that semantics as we are defining it depends to closely on truth (and satisfaction). If truth applies to sentences only as interpreted, conceiving of interpretation in terms of truth sounds awfully close to being circular. More precisely, we have a *diallele* holding between truth and meaning.[42] I will return to this problem in §3, and then later in §7.

1.3.5. Sentences as truth bearers

We now have a good understanding of our truth-bearer, the sentence. A sentence of a language L is a phonological string, a wff^0 (well-formed formula with zero free variables), considered in relation to a semantic function $\lambda x.|x|^L$ (which might be shorthand for a satisfaction predicate). The language relativity of sentences yields the following, perhaps surprising, result:

(F) **Individuation of sentences**: let s_1 be a sentence of the language L_1, and s_2 a sentence of L_2. If $L_1 \neq L_2$ then $s_1 \neq s_2$.

If s_1 and s_2 are the same phonological string, we say they are *homophones*. If $|s_1|^{L_1} = |s_2|^{L_2}$ we say that s_1 and s_2 are *synonyms*. But by (F), homophony and synonymy are not sufficient for identity. Two sentences are the same only if they are in the same language.

With these preliminary stipulations and definitions in hand, we can move on to our main problem, that of defining truth.

---

[42] A *diallele* (δὶ ἀλλήλων) is the circumstance in which a concept is understood in terms of another and vice versa. It is one of the skeptical tropes used by Sextus Empiricus to undermine philosophical theses. Particularly relevant to our case is his attack on the Stoic concepts of truth and of the *grasped representation* (φανταζία καταληπτική). See his *Against the Logicians*, I.426, II.85. In brief, his complaint is that the Stoics defined the grasped representation as that which represents what exists, and they define what exists to be that which is given in a grasped representation. I was led to this notion by Kant (see footnote 44).

2 Naïve Truth

The previous chapter was concerned with the formal correctness of definitions. But a definition might be flawless formally and still fail as an explication of a concept, if it doesn't capture the content of the concept to be defined. A definition must therefore conform to something which is somehow given to us beforehand. This conformity is called by Tarski the *material adequacy* of a definition. What it means in general for a concept to be "given to us beforehand" is a difficult question. In the case of the concept of truth, the starting point is what can be called *the naïve semantic conception of truth*. The present chapter discusses the naïve conception and its refutation in the form of the liar paradox, corresponding in essence to §1 of *CTFL*.[43]

2.1. The naïve-semantic conception of truth

2.1.1. Statement

We are looking for a formula of the form:

(1) $True(x) \leftrightarrow \phi(x)$,

where "ϕ" is replaced by a statement of the necessary and sufficient conditions for an arbitrary sentence x to be true. Loosely speaking, a sentence is true just in case the state of affairs that the sentence reports indeed holds. In other words, a sentence expresses its own truth conditions, for example:

(2) $True(\text{"}the\ sea\ is\ blue\text{"}) \leftrightarrow\ the\ sea\ is\ blue.$

It follows that for every sentence, we have a statement of its truth-conditions in reach: the sentence itself. It should be straightforward to generalize this notion into a statement of the form (1). For consider an analogous case. Imagine that the following is a satisfactory definition of the concept bachelor:

(3) $Bachelor(x) \leftrightarrow Male(x) \wedge (age(x) \geq 18) \wedge \neg\exists y Wife(y, x)$.

[43] The English term "materially adequate" is from Tarski (1944), and is used in the (1956) English translation of *CTFL*. In the Polish version of *CTFL* we find "merytorycznie trafna" and in the German "sachlich zutreffende". A more literal translation would be "contentually accurate" (if "contentually" were a word).
Carnap, in the English version of *The Logical Syntax of Language* (1937), uses "material" to translate "inhaltlich" (see §§77-81), which, like "sachlich", would more literally would be rendered "contentual". This might be the origin of Tarski's "materially". I remember reading Quine reminisce about being the one who suggested to Carnap "material" for "inhaltlich", but I can't find the place. See Hodges (2004), (2008), Patterson (2008b), (2012, §4.1.1), for discussion.

This is a statement of necessary and sufficient conditions for bachelorhood for an arbitrary person. It follows from this definition that for every person, say Jan, we have a particular statement of these conditions:

(4) $Bachelor(Jan) \leftrightarrow Male(Jan) \wedge age(Jan) \geq 18 \wedge \neg\exists y Wife(y, Jan)$.

The relation of (3) to (4) is that of the universal to the singular. We would think that (1) and (2) would be related in the same way, but this is not the case. There is no way to get (1) from (2) by generalizing over a singular term. The problem is that although a sentence expresses its own truth conditions, it is not mentioned in the statement of those truth conditions in the way that John is mentioned in the statement of his bachelorhood conditions. This is *the generalization problem*.[44]

In various fields in and around mathematical logic we often use the device of a schema in order to generalize over sentence position. This lets us formulate the *naïve-semantic conception of truth*, which says that the concept of truth is such that we expect it to entail all instances of the *disquotational T-schema:*

(A) **Disquotational T-schema**: $True(\ulcorner\phi\urcorner) \leftrightarrow \phi$,

where an instance is had by replacing "ϕ" with some sentence of the language in question *L*, and "$\ulcorner\phi\urcorner$" with a name of that sentence.[45] These instances are called *disquotational T-sentences*. Sentence (2) is the instance had by replacing "ϕ" with

(5) The sea is blue.

We say that (2) is a T-sentence *generated* from (5). The naïve conception of truth amounts to saying that for every sentence of the language in question, the T-sentence generated from it is entailed by (the

---

[44] This problem is stated by Kant in the *Critique of Pure Reason*, A58-9/B83:

> If truth consists in the agreement of a cognition with its object, then this object must thereby be distinguished from others; for a cognition is false if it does not agree with the object to which it is related even if it contains something that could well be valid of other objects. Now a general criterion of truth would be that which was valid of all cognitions without any distinction among their objects. But it is clear that since with such a criterion one abstracts from all content of cognition (relation to its object), yet truth concerns precisely this content, it would be completely impossible and absurd to ask for a mark of the truth of this content of cognition, and thus it is clear that a sufficient and at the same time general mark cannot possibly be provided. Since above we have called the content [Inhalt] of a cognition its matter [Materie], one must therefore say that no general sign of the truth of the matter of cognition can be demanded, because it is self-contradictory.

See Prauss (1969, p. 177ff) for a discussion of the similarity between this passage and Tarski's generalization problem.

[45] On this formulation, the occurrence of "ϕ" within "$\ulcorner\phi\urcorner$" is inert, and is just meant to remind us that the latter is a name of the former. It would also have been possible to treat "$\ulcorner\phi\urcorner$" as a composite expression in which the corner quotes express an expression-naming device (not exactly the Quine quotes of §1.3.1, but similar enough). See below for expression naming.

definition of) the concept of truth. In another terminology, we can say that they are *analytic* to the concept. This is a condition on material adequacy: a definition that doesn't entail all the T-sentences is materially inadequate.

The main negative result of *CTFL* is that the naïve conception is untenable. The reasons will be given below (§2.3). But if it had been tenable, the naïve conception would have had important philosophical benefits. That's why many philosophers, especially since the 1970s, have tried to maintain it by negotiating some other assumption made by Tarski, usually about the logical framework. Except in passing, I will not discuss the other strategies in this work. I turn now to some of the philosophical benefits a tenable naïve conception would have had. The first is the fact that it is intuitive. As its name suggests, the naïve conception expresses our "pre-theoretic" concept of truth.[46] The other benefits take more explaining.

2.1.2. Escape from relativity

If sentences are our truth-bearers, then since sentences are only such relative to a language, truth becomes relative to a language. This is disturbing. Surely truth should be the first and foremost objective concept, dependent on how things are and not on how they are expressed. It is a virtue of the naïve conception that under it we can make the relativity to language benign, at least with respect to a certain privileged class of languages.

In §1 of *CTFL* Tarski discusses something like the naïve conception in relation to a language he calls *everyday language*.[47] In Tarski's usage this is not a general term under which several particular objects such as English and Hebrew fall, but a certain mode of discourse, something like unregimented discourse.[48] Tarski describes everyday language as *universal*, in the sense that anything that can be expressed at all can be expressed in it:

(a) **Universality thesis for everyday language:** everything that can be expressed at all can be expressed in everyday language**.**

---

[46] See, e.g., McGee (1991): "[the disquotational T-schema] is so deeply embedded in our ordinary thinking about truth that we might fear that, once we decide to give [it] up, we should become so badly disoriented that we should not be able to talk about truth at all" (p. vii).

[47] "Język potoczny" in the original Polish; "Umgangssprache" in the German; "colloquial language" in the (1956) English. I take "everyday language" from Tarski (1944), originally published in English. In that paper Tarski uses both the terms "everyday language" and "natural language", though not, I think, interchangeably. See below.

[48] That for Tarski everyday language means unregimented discourse and not an object such as English is visible in places such as (1956, p.60). See also *CTFL*, p.267, where he speaks of philosophers who think that "the one natural language [is] colloquial language"; the Polish for "colloquial language" here reads "język życia potocznego", more precisely translated as "the language of everyday life" (p.116).

One also finds in the philosophical literature heavy use of the expressions *natural language* and *ordinary language.* They are in general used interchangeably, and in two senses. Sometimes they are used in the sense of Tarski's "everyday language", to refer to unregimented discourse.[49] But often we find expressions such as "English and other natural languages".[50] Here it seems that not a mode of discourse is meant, but a general concept under which several particulars fall. In order to be clear about the distinction, let's call languages such as English, Hebrew, etc. *world-languages*. When I don't care to distinguish between everyday language and world-languages, I use the term "ordinary language."[51] It is often assumed in the literature, sometimes implicitly, that world-languages are universal, in the sense described above. This yields:

(b) **Universality thesis for world languages:** everything that can be expressed at all can be expressed in any world-language.

From this thesis it follows that every two universal languages are intertranslatable: for every sentence in the one there is a synonymous sentence in the other.[52] Under this universality thesis, we can treat all world-languages as identical up to translation.

Another frequent assumption in the literature is that ordinary language is the primary object of study for the philosophy of language. Formalized languages are perceived by many as only peripherally important to major philosophical concerns. This gives rise to another thesis:

(c) **The primacy of ordinary language thesis:** ordinary language is our primary object of study.

This is actually two theses, corresponding to the two senses of "ordinary language". Combined with the appropriate universality thesis, the significance of the primacy thesis is that in the important case, that of ordinary language, we don't have to worry about language relativity.[53] This is an important advantage of a conception of truth that allows universality, such as the naïve conception, over one that doesn't, such as Tarski's.

---

[49] This seems to be the use associated with the early "ordinary language" philosophies and with Wittgenstein.
[50] See, e.g., Soames (1999, p.52), Priest (1984, p.117).
[51] The term "natural language" I reserve for something else, which will not be discussed in this work. Briefly, it seems most fitting for what linguists understand by it, a certain natural cognitive faculty. This is not unambiguous either. It might refer to an innate cognitive faculty (encoding a universal grammar?) or to the grammar implemented in the mind of an individual at a time (an I-language?), or maybe to something else. Notice that neither universal grammar nor I-language is the same as a world-language.
[52] I ignore the possibility that the unit of translation may be greater than the sentence.
[53] Most philosophers who subscribe to the primacy thesis subscribe to the universality thesis as well. An exception might be Kripke (1975), if his footnote 34 (p.714) is something to go by. See Burge (1979, p. 174fn) for a brief critique, also McGee (1991, p.91). Kripke says that "natural language in its pristine purity, before philosophers reflect on its semantics" might not be universal, but I'm not sure what he means by that.

2.1.3. Disquotationalism

The naïve conception is also important for the family of philosophical positions known as *deflationism.* The common core of deflationist theories is the claim that the collection of T-sentences defined by the T-schema exhaust the content of the concept of truth, and that therefore we should not expect a more informative definition that entails them. On this view the T-schema functions, not as a material adequacy condition on definitions, but as a definition in its own merit.

Where the truth-bearer is the sentence, deflationists sometimes call themselves *disquotationalists*.[54] A common philosophical gloss on disquotationalism is that truth, though it has no philosophical content, has a certain "function", sometimes dubbed logico-linguistic.[55] This function is to allow the indirect assertion of sentences, when for some reason the sentences themselves are not available for assertion, e.g. when they are not known or when there are too many of them to assert.

For example, consider the two sentences:

(6) Everything Alfred says is true,

(7) Alfred says that war is evil.

From (6) and (7),[56] using the disquotational T-schema (A), we can derive the sentence:

(8) War is evil.

So a person asserting (6), using the truth concept, will find himself committed to (8), even if he never as much as entertained (7), e.g. when he is endorsing all of Alfred's assertions on authority. In this way the naïve conception can serve as the engine for indirect assertion, without us needing to come up with a definition.

2.2. Syntactic closure

Unfortunately, the naïve-conception faces some serious difficulties. Most notably, the requirement it places on an adequate definition of truth is such that no formally correct definition can fulfill it. The problem is that it presupposes that the language in question *L* is semantically closed. A language is *semantically closed* if it expresses its own semantic theory, which for us comes down to being *syntactically closed* (expressing its own syntactic theory) and expressing its own truth predicate.

---

[54] See, e.g., Leeds (1978), David (1994), Williams (1999), Horsten (2011). The term “disquotation” is from Quine, and some ascribe to him the view as well.

[55] See, e.g. Williams (1999, p.547f).

[56] Regimenting (7) as direct discourse.

Syntactic theory is definable in phonology, so it's enough that L express its own phonological theory. For consider an instance of the disquotational schema (A), a disquotational T-sentence such as (2), reproduced here:

(9) $True(\text{"}the\ sea\ is\ blue\text{"}) \leftrightarrow\ the\ sea\ is\ blue.$

According to the instructions adjoined to the T-schema, the two occurrences of the string "$the\ sea\ is\ blue$" in (9), once mentioned and once used, are occurrences of the same sentence. This means that they are in the same language, by §1.3.5F. The used occurrence of the string is a subformula of the T-sentence, which means that they too are in the same language. Therefore, on the naïve conception, the language *for* which truth is defined is the same as the language in which it is defined.[57]

This section is about syntactic and phonological closure. The task is to define, in a formally correct way, a syntactically closed language. More generally, we will show how, on the basis of an arbitrary language L, we can define its syntactic closure $SynC(L)$. This is hardly a new result, but it will be instructive to reproduce it in the current setup.

2.2.1. A syntactically closed language

Let L_{Ws} be a language with a unary predicate "W" expressing whiteness and an individual constant "s" referring to snow. The domain contains snow together with the collection of strings over the alphabet of L_{Ws}. In a weak sense, L_{Ws} is phonologically closed since it quantifies over its own strings. But in order to define its own syntactic theory (the predicate wff,see §1.3.3) it needs to be able to refer to strings individually. It needs to have, not only the ontology of phonology, but also its ideology. We want to extend L_{Ws} to a syntactically closed $SynC(L_{Ws})$ that can also speak about snow being white.

Let $L_{\underline{Ws}}$ be like L_{Ws} with the addition of the phonological concepts for L_{Ws}: names for the phonemes and a symbol for the concatenation operation. A name for an L_{Ws} phoneme will be an underlined version of that phoneme: "$\underline{W}$" denotes "W" and "$\underline{\forall}$" denotes "$\forall$" and so forth. Concatenation will be marked by a centered dot "$\cdot$".[58] Between underlined names I omit it, so that both "$\underline{W} \cdot \underline{s}$" and "$\underline{Ws}$" denote the string "$Ws$". The lexicon of $L_{\underline{Ws}}$ then consists of the symbols $W, s, \forall, \downarrow, =, \underline{W}, \underline{s}, \cdot$. This allows us to define wff for L_{Ws} in $L_{\underline{Ws}}$, along the lines of §1.3.3. However, $L_{\underline{Ws}}$ does not have names for all of its own symbols; in particular, there is no way to refer individually to the underlined phonemes and to the

[57] Strictly speaking, we don't have to assume a full phonological theory. It is enough that L has names only for its sentences and not for other strings. However, since L will have infinitely many sentences, the only way to name all of them is through a full phonological theory. If this full phonological theory is given only in a strictly stronger metalanguage (as is the procedure of Gupta (1982)), then I don't say that the language is syntactically closed.

[58] To make concatenation a total function, let's say that the concatenation of anything with snow yields snow again.

centered dot. "$\underline{W}$" denotes "W", but nothing in the lexicon denotes "$\underline{W}$". $L_{\underline{Ws}}$ is not yet phonologically closed.

It obviously won't do to simply add further names to designate the unnamed phonemes (say, "$\underline{\underline{W}}$", "$\underline{\forall}$"), since the problem would arise again for these new names. This way of going about things leads to a regress:

(B) **The naming regress:** If every phoneme is named by a single phoneme, then a syntactically closed language will have an infinite alphabet.

This is because some phonemes will not name phonemes (e.g. "="), so we need a bijection between the alphabet and a proper part of it, which is only possible with an infinite alphabet.

We can avoid the regress if, instead of letting phonemes denote phonemes piecemeal, we systematize this relation and define an operator which takes an expression to its name. More precisely, since an object may have many names, we define an operator q which takes an expression to its *canonical name*. Here is an informal recursive definition:

(C) **Recursive definition of canonical name:**

a. The canonical name of snow is "s"; the canonical name of "s" is "$\underline{s}$"; etc.

b. The canonical name of an underlined phoneme α is $\ulcorner q(\alpha) \urcorner$, e.g. "$q(\underline{s})$" canonically denotes "$\underline{s}$".

c. The canonical name of a string is the concatenation of the canonical names of its elements in order, e.g. "$\underline{q} \cdot \underline{(} \cdot q(\underline{s}) \cdot \underline{)}$" canonically denotes "$q(\underline{s})$" (which is also denoted non-canonically by "$q(q(\underline{s}))$").

Let L^q_{Ws} be like $L_{\underline{Ws}}$ with the addition of three symbols: "q" for the operator q; "$\underline{q}$" for the phoneme "q"; and "$\underline{\cdot}$" for the concatenation symbol "·". The domain of L^q_{Ws} will include all strings over the enlarged alphabet (as well as snow). It is easy to see that every string of L^q_{Ws}, and in particular every sentence string, is not only an element of the domain, but also has a name in L^q_{Ws}. On this basis we can define L^q_{Ws}'s syntactic theory, and we have $L^q_{Ws} = SynC(L_{Ws})$.

The operator q was defined informally. We want to be sure that it can be defined in a formally correct way. In the literature we find two ways to mention expressions: quotation and structural-descriptive naming (arithmetization is a case of the latter). I'll examine them in turn.

2.2.2. Proper quotation

Quotation is often assumed to be an unproblematic, transparent means for naming linguistic expressions, so much so that its converse "disquotation" is held to be clear enough to explicate truth. But the problem of quotation is a difficult one, and its difficulties are not unrelated to the ones plaguing the concept of truth. These problems were already noticed by Tarski in *CTFL*, but have not played an important role in the debate on truth.

Sometimes "quotation" is used as a generic term for a device for naming strings, including under it structural descriptive naming as well. I wish to characterize *proper* quotation. To see what I mean by proper quotation, notice that in this very sentence, the word "sentence" is mentioned (as well as used). However, the letter "s" is apparently not mentioned in that sentence. On the structural-descriptive method this is impossible, as we will see shortly, since the name of a complex expression is composed of the names of its letters. This suggests the following characterization of proper quotation:

(D) **Characterization of Quotation:** Quotation takes an expression in use to its name.

If we want to make q into a pure quotation operator we need to define it so that it takes as argument, not a name of an expression, but the expression itself, and returns its name, e.g.

(10) $|q(Ws)| = "Ws"$.

This seems to conform to how we perceive quotation: in quoting an expression I am not using names for the phonemes making up the expression, though I am certainly using the phonemes themselves.

Implementing proper quotation faces some difficulties. Let L_q be like L_{Ws} with the addition of a proper quotation operator "q". The first problem is with syntactic theory. If L_q is to be syntactically closed, then the predicate wff should be definable in it. There is no difficulty in formulating the definition from §1.3.3(C). The sentence that says that the string "Ws" is a wff is expressed in L_q by the string:

(11) $wff(q(Ws))$,

and as desired, it follows from (10) and from the definition of wff. However, (11) itself is supposed to be a wff of L_q, but it is not captured by the definition in §1.3.3(C), because the expression "$q(Ws)$" is not well-formed: Its argument form is a sentence instead of a singular term.[59] One way to respond to this problem is to loosen our definition of wff so as to allow predicates to take arbitrary strings as arguments. Apart from being ad-hoc, this would overgenerate, since now a strings such as "$W(Ws)$"

[59] Recall that "$q(Ws)$" is the abbreviation (§1.3.2) of an expression containing a predicate "$P_q(Ws, x)$". But this last expression is an atomic formula that doesn't fall under the base clause of the inductive definition in §1.3.3(C).

would also be a wff. Another possibility is to add a clause in the definition of wff expressly for the operator q:

(12) If α is any string, then $\ulcorner q(\alpha) \urcorner$ is a singular term.[60]

This would give q a special status, something like a logical operator. Maybe this is the right way to go about it, but it at least deserves some discussion.

Similar problems arise at the level of semantics. Recall that by §1.3.4D (compositionality), the meaning of a complex expression has to depend on the meanings of the parts. Normally, then, we would expect for q a clause in the form of:

(13) $|\ulcorner q(\alpha) \urcorner| = f(|\ulcorner \alpha \urcorner|)$,

for some function f on meanings. But there is no function on the meaning of expressions which can do the work of quotation, as is evident from the fact that two different strings often have the same meaning. The semantic value of a quoted expression is not computed from the semantic value of the expression quoted, but from the expression itself:

(14) $|\ulcorner q(\alpha) \urcorner| = \ulcorner \alpha \urcorner$.

This means that "q" is not an extensional operator. But worse, it isn't any kind of intensional operator either. Like in the syntactic case, we would have to accommodate it expressly in our general semantic framework and give it the status of a logical operator.

These departures from our usual definitions are perhaps bearable. A more difficult problem is that having q on board would compromise the unambiguity of the language. Consider the L_q expression:

(15) $q(W) \cdot q(s)$.

The string (15) is given two different meanings by the semantics of L_q. We could use (14) to calculate the meaning of (15) if we knew what in (15) replaces "α" in (14). But (15) has two readings: on the first we have one application of q with $\alpha =$ "$W) \cdot q(s$", and on the second we have the concatenation of two applications of q, one with $\alpha =$ "W" and the other with $\alpha =$ "s", together yielding the string "Ws". In general, the problem is that if we allow any string whatever within the scope of q, then the string signaling the end of that scope, namely ")", becomes ambiguous, obliterating the distinction between use and mention.[61]

[60] Strictly: "if α is any string and x is a variable, then $P_q(\alpha, x)$ is a wff".

[61] Using Polish notation would only make things worse, since it depends even more substantially on expectations about how the function's argument will look like.

To sum up, proper quotation is not as innocent as one might assume from reading the literature on truth. Pending a responsible treatment, we shouldn't help ourselves to it. (I'll continue to use quotation in the informal discussion.) [62]

2.2.3. Structural-descriptive naming

The informal definition (C) of the canonical naming operator q uses quotation names, for example in the clause:

(16) The canonical name of snow is "s"; the canonical name of "s" is "$\underline{s}$"; etc.

In view of the problems with quotation, we would like to see whether q can be defined differently. In view of the naming regress, we can't use single-phoneme names instead ((16) would be formalized as "$q(s) = \underline{s} \wedge q(\underline{s}) = \underline{\underline{s}} \wedge$ ..."). The only way to avoid the regress is to reject its hypothesis that every phoneme is named by a single phoneme. The hypothesis follows from our practice of using a symbolized language in which every lexeme is a single phoneme (§1.3.2A). Rejecting it means that the lexicon will contain multi-phoneme strings: "for all", "snow", etc. Among these we can include multi-phoneme names of single phonemes, for example "ess" for "s".

Accordingly, let us define a language L_{SD}, whose lexicon contains the name of a binary function "$conc$" for concatenation, and names for all the phonemes: "ess" for "s", "$Doubleyoo$" for "W", etc. Names of strings are had by concatenating names of phonemes, e.g.

(17) $|Doubleyoo\ conc\ ess| = "Ws"$.[63]

In this way L_{SD} will have names for all strings over its own alphabet without recourse to a special quotation operator. Among other strings it will have names for its own lexemes, and among the lexemes are the names of the phonemes:

---

[62] These problems are raised in *CTFL*, §1. There is a swell of literature on the subject in the past two decades. See Saka (2013) for a good overview with references. A relatively early influential treatment is Davidson (1979) (it involves an appeal to demonstratives and to an unexplicated relation of x being a token of y). The problem with this literature is that it is concerned with a different problem, that of accounting for various quotation phenomena in natural language. See, e.g., the introduction the collection Brendel, Meibauer and Steinbach (2011), as well as the opening chapters of Cappelen and Lepore (2007) for a glimpse of what this literature tries to accomplish.
Boolos (1995) has a solution to the quotation problem for regimented languages, on which the form of the quote mark is determined according to the content of the quotation. His suggestion amounts to superscripting numerical indices to the parentheses of q, so that if α is a string and n is the greatest number such that α contains $\ulcorner)^n\urcorner$, then α will be quoted using parentheses $\ulcorner(^m\urcorner$ and $\ulcorner)^m\urcorner$ with $m > n$. This allows us to parse a string unambiguously in an effective way. But compositionality is thereby compromised, since now the meaning of an expression depends not just on the expression itself, but on an expression of which it is a part.

[63] When concatenating multi-phoneme lexemes I put a space between them. For uniformity, we can also think of the space existing between concatenated single phoneme lexemes.

(18) $|ee\ conc\ ess\ conc\ ess| = "ess"$.

Compare:

(19) $|ess| = \text{“}s\text{”}$,

(20) $|s| = snow$.

It is easy to see that L_{SD} is adequate to express the phonology and the syntax of any language, including itself. We introduce abbreviations:

(21) "$\underline{s}$" abbreviates "ess"; “ $\cdot$ ” abbreviates “$conc$”; "∀" abbreviates “forall”, etc.

(22) Convention: “ $\cdot$ ” is omitted between underlined abbreviations.

So that:

(23) $|\underline{s}| = |ess| = \text{“}s\text{”}$,

(24) $|\underline{ess}| = |\underline{e} \cdot \underline{s} \cdot \underline{s}| = |ee\ conc\ ess\ conc\ ess| = \text{“}ess\text{”}$.

The function q of canonical naming can be explicitly defined in these terms. With these abbreviations L_{SD} looks and feels like our usual symbolic languages, but under the hood it has a solid structural-descriptive engine that takes care of phonological closure.[64]

2.2.4. Conclusion

If L_1 and L_2 are languages, let $L_1 \cup L_2$ be the language whose ideology is their combined ideologies and whose ontology is the union of their ontologies. We wanted to find the syntactic closure of L_{Ws}, $SynC(L_{Ws})$. It can now be defined as $L_{Ws} \cup L_{SD}$. Quite generally, for every language L, we have $SynC(L) = L \cup L_{SD}$. This concludes the task of this section.

We also get some support for the phonological conception of language as against the intelligible conception (see §1.3.2). On the intelligible conception, language doesn't include a phonological aspect. The motivation for this view is the fact that phonology is representationally inert and the claim that language is essentially representational. But if the reasoning of this section is sound, then in order to have syntactic closure we must make a detour through some representationally inert medium (numbers

[64] The difference between this method and the more common method of arithmetization of syntax is that on this method we have direct reference to strings. To say of a string s that it is a wff can be true in a simple sense. By contrast, the method of arithmetization relies on a correspondence between strings and numbers. This is unproblematic in proof-theoretic contexts, but less natural when dealing with semantics philosophically. See Quine (1940, ch.7) for a phonological approach.

or strings). Otherwise, because of the naming regress (B) we fall foul of the finitude constraint (§1.3.3B).[65]

2.3. Refutation of the naïve conception

The naïve conception, we said, requires that the language it is about be semantically closed. A language is semantically closed if it is syntactically closed and has its own truth predicate (a predicate that conforms to the naïve conception). We recently managed to define a syntactically closed language $SynC(L_{Ws})$. Let L_{sem} extend $SynC(L_{Ws})$ with a predicate "$True(x)$" which entails all disquotational T-sentences as in (A). L_{sem} is semantically closed. We show it to be impossible.

2.3.1. The liar paradox

The problem is of course the semantic paradoxes. The informal version of the liar paradox ("this sentence is not true") is not reproducible within our regimentation scheme, but the paradox can be formulated using diagonalization in any semantically closed language.[66] We'll set up a general diagonalization schema for L_{sem} and then apply it to get the paradox. First, we define a restricted substitution function:

(25) $Sub(x, y) =$ the result of substituting a string x for the free occurrences of a variable "x" in a string y.[67]

For example, $Sub(\underline{Wx}, \underline{s}) = \underline{Ws}$. This function is explicitly definable in strictly phonological terms, along the lines of Quine (1936).[68] Now observe that the formula $\gamma = Sub(x, q(x))$ denotes the result of substituting a name of x for the variable "x" in x. In other words, if δ is a formula with "x" free, then $\gamma(\delta)$ denotes the self-application of δ. For example, $Sub\left(\underline{Wx}, q(\underline{Wx})\right) = \underline{W} \cdot q(\underline{Wx})$, which is a sentence which says that the string "Wx" is white.

---

[65] Compare Thomason (1975, p.127) and Richard (1986) who reach similar conclusions (Richard's term "quotation" means structural-descriptive naming). See also 201D8-202C6 of Plato's *Theaetetus* (Socrates' dream) for something like the intelligible vs. the phonological conception of language.

[66] Gupta (1982) shows that a language can have names for all its sentences, and a truth predicate, and fail to produce the paradox. But his language is syntactically closed only in a degenerate sense: it has names for its sentences but no string manipulation resources, so it can't express diagonalization. (See footnote 62.)

[67] The used variable "x" in the definition is unrelated to the mentioned variable "x".

[68] I'll give a sketch. Let $Fsub(x, y)$ return the result of substituting x for the final free occurrence of "x" in y. Then the following is a recursive definition of $Sub(x, y)$:

a. $Sub(x, y) = x \quad (if\ "x"\ is\ not\ in\ y)$
b. $Sub(x, y) = Sub(Fsub(x, y), y) \quad (otherwise)$.

This recursive definition satisfies the conditions stated in footnote 36 and can be transformed into an explicit definition. See Quine (1936) for (rather complicated) details.

Next let $\phi(x)$ be some wff^1 (unary formula) of L_{sem}, with "x" as free variable. Then

(26) $\phi(Sub(x, q(x)))$

means that the result of substituting a name of x for "x" in x (i.e., x's self-application) is ϕ. For example, if we take ϕ to be "Wx" again, then "$W(Sub(x, q(x))$" is itself a unary predicate which says that the self-application of its argument is white. For an appropriate choice of ϕ, (26) is an L_{sem} formula. Since L_{sem} is syntactically closed, there is a name for (26) in L_{sem}. Let "$a(\phi)$" abbreviate that name. Now since (26) is a wff^1, we can apply it to itself, i.e. substitute "$a(\phi)$" for the variable in (26), to get:

(27) $Sub(a(\phi), q(a(\phi)))$.

For an appropriate choice of ϕ, (27) denotes in L_{sem} a string, of which we can then predicate ϕ. This is the diagonalization of ϕ:

(28) $\phi(Sub\big(a(\phi), q(a(\phi))\big))$,

for which we give an abbreviation in L_{sem}:

(29) $d(\phi) \leftrightarrow \phi(Sub\big(a(\phi), q(a(\phi))\big))$.

For an appropriate choice of ϕ, the diagonalization in (28) has a name in L_{sem}, which we abbreviate $q(d(\phi))$. Note that (by the definitions above of $Sub(x, y)$ and $q(d(\phi))$):

(30) $Sub\big(a(\phi), q(a(\phi))\big) = q(d(\phi))$.

Consequently (by (29) and substitution of identicals), we have:

(31) $d(\phi) \leftrightarrow \phi(q(d(\phi))$.

This is the well-known *fixed-point* or *diagonal lemma*. It is a lemma schema – we get one for every unary formula ϕ.[69]

To get the paradox, we take L_{sem} formula "$\neg True(x)$" and put it for "ϕ" in (29) to get its diagonalization $d\big(\neg True(x)\big)$. This is a sentence of L_{sem} (the liar sentence). We show that this sentence is interderivable with its negation. For assume:

[69] The essence of the diagonalization schema is of course Gödel's procedure in (1931) for the case of provability. The generalization to a schema applicable to any predicate is in Carnap (1934, §35).
I have relied on Sereny (2006) for my formulation. The proof that the diagonal schema does what I say it does is given in any competent logic text book (e.g. in Mendelson (1997, p. 203)). Notice that on my presentation we get a sentence that genuinely predicates ϕ of itself (of its string), and not just a sentence provably equivalent to the predication of ϕ of it.

(32) $d(\neg True(x))$.

From (29) and (30), with substitution of identicals and equivalents, we have:

(33) $\neg True(q(d(\neg True(x))))$.

By the disquotational T-schema (A), (33) yields:

(34) $\neg d(\neg True(x))$.

This completes one direction. For the other, see that (34), together with (29) and double negation elimination get us:

(35) $True(q(d(\neg True(x))))$.

And using the T-schema again we go from this back to (32). Therefore (32) and (34) are interderivable contradictories, which with the law of excluded middle lands us in outright contradiction.[70]

⊠

The only premise used in these derivations beyond classical logic and facts about phonological closure was the T-sentence generated by the liar sentence. The availability of this T-sentence was a consequence of the assumption that the predicate "$True(x)$" in L_{sem} conformed to the naïve conception of truth. It follows that no predicate can conform to this conception; the requirement it puts on definitions of truth is inconsistent. This concludes the refutation of the naïve conception.[71]

2.3.2. The significance of the paradox

How should we interpret this result? The first thought is that the concept of truth turns out to be an incoherent concept. It seems to me that an incoherent concept is not strictly speaking a concept at all, except in name.[72] This is because concepts are themselves the vehicles and building blocks of intelligibility and coherence. In order to say what an incoherent concept is, we would have to have a

[70] We would get this result with a principle weaker than the T-schema, e.g. $True(\ulcorner\phi\urcorner) \Leftrightarrow \phi$ (*T-interderivability*). Consequently, any predicate governed by a principle at least as strong as T-interderivability, for example a necessity predicate, would entail a contradiction when diagonalized in this way. For the necessity predicate this is essentially Montague's (1963) result.

[71] Most purported solutions to the semantic paradoxes attempt to preserve the naïve conception at the cost of weakening classical logic so that the contradiction fails to follow logically from the assumptions. This has led to a wealth of insight about the nature of logic, but has not yet borne a satisfactory solution to the paradoxes. See Murzi and Carrara (2015) for a comprehensive survey of the effort of logical revision in the face of the semantic paradoxes, with a dismal (though tentative) conclusion about the prospects of this project.

[72] In the sense of Aristotle:"a dead man is a man only in name" (389b31-32).

theory of the constitution of concepts, of the way in which a concept is built up from more basic elements. Then we could show how some combinations of these elements fail to add up to a proper concept, and these combinations would be incoherent concepts.

I don't know what the elements of concepts are. In this work I am abstaining from discussing concepts directly (see §1.1). My proxy for a concept is the definition, which I consider to be a declarative sentence, in other words a theory. To each concept there corresponds a theory, and the incoherence of a concept can be explicated by a property of its corresponding theory, namely inconsistency. Now it seems to me that, strictly speaking, an inconsistent theory is no more a theory than an incoherent concept is a concept, and for the corresponding reasons. A theory tells me something about the world, and an inconsistent theory tells me nothing. So in order to understand what an inconsistent theory is we need an account of the constitution of theories from more basic elements, which would show how some combinations of these elements do not successfully add up to a theory.

As it happens, we have such an account. The elements of theories are sentences in a language, and inconsistency is defined as entailment of a contradiction, a sentence of the form $\ulcorner \alpha \wedge \neg \alpha \urcorner$ for some sentence α. The fundamental notion that underlies theories and concepts is therefore that of language; and a language can express contradictions without thereby ceasing to be a language except in name. It is in terms of (sentences in a) language that we can explicate failed theories and concepts. The reasoning of the previous section shows that any theory that conforms to the naïve conception entails a contradictory sentence, and is therefore inconsistent. And this means that the naïve conception defines an incoherent concept.

But the plot thickens. In §1.3.4E (truth-sufficiency) we assumed that a semantic theory for a language must be such as to entail a truth theory. If this is so, then the inconsistency of a truth predicate for L_{sem} implies the inconsistency of a semantic theory for it. This is the real consequence of the liar paradox. Unlike a mere contradictory sentence, which effectively destroys a theory that entails it but is unproblematic for the language in which it is expressed, a paradoxical sentence entails a contradiction in the semantic theory of that very language. A semantic theory for a language like L_{sem} is simply not possible. Semantically closed languages are inconsistent. Now if we accept the assumptions of §2.1.2 concerning ordinary language, this means that ordinary language is inconsistent. This is how one usually interprets Tarski's result in §1 of CTFL.[73]

Some thinkers have embraced this conclusion and advanced approaches that accommodate it or capitalize on it.[74] The reasoning seems to be that, although we have strong intuitions about the semantic

[73] E.g. in Soames (1999, p.49ff).

[74] Most notably Priest (1984). Eklund (2002), Patterson (2009), and Scharp (2013) also seem to offer an inconsistent language approach, but they think of language more in terms of a cognitive faculty. Their notion of

closure of ordinary language, we don't have any regarding its consistency. We should therefore learn to stop worrying and love the bomb. The insistence on consistency is a mere logician's hang-up that we are better off without: ordinary language is happily inconsistent.[75] The problem with this strategy is that since (interpreted) languages are individuated by their semantics (see §1.3.4), a language for which no semantic theory can be given is not strictly speaking a language at all, except in name. An inconsistent language is not a language in which inconsistencies can be expressed, but a language which cannot consistently be given. A more precise way to express the upshot of the paradox is to say that semantically closed languages, and with them universal and ordinary languages, simply do not exist. This is, after all, what we end up saying about proper classes relative to ZF set theory.[76]

As a final remark, notice that this makes the concept of truth different from other concepts. Whereas in order to explain the failure of ordinary concepts we could rely on the notion of language, the failure of the concept of truth makes the very notion of language impossible. In this sense the concept of truth is more fundamental than other concepts, and its breakdown a graver predicament. This I take to be the main import of §1 of Tarski *CTFL*.[77]

2.3.3. The generalization problem again

It may be that we will find the precise tweak of logic that will solve the paradox. This is difficult for me to imagine, but it is hard, as the Danes say, to prophecy, especially about the future. Regardless of

---

language does not satisfy the truth-sufficiency requirement from §1.3.4E. As a response to the paradox, such an approach amounts to changing the subject. (The same cannot be said about Priest, at least not easily.)

[75] Priest (1984, p.128), quotes Wittgenstein (1953) to the effect that the classical logician lives in "superstitious fear and awe of contradiction".

[76] This is more or less Herzberger's point in (1966) and (1967). Priest (1984, p. 120), has a rejoinder with which I will not deal here.

[77] This is a good place to point out a wide-spread misreading of Tarski, which has been hinted at earlier. In the literature on truth and on Tarski one constantly encounters statements such as "Tarski said… that English is 'inconsistent'" (Patterson 2009, p.388; Heck 1997, 545fn), and variations thereon. At least in *CTFL* and in (1944) I have found no statement by Tarski to that effect. In *CTFL* he says that everyday language (or "colloquial language", "Umgangssprache", "język potoczny") is inconsistent, but as I've claimed in §2.1.2, this should be understood as referring to unregimented discourse, not to such languages as English or Polish. In (1944) Tarski refines his position towards ordinary language, and in fact uses two terms: "everyday language" (e.g. p.349) and "natural language" (e.g. p. 347). Given Tarski's well-known meticulousness with words, I assume that he is making a conceptual distinction corresponding to the terminological one, though it is admittedly a subtle one which he never elaborates. About natural language he says that it is the object of empirical research (p.365). Maybe this is what I called (§2.1.2) "world-languages", like English (though see also footnote 51).
Now in (1944), unlike in *CTFL*, there is no direct argument that either everyday language or natural language is inconsistent. In both cases the claim is that their structure is not exactly specified, and therefore that "the problem of [their] consistency has no exact meaning" (p.349). So I think we should be wary of attributing to Tarski the claim that, say, English is inconsistent. More likely, his position is that the question simply doesn't arise for English.
You may object that great advances have been made in recent decades in linguistic semantics (based on Tarski's work) and that now an exact specification of the structure of natural language is at least conceivable. But nothing in these exact specifications (as implicit, e.g., in semantics textbooks such as Heim and Kratzer (1998)) hints that natural language is semantically closed. These themes are taken up again briefly in the Coda.

the paradox, there is a difference between thinking of the naïve conception as an adequacy criterion on definitions and, like the deflationists, thinking of it as a theory. On the deflationist or disquotationalists view of truth (see §2.1.3), the content of the concept of truth is exhausted by the collection of T-sentences, or the theory that has it as an axiomatic basis, and the function of truth is to allow us indirect assertion, for example when certain sentences are unknown to us or if there are too many of them to assert directly. The collection of T-sentences is given by the T-schema:

(36) $True(\ulcorner\phi\urcorner) \leftrightarrow \phi.$

This is not a formally correct way to give a collection of sentences. For example, if the language under discussion is L_{Ws}, then the expressions "ϕ" and "$\ulcorner\phi\urcorner$" are not part of it. They are place holders for sentences and sentence names, and we get a T-sentence by "inserting" a sentence and its name into the place-holders. I scare-quoted "inserting" because we don't actually *do* anything. If x is a sentence, let $Tsent(x)$ denote its disquotational T-sentence. The insertion metaphor is just an intuitive way to say that we assert all of the instances that we would get if we *would have* inserted etc. A formally correct formulation of the collection of T-sentences would be a collective assertion of them. But this is, on the deflationist view, precisely the case for which we need the truth predicate to begin with. The real form of the T-schema would be something like this:

(37) $\forall x(wff^0(x) \rightarrow True\big(Tsent(x)\big))$

where x ranges over expressions of L_{sem}. But then the disquotationalist view of truth is circular.

This is not just an unfortunate coincidence. We resorted to a schema because of the generalization problem (§2.1.1). if it was possible to define the schema using more innocent means, we would have resorted to those means instead. What the truth predicate provides is *theoretical unity*, the possibility to assert many sentences in one. Although every T-sentence is expressible without a working truth-predicate, expressing their totality requires a truth predicate. But there are other ways to get theoretical unity beside through a truth predicate. For consider the naïve conception of truth, not on its deflationist interpretation. If the naïve conception is understood, not as a theory, but as a criterion for the material adequacy of a truth definition, then we don't need to *assert* every T-sentence in order to express it. It is enough to assert the relation of logical consequence between a (variable) definition and the set of T-sentences. This is *conditional theoretical unity* and it relies, not on the concept of truth, but on the concept of logical consequence. We will see below (§4.2) that this makes a big difference.[78]

[78] See Gupta (2011b) for a similar point.

3 Stratified Truth

The naïve conception of truth is untenable because it implies commitment to the existence of semantically closed languages, and such languages do not exist. If we want to make progress in the elucidation of the concept of truth, we need to look for a characterization which preserves as much as possible of the content of the naïve conception, but which doesn't presuppose semantic closure. We distinguish between the language *L for which* truth is to be defined (the *object-language*) and the language *M in which* the definition is formulated (the *metalanguage*). We assume nothing at the outset concerning their identity or non-identity. This is the characteristic feature of *the stratified conception of truth.*

The central task of this chapter is to show that a definition of truth is possible on the stratified conception, and therefore that the stratified conception, unlike the naïve one, is a viable philosophical account of truth. In §2.1.2 the hope was expressed that the relativity of truth to language could be circumvented. The failure of the naïve conception belied that hope. Truth will have to be defined anew for every language. This makes it pressing to have a material adequacy criterion to replace the naïve conception. This criterion will be put forth in §3.1. §3.2 we will construct a formally correct definition of truth for our toy language L_{WS} from Chapter 2, and discuss the conditions for the possibility of such a definition. §3.3 will raise some problems for the stratified conception, the resolution of which will require the whole of Part Two.

3.1. Convention T

The undoing of the naïve conception was due to the commitment it implies to the existence of semantically closed languages. This commitment is engendered by the disquotational T-schema:

(1) $True(\ulcorner\phi\urcorner) \leftrightarrow \phi$.

Since the sentence mentioned on the left-hand side (the *object sentence*) is required to be identical to the sentence used on the right-hand side (the *truth condition*), and since identical sentences belong to the same language (see §1.3.5), this forces the language of the truth predicate to be the same as the language of the object sentence. If we want to avoid this, we need to characterize T-sentences without requiring the identity of the object sentence with the truth condition. What suggested this requirement in the first place was the observation in §2.1.1 that a sentence expresses its own truth conditions. The observation is no doubt sound in itself, but in this case misleading, since a sentence s is not the only sentence to express s's truth conditions. What we need is to generalize (1) so that the truth condition does not have to be identical to the object sentence, just a synonym or a translation of it. The following

criterion, Tarski's famous *Convention T*, generalizes the naïve conception by dropping the requirement of semantic closure:

(A) Convention T:

A definition of the predicate "$True(x)$" in a language M is an adequate truth definition for a language L if it entails all instances of the *translational T-schema*:

The Translational T-schema: $True(\alpha) \leftrightarrow \phi$,

where "α" is replaced by the name of an L-sentence and "ϕ" by the translation of that sentence into M.

Convention T is not yet a definition, but there are some conclusions about truth that we can draw already at this point. If M defines a truth predicate for L we say that *M mentions L*, in symbols $M \succ L$. Clearly for M to mention L, M must be able to express a T-sentence for every sentence of L. It must therefore contain a translation for every sentence of L. We then know the following:

(B) Hierarchy in the abstract:

(a) $If\ M \succ L\ then\ M \neq L$ (irreflexivity),

(b) $If\ M \succ L\ then\ L \not\succ M$ (asymmetry),

(c) $If\ M \succ L\ and\ \overline{M} \succ M\ then\ \overline{M} \succ L$ (transitivity).

Clause (a) can be proved by a reasoning parallel to that of the liar paradox in §2.3.1 (except that now there is no paradox, see below). Clause (c) follows from the fact that $\overline{M}$ has a translation for every M sentence, and in particular for M's truth predicate for L. Clause (b) is a consequence of (a) and (c). Mentioning is thus a partial ordering. This is the (in)famous Tarskian hierarchy of metalanguages.

Two immediate corollaries of (B) stand out:

(B) Corollaries:

(d) There is no semantically closed language,

(e) There is no universal metalanguage.

Clause (d) is the contrapositive of (a) above, and is known as Tarski's Indefinability Theorem. It is important to see the difference between this result and the failure of semantic closure in the case of the naïve conception. The naïve conception of truth *presupposes* the semantic closure of the languages it discusses, in the sense that its very formulation doesn't make sense for languages which are not semantically closed. That's why the proof of a contradiction causes a total breakdown in the naïve truth concept. At the end of §2 we were left in *aporia*, with no way to talk about truth at all. In the parallel

case of (B)(a), the derivation of a contradiction serves merely to exclude the special limiting case in which $M = L$. Convention T itself is still usable and, I think, correct. The paradox is not only defused, it is put to work.

By a "universal metalanguage" U in clause (e) I mean a language that mentions all languages. Assume such a language exists. Then it must mention itself, contrary to (d). Or we might say it is universal in a weaker sense, that it mentions all languages beside itself. But then, if U exists, then there is a semantic theory for it, which will have to be formulated in a metalanguage $\overline{U}$. Since mentioning is a partial ordering, this will engender a new hierarchy of languages that mention U, and U can't be said to be universal even in the weaker sense. Therefore there is no universal metalanguage.

The results in (B) are said to be "in the abstract" because they follow from Convention T independently of any particular truth definition. Now if $M \succ L$, then this must be related somehow to the contents, the expressive powers (ideology and ontology), of the two languages. M must have some kind of expressive advantage over L. This is what Tarski calls *essential richness*.[79] But speaking abstractly we have no insight into the nature of this expressive advantage, and so have no way to interpret it philosophically. It is only once we have a particular definition in front of us that we can say what the stratification of truth consists in. Our task in what follows is to observe one way of defining truth (essentially Tarski's) in order to see what expressive resources are involved. We can phrase this as a guiding question for this chapter:

(C) **Essential Richness Question:** If $M \succ L$, what expressive advantage does M have over L?

3.2. Defining truth

3.2.1. Expressing the T-sentences: a puzzle

More fully phrased, (C) asks what expressive resources a metalanguage M must possess in order to express a definition of truth for a language L that conforms to Convention T. According to Convention T, a definition is adequate if it entails all of the T-sentences for L. First of all, then, M must be able to express the T-sentences. A T-sentence contains nothing more than the name and the translation of its object sentence.

Here we face a certain embarrassment. Truth is relevant for interpreted sentences, not for mere phonological strings. On our framework an interpreted sentence is a string together with the semantic theory for the language in which it is to be interpreted. But a semantic theory already entails a truth

[79] Tarski (1944, §10).

theory for the language (§1.3.4E). If M refers to L sentences *as interpreted*, then by hypothesis we are already in possession of a truth predicate, and there is no need to define anything. In case of a truth-conditional semantic theory, interpretation is defined in terms of truth, and we have a flat-out circularity, or a *diallele* (See §1.3.4). Tarski is in the same predicament: in his informal discussion he stresses that truth applies to interpreted sentences (*CTFL*, p. 166), and he relies on the notion of translation in his formulation of Convention T (as I have above), but he leaves these notions unexplicated. It was Davidson's insight in (1967), and less explicitly, Montague's (1970),[80] that the dependence of Tarskian truth on interpretation can be turned on its head: if truth is taken as primitive then the definition can be used to elucidate meaning. However, we can't define truth in terms of interpretation and vice versa at the same time.

Our procedure will be the following. We do not suppose that a semantic theory for L is already present in M. Instead, we let the intended truth predicate in M apply to uninterpreted sentence strings. For these strings we construct a single definition, which appeals neither to truth nor to meaning, and which will serve at the same time both as a definition of truth and an interpretation. In this way truth and meaning grow together like two flowers on a single stem, or, if you prefer a different figure, are like the duck and the rabbit of the ambiguous image. If you dislike figures altogether, wait until §7.5, where the issue of the diallele will receive a happy solution in straight terms.

With this much in the way of alleviating remarks, let us resume the thread of investigation. M, we said, must be able to express the T-sentences, and a T-sentence contains nothing more than the name and the translation of its object sentence. M must therefore be able to name all sentence strings and to express translations of all L sentences. The first requirement is taken care of if we let M have full phonological structural-descriptive naming capacities, i.e. to include L_{SD} (see §2.2.4). In order for M to have translations of all L sentences, M's ontology must match or exceed L's ontology (i.e. $D^L \subseteq D^M$), and M's ideology must match or exceed L's ideology (i.e. for every predicate of L there is a formula of M with the same extension). In short, in order to express every T-sentence for L, a language M must (a) include phonology and (b) match or exceed L in expressive power.

We have already encountered a language that fulfills these conditions: the syntactic closure $SynC(L)$ (§2.2.4). We can easily see that $SynC(L)$ does not only fulfill these conditions for L, but also for itself. However, if these conditions sufficed for defining truth, then $SynC(L)$ would be semantically closed, which is impossible. Therefore, it is not sufficient, in order to define truth for L, to contain phonology and to match L in expressive power. We are faced with the following puzzle. On the one hand:

[80] But already present in Dummett (1959). See also Etchemendy (1988, §1.2).

(a) For M to adequately define truth for L, it is necessary and sufficient that it express a sentence that entails every T-sentence for L. (by Convention T)

On the other:

(b) For M to adequately define truth for L, it is (necessary but) not sufficient that it express every T-sentence for L. (since $SynC(L)$ is not semantically closed)

The difference between the sufficient condition stated in (a) and the insufficient condition stated in (b) is that (a) requires there to be a definition, a single sentence from which all T-sentences follow. It is one thing to be able to express every sentence of a collection individually; quite another to express the whole collection in a single sentence. The T-sentences for L exhaust, in a sense, the content of the concept of truth for L. But something more is needed in order to unify that content into a single concept. That extra something, which above in §2.3.3 we called theoretical unity, is what systematically eludes a language with regard to its own truth predicate. Our guiding question for this section can be refined accordingly:

(C') Essential Richness Question (refined): If $M \succ L$, what expressive advantage does M have over $SynC(L)$?

3.2.2. The inductive definition

If we had finitely many T-sentences, we would be able to conjoin them into a single sentence which would entail each of them. But there are as many T-sentences as sentences of L, which is to say infinitely many, and there are no infinite conjunctions.[81] The problem of truth turns out to be another one of the problems with the idea of the infinite. It will be instructive to take a look at how concepts that apply to the most basic of infinite systems, the system of natural numbers, are handled formally. The relevant devices are those of an inductive definition, a recursive definition, and a proof by induction. I'll sketch some trivial examples with numbers to serve as a roadmap for the case of truth.

The natural numbers are defined inductively:

(2) Inductive definition of number:

(a) Zero is a *number* (*base clause*),

(b) For all x, if x is a number, then x's successor is a *number* (*step clause*),

(c) Nothing else is a *number* (*extremal clause*).

[81] More precisely, we would need a definition of the form "$True(x) \leftrightarrow (x = \ulcorner\phi\urcorner \wedge \phi) \vee (x = \ulcorner\psi\urcorner \wedge \psi) \vee$ .... See *CTFL* p.188.

We assume that *zero* and *successor* are previously understood, and written "0" and "$s(n)$", respectively. The idea behind such a definition is that we start with zero as root and generate the numbers successively and indefinitely using the successor function. The successor function is understood to produce different successors for different numbers, and zero is understood to succeed no number. The extremal clause limits the numbers to the objects that were generated through this procedure.

A recursive definition of a numerical concept *P* states the conditions for applying *P* to an arbitrary number n in terms of the generation sequence that leads from 0 to n; more precisely, in terms of a condition on the predecessor of n, or directly in the case of 0. The operator of addition, for example, or the relation of being a sum, is defined recursively as follows:

(3) Recursive definition of addition:

(a) $x + 0 = x$ (base clause)

(b) $x + s(y) = s(x + y)$ (step clause)

Put into a single sentence, this definition becomes:

(4) $x + y = z \leftrightarrow (y = 0 \wedge x = z) \vee \exists v(y = s(v) \wedge z = s(x + v))$.

Of course, some concepts are definable explicitly on the basis of concepts defined recursively, for example:

(5) $Even(x) \leftrightarrow \exists y(y + y = x)$.

Such concepts are still to be considered recursive. Inductive and recursive definitions justify proofs by induction. In this way assertions about infinite collections can be proved using finite means. For example, we can prove that the definition (5) has the following desired consequence:

(6) $\forall x(Even(x) \leftrightarrow \neg Even(s(x)))$.

This is the model according to which we will define truth. The inductive basis, parallel to (2) above, is the collection of strings (see §1.3.1). Truth is a concept that applies to sentence strings. The sentences were defined explicitly, in analogy with (5) above, in terms of the concept of wff, which was defined recursively like (3) above. In the same way we will define truth, the semantic concept for sentences, explicitly in terms of the semantic concept for wffs, satisfaction. The relation of satisfaction will be defined recursively. This will allow us to derive T-sentences for all *L* sentences, in analogy with (6).

What is satisfaction? Intuitively, a wff^n (a formula with n free variables) α expresses a relation R_α which holds (or fails to hold) between n objects, as assigned to "roles" or "places" in the relation. The free variables $x_1, \dots, x_n$ in the formula represent the roles. Satisfaction is then a relation that holds or fails to hold between a wff^n and an assignment (mapping) of objects to the variables $x_1, \dots, x_n$. An

assignment σ satisfies a $wff\ \alpha$ if and only if the objects assigned to the roles designated by the variables stand in the relation R_α. I will sometimes speak as if it is the objects (or sequences of them), and not the mappings, that satisfy formulas. Where α is a wff^1, R_α is not strictly speaking a relation but a unary concept. Then I will say that it is the object (assigned to the free variable) that satisfies or fails to satisfy the formula. If α is a wff^0, a sentence, then again R_α is not properly speaking a relation. In this case it makes no sense to speak of satisfaction. However, for uniformity, we will say that it is either satisfied by all or by no variable assignments.

We will formally define the binary predicate $sat(x, \sigma)$ for the language L_{Ws}, where x is a formula and σ an assignment function. For readability I use lowercase Greek letters "σ" and "τ" for variables that range over assignment functions, but they are to be read as standard first-order variables. We let the metalanguage M possess structural-descriptive resources and homophonic translations of L_{Ws}'s non-logical vocabulary. Thus "s" in M refers to snow, and "W" denotes whiteness, just like in L. We need some auxiliary notions (in M):

(7) $\sigma(x)$ = the object that assignment σ assigns to variable x. We decree that $\sigma(\underline{s}) = snow$, for any σ.
(8) $\sigma^{x \to y}$ = the assignment that is like σ except (maybe) in assigning y to the variable x,
(9) $ST(x)$ if and only if x is a variable or $x =$ "s" (x is a singular term).
(10) $D^L(x)$ if and only if x is in the domain of L.

The base clause of the recursive definition gives satisfaction conditions for atomic formulas. In L_{Ws} we have two predicates: the logical ⌜$\alpha = \beta$⌝ and the non-logical ⌜$W\alpha$⌝; and two classes of singular terms: variables and "s". Correspondingly, we have two base clauses:

(D) Recursive definition of satisfaction (base clauses)

(a) $base_{=}(x, \sigma) \leftrightarrow \exists z(ST(z) \wedge (x = \underline{W} \cdot z \wedge W(\sigma(z))))$,
(b) $base_W(x, \sigma) \leftrightarrow \exists z \exists w(ST(z) \wedge ST(w) \wedge (x = z \cdot \underline{=} \cdot w \wedge \sigma(z) = \sigma(w)))$.

The step clauses correspond to the two ways of composing formulas: truth-functional connection ($\phi \downarrow \psi$) and quantification ($\forall \alpha \phi$):

(D) Recursive definition of satisfaction (step clauses)

(c) $step_{\downarrow}(x, \sigma) \leftrightarrow \exists z, w(wff(z) \wedge wff(w) \wedge x = z \cdot \underline{\downarrow} \cdot w \wedge (sat(z, \sigma) \downarrow sat(w, \sigma)))$,
(d) $step_{\forall}(x, \sigma) \leftrightarrow \exists z, w(wff(z) \wedge var(w) \wedge x = \underline{\forall} \cdot w \cdot z \wedge \forall v(D^L(v) \to sat(z, \sigma^{w \to v})))$.

Given a formula x and a variable assignment σ, we can calculate whether $sat(x, \sigma)$ by going backwards through x's construction tree up to the atomic formulas. Seeing the four clauses as individually sufficient and disjointly necessary, we can put the recursive definition in the form of a proper definition:

(D') Recursive definition of satisfaction (full)

(e) $sat(x, \sigma) \leftrightarrow base_{=}(x, \sigma) \vee base_{W}(x, \sigma) \vee step_{\downarrow}(x, \sigma) \vee step_{\forall}(x, \sigma)$.

The disjunction in this definition is exclusive. It is a phonological fact that (for any assignment σ) a string x can fall under at most one of the disjuncts.[82] Defining truth for sentences is now straightforward:

(E) Definition of truth for L_{Ws}

$True(x) \leftrightarrow wff^{0}(x) \wedge \forall\sigma(sat(x, \sigma))$.

From this definition all T-sentences follow. Let's demonstrate for the object sentence "$\exists x_1 \exists x_2 (Wx_1 \wedge \neg(x_1 = x_2))$" ("there exist two things, one of which is white"). The following sentences are all equivalent:

(a) $True(\underline{\exists x_1 \exists x_2 (Wx_1 \wedge \neg(x_1 = x_2))})$

(b) $\forall\sigma(sat(\underline{\exists x_1 \exists x_2 (Wx_1 \wedge \neg(x_1 = x_2))}, \sigma))$ by (E)

(c) $\forall\sigma \exists z(D^{L}(z) \wedge sat(\sigma^{1\rightarrow z}, \underline{\exists x_2 (Wx_1 \wedge \neg(x_1 = x_2))}))$ by (D)(d)

(d) $\forall\sigma \exists z \exists w(D^{L}(z) \wedge D^{L}(w) \wedge sat(\sigma^{1\rightarrow z, 2\rightarrow w}, \underline{Wx_1 \wedge \neg(x_1 = x_2)}))$ by (D)(d)

(e) $\forall\sigma \exists z \exists w(D^{L}(z) \wedge D^{L}(w) \wedge sat(\sigma^{1\rightarrow z, 2\rightarrow w}, Wx_1) \wedge \neg sat(\sigma^{1\rightarrow z, 2\rightarrow w}, \underline{\neg(x_1 = x_2)}))$ by (D)(c)

(f) $\forall\sigma \exists z \exists w(D^{L}(z) \wedge D^{L}(w) \wedge Wz \wedge \neg(z = w))$ by (D)(a),(b)

(g) $\exists z \exists w(D^{L}(z) \wedge D^{L}(w) \wedge Wz \wedge \neg(z = w))$ since $\forall\sigma$ is vacuous

Therefore, we have derived:

(11) $True\left(\underline{\exists x_1 \exists x_2 (Wx_1 \wedge \neg x_1 = x_2)}\right) \leftrightarrow \exists z \exists w(D^{L}(z) \wedge D^{L}(w) \wedge Wz \wedge \neg(z = w))$

$\boxtimes$

Inspect it to see that it is really the T-sentence generated by our object sentence. The sentence on the right-hand side of the biconditional in (11) is an (almost homophonic) translation of the sentence mentioned on the left-hand side (it is not completely homophonic because of the domain restriction).

[82] This fact is crucial for deriving the T-sentences.

By induction we can show that for every L sentence the appropriate T-sentence can be derived. The inductive definition seems therefore to conform to Convention T.

3.2.3. A clue

If (E) is a materially adequate truth definition for L in M, then we should be able to answer the essential richness question (C): what expressive resources, beyond the ontology and ideology of $SynC(L)$, are required for the definition of truth for L? By inspection of definitions (D) and (E) we find that the only item not in $SynC(L)$ is the ontology of variable assignments. On the face of it this seems to constitute a genuine advantage in expressive power. Variable assignments are mappings from the countable set of variables to the domain D^L. The cardinality of the collection of assignments is therefore $|D^L|^{\aleph_0}$, which is often greater than the cardinality of D^L. For example, if D^L is denumerable, then the collection of assignments will have the size of the continuum. This is a significant difference, and suggests a concrete answer to (C): the advantage of M over $SynC(L)$ that allows it to define truth is the ontology of variable assignments.

This answer is wrong. The assignment functions I used in the definition of satisfaction were had the entire infinite set of variables as their domain, but only a finite number of variables occurs in every formula. It is possible to define satisfaction, a little less elegantly perhaps, with finite variable assignments or finite sequences of objects.[83] Unlike the set of all total variable assignments, the set of finite sequences over D^L is of the same size as D^L if D^L is infinite, and can be coded in it. Since a syntactically closed language will always have an infinite domain (since it contains phonology), it will always be possible to code finite sequences in it. There is therefore nothing in the recursive definition of satisfaction that requires more expressive resources than those present in $SynC(L)$. We are back at the puzzle of §3.2.1: what, in addition to the expressive resources of $SynC(L)$ is required for defining truth for L?

The answer lies in the fact that inductive and recursive definitions are not, strictly speaking, definitions at all. An inductive definition works by generating the items one by one, and a recursive definition works by tracing the items up their generation history. But this talk of generation and of tracing up a history is no more than metaphor.[84] On our framework a definition is a statement of application conditions, a statement that appeals to concepts, objects, their interaction in the form of predication, and the logical composition of judgments using connectives and quantifiers. Temporal or causal notions such as "tracing up a generation history" are not part of it. If we do formulate a recursive definition in the grammatical form of a statement of application conditions (as we did for addition in (4) and for

[83] See *CTFL*, p. 195fn; also Popper (1955).
[84] See §1.3.1, footnote 21 for genetic inductive definitions.

satisfaction in (D)(e)), then this statement does not constitute a formally correct definition, since the step clauses contain the definiendum and thus render the definition circular. On the one hand, the temporal metaphor that accompanies an inductive or a recursive definition does not meet our standards of clear and explicit discourse, making the definition formally incorrect. On the other hand, since the resources used in the inductive definition of truth don't exceed those of $SynC(L)$, and since we know that truth can't be defined in $SynC(L)$, the temporal metaphor is doing some work. We can't just get rid of it. It neither discloses nor conceals the answer to our essential richness question; it gives us a sign.

3.2.4. The explicit definition

What we need is to convert the materially adequate but formally incorrect recursive definition into an explicit definition. In §1.3.3 (footnote 36) I mentioned a method by Quine to transform recursive definitions into explicit ones under certain conditions. The condition is that we can represent the metaphorical generation sequence in a (finite) string. Unfortunately, although we could apply this method to the syntactic concept wff, it won't work for the corresponding semantic concept of satisfaction. The reason is that whether an assignment function σ satisfies a universal formula $\ulcorner \forall x \phi \urcorner$ depends on whether $\sigma^{x \to a}$ satisfies $\ulcorner \phi \urcorner$ for every $a \in D^L$, and D^L might be infinite. If we wanted to represent the "construction sequence" in a string, in the general case the string would have to be infinite, which is impossible. There is, however, another strategy for converting inductive or recursive definitions into explicit ones. This is the strategy suggested by Frege's attempt to redeem the concept of number from the constructive metaphor of inductive definition, and it is the one used by Tarski in *CTFL* (though he doesn't spell out the details). First, the idea is to view the base and step clause of the inductive definition of number in (2), not as describing a construction procedure, but as a statement to the effect that the concept of number includes zero and is closed under successor:

(12) $N0 \wedge \forall x(Nx \to Ns(x))$.

This is read as a statement about the concept N, expressing what we can call its *closure property*. The next step is to identify N with this property. By generalizing over N we make the closure property a condition on concepts. We then define N as the conjunction of all of these concepts. Formally:

(13) $Nx \leftrightarrow \forall F(F0 \wedge \forall y(Fy \to Fs(y)) \to Fx)$.

In words: an object falls under the concept of number if and only if it falls under every concept that has the closure property, i.e. includes zero and closed under successor.

We can apply this general strategy to recursive definitions,[85] for example to the case of addition ((3) above). For readability, let's first put forth the following abbreviations:

(14) $base_+(R) \leftrightarrow \forall v R(v, 0, v)$,

(15) $step_+(R) \leftrightarrow \forall v \forall u \forall w \left(R(v,u,w) \rightarrow \exists t\big(t = s(w) \wedge R(v, s(u), t)\big)\right)$.

The explicit definition of addition is then:

(16) $x + y = z \leftrightarrow \forall R(base_+(R) \wedge step_+(R) \rightarrow R(x,y,z))$.[86]

Explicit definitions defined in terms of recursive definitions (e.g., (5)) can stay as they are, except that we need to remember that they now implicitly involve quantification over predicate position. Definitions such as (13) and (16) clearly allow for proofs by induction.

We apply this procedure to the case of satisfaction. The idea is the same: transform the recursion clauses into a closure condition on relations, this time between variable assignments and formulas. The closure clauses are simple transformations of the induction clauses:

(17) Second-order closure clauses for satisfaction concepts:

(a) $base^2_W(R) \leftrightarrow \forall x \forall \sigma (\exists z (ST(z) \wedge \left(x = \underline{W} \cdot z \wedge W\big(\sigma(z)\big)\right)) \rightarrow R(x, \sigma))$,

(b) $base^2_{\underline{=}}(R) \leftrightarrow \forall x \forall \sigma (\exists z \exists w (ST(z) \wedge ST(w) \wedge \big(x = z \cdot \underline{=} \cdot w\big) \wedge \big(\sigma(z) = \sigma(w)\big)) \rightarrow R(x, \sigma))$,

(c) $step^2_{\downarrow}(R) \leftrightarrow \forall x \forall \sigma (\exists z \exists w (wff(z) \wedge wff(w) \wedge (x = z \cdot \underline{\downarrow} \cdot w) \wedge \big(R(z, \sigma) \downarrow R(w, \sigma)\big)) \rightarrow R(x, \sigma))$,

(d) $step^2_{\forall}(R) \leftrightarrow \forall x \forall \sigma (\exists z \exists w \left(wff(z) \wedge var(w) \wedge (x = \underline{\forall} \cdot w \cdot z) \wedge \forall v \big(D^L(v) \rightarrow R(z, \sigma^{w \rightarrow v})\big)\right) \rightarrow R(x, \sigma))$.

The explicit definition is, as in the case of addition, a conjunction of all concepts that fulfill the closure conditions:

(18) Explicit definition of satisfaction (second order):
$sat(x, \sigma) \leftrightarrow \forall R(base^2_{\underline{=}}(R) \wedge base^2_W(R) \wedge step^2_{\downarrow}(R) \wedge step^2_{\forall}(R) \rightarrow R(x, \sigma))$.

This definition is second-order, and as such not yet formally correct according to our regimentation scheme. We can improve on it by observing how it avoids the circularity of the recursive definition

[85] For the case of numbers and strings I don't think we should, see §1.3.1.
[86] Compare the formally incorrect (4).

(D)(e). That definition was circular since the definiendum sat appeared in the definiens, in the step clauses. In (18) we replace the definiens in the definiendum with a variable over relations, and quantify over it. In order for this definition to succeed, concepts have to be quantified over, i.e. they have to be objects in the second-order domain. [87]

But, importantly, in order for a definition of truth for L to succeed, M doesn't have to have *all* of its concepts as objects. We can get a formally correct definition, and in addition a more fine-grained understanding of what is required for a truth definition, by letting the extensions of the necessary concepts in the first-order domain of M. We need to revise the closure clauses in an obvious way: in (17) we change the second-order variable "R" into a first-order variable "r", and the predications $\ulcorner R(x,\sigma)\urcorner$ into set-membership statements $\ulcorner\langle x,\sigma\rangle \in r\urcorner$. This makes (18) into a first-order definition :

(F) Explicit definition of satisfaction (first-order):

$$sat(x,\sigma) \leftrightarrow \forall r(base^1_{=}(r) \wedge base^1_{W}(r) \wedge step^1_{\downarrow}(r) \wedge step^1_{\forall}(r) \rightarrow \langle x,\sigma\rangle \in r).$$

The derivation of the T-sentences from this definition is similar to the derivation from the recursive definition, and a proof by induction that all T-sentences are derivable goes through in the same way. The definition of truth in (E) can be kept if we understand the satisfaction predicate it employs to be defined as here. We therefore have a formally correct and materially adequate definition of truth for L_{Ws}.

3.2.5. Some remarks

The central task of this chapter, and of Part One, has been achieved: we have provided a formally correct and materially adequate definition of truth. The definition applied only to a particular language L_{Ws}, but it is clear how to apply it to other languages. I close this section with some comments.

With respect to the central question of this section, concerning the expressive resources necessary for defining truth, the first-order definition (F) has several advantages over the second-order one (18).[88] According to the second-order definition, in order to define truth a metalanguage has to have variables of a higher-order than the object-language. Essential richness is here a question of logic. This makes it incompatible with a first-order regimentation scheme, and in fact with any regimentation scheme of

[87] This is due to the fact that L is first-order, and therefore that variable assignments are also first-order objects. If L had been second-order, M would have to be third-order, etc.

[88] Tarski's own definition in the main text of *CTFL* is higher-order. In the postscript from 1936 he expresses a preference for first-order definitions. See Schiemer and Reck (2013) for a historical account of the general shift from type-theoretic to model-theoretic logic.

bounded order. By contrast, the first-order definition cashes essential richness out in ontological terms. The metalanguage is a language of logically the same kind as L, but it needs to have more objects in its domain (objects of a higher order). We can stay with a simple regimentation scheme that recognizes only concepts and objects.

The second advantage is that we can give a more fine-grained answer to the question of which new objects need to be countenanced. The domain of a second-order variable usually includes all sets over the first-order domain, or at least some natural collection of such sets, for example the definable sets. On the first-order definition we see that there is a single object which needs to be in the domain of M in order for the definition of truth to succeed: the extension of the satisfaction relation, $|sat|$. If $|sat| \in D^M$, then (and only then) the definition will yield the correct result. This fact is of philosophical significance. In §1.3.4 we identified a language with the extension of its satisfaction relation. The necessary and sufficient condition, then, for M to be able to express a truth definition for L according to the first-order definition, is that L be an object in the domain of M. The term *object-language*, on this way of defining truth, takes on a literal sense.

Another issue that bothered us was the interdependence of truth and meaning, the diallele. In §3.2.1, in order to sidestep this problem, we decided try to give a definition that relies neither on the concept of truth nor on the interpretation of sentences. This definition was then to be used both for truth and for interpretation. For the case of truth, it was shown that our definition is successful. It remains to show that it will work also as a semantic theory. We put forth two styles of meaning theory: truth-conditional and model-theoretic (see §1.3.4). The definition of satisfaction is of itself already a definition of truth-conditional semantics. Defining model-theoretic semantics is also straightforward, but requires more ontology. In general, we define the semantic value of a predicate to be the set of things that satisfy it:

(19) $|\phi| = y \leftrightarrow (x)(x \in y \leftrightarrow sat(\ulcorner \phi x \urcorner, x))$.[89]

The added ontological assumption is that such a set exists in D^M. For every concept definable in L we need to have a set in D^M. In other words, the requirement that model-theoretic semantics puts on the ontology of a metalanguage is greater than the minimal one of having L as an object in D^M. It is that every formula in L be an object in M. In other words, that M's ontology contain L's ideology.

[89] Here I speak as though it is objects and not assignment functions that stand in satisfaction relations with formulas. This definition is straightforwardly generalizable to n-ary predicates.

In the literature on truth after Tarski explicit definitions are often not required, or nor emphasized. Sometimes an axiomatic approach is adopted and sometimes recursive or inductive definitions are put forward without discussion of their nature or formal correctness. Maybe the motivation is a hope to thereby avoid stratification. But in §3.1 we saw that stratification of languages is forced upon us even before we contemplate a particular style of definition. Axiomatic treatments therefore neither avoid stratification nor explain it. By following through Tarski's own procedure of defining truth explicitly, we can go beyond the conclusions derived abstractly from Convention T, that the languages constitute a partial ordering. The explicit definition tells us precisely what the ordering consists in: mentioning a language means referring to it. The impossibility of semantic closure then becomes a case of the more general assumption of the well-foundedness of collections.

One might feel uneasy with my quick dismissal of inductive and recursive definitions as metaphors or heuristics in the service of explicit definitions.[90] After all, inductive definitions are very commonly used and certainly enjoy a firm set-theoretic grounding.[91] But it is important to recognize that on the set-theoretic accounts, inductive definitions are strictly speaking not definitions, but intuitive ways to present certain relations between sets that are defined explicitly: usually the set defined by the base clause of an inductive definition and the sets defined by the iterated application of some operation or set-theoretic function on the base set, until a fixed-point is reached. These operations should not be understood as actions (except metaphorically) any more than we say that, say, the union operation is an action. Kripke (1975) exploits the temporal metaphor of the inductive definition in order to align his definition of truth with some intuitions. But at the end of the day, his definition too is conceived as an explicit definition in a strictly stronger metalanguage.

3.3. Tarski's revenge paradox

The post-Tarskian literature on the liar paradox is plagued by the phenomenon of revenge paradoxes. The typical plot is this: a new kind of language L is defined so as to express a predicate "$T(x)$" which holds of all and only the true sentences of L. The liar paradox is avoided by weakening the logic in L, at least for certain "pathological" cases. The liar sentence is declared to be one of those cases, and the inference to contradiction fails. L, it seems, *pace* Tarski, is semantically closed. However, it is then shown that a paradoxical sentence can be formulated using the terms of the solution. This is the revenge paradox. Or, according to an alternative analysis, the problem is that the predicate "$T(x)$" only approximates the truth predicate for L. For example, on theories which block the inference to contradiction by rejecting bivalence the predicate "$T(x)$" is defined so as to have the same extension as

[90] See the footnotes on pp.175f, 177, 180, 182 in *CTFL* for comments on this issue.
[91] E.g. Moschovakis (1975), Aczel (1975).

the truth predicate for L. However, if bivalence is rejected, then the extension of a predicate is not enough. In order to faithfully express truth for L, "$T(x)$" would have to share both the extension and the anti-extension of truth for L, and this it cannot consistently do. If defined precisely, the proper truth predicate for the object-language on such solutions will generally be defined in a richer metalanguage, leaving us with "the ghost of" stratification.[92]

The lesson we should draw from the revenge phenomenon is that in order to assess a proposed account it is important to examine the actual definitions and the language they're given in, and not so much the behavior of the defined predicate if it is expressed in a different language. Not every theorist provides a clear statement of the metalanguage, so it is not always easy to assess a solution.[93] One gets the feeling, however, that the second lesson to draw from the revenge phenomenon is that stratification is unavoidable. It might therefore come as a surprise, a quite disturbing one, to find out that accepting stratification doesn't grant immunity from revenge; for the stratified conception too faces a number of serious difficulties, one of which is an instance of the revenge problem. This is disturbing because then there is no other conception to fall back on.[94] I will close Part One with a brief presentation of what I take to be the most serious problems. Part Two will be devoted to their solution.

The first problem with the stratified account is that it doesn't do what it set out to do. Our task was to provide an analysis of the concept of truth, and we stipulated that concept analysis takes the form of a definition of a truth predicate (see §1.1). In a sense, we have succeeded too much: we asked for one definition and were presented with a swarm of them. For each object-language-metalanguage pair a different definition is required. The impossibility of semantic closure and of a universal metalanguage makes it the case that this multiplicity and relativity to language cannot be reduced. In the absence of an account of why all of these definitions are definitions of the same concept, we can't say we have a philosophical analysis of the concept of truth. As Putnam complains, it seems that while Tarski's definitions give us the *extension* of the concept of truth, they don't capture its *sense*, i.e. the intuitive notion of truth.[95] This is therefore one problem facing a stratified account, which I will refer to as *the unity problem*: what unifies the various definitions into a single concept?

The unity problem seems to have an easy answer, to which Putnam's complaint points the way. Putnam wants the *sense* of the concept of truth. I don't know what senses of concepts are, but one popular way

---

[92] This sketch is inspired especially by Kripke (1975), but it applies to many other theories as well. See Beall (2007a) for various approaches to the revenge phenomenon and Beall's introductory essay for an overview.
[93] See Leitgeb's critique of Field's solution in Beall (2007).
[94] Except perhaps Priest's, with which I won't engage in this work.
[95] Putnam (1985, p. 64). Tarski himself presents his project as a search for a philosophical analysis of truth, see e.g. (1944, p. 341).

to think of the senses of predicates is as functions from circumstances of evaluation (possible worlds, times, etc.) to their extensions. These functions are called *intensions*, and they may be said to provide the unity of a predicate that has different extensions in different circumstances. For example, the extension of "president" at a time not so long ago has Obama as a member up, but it doesn't anymore. Yet it is the same predicate then and now. What defines it is therefore not its extension but its intension, the function that maps 2016 to a collection that includes Obama and 2017 to a collection that excludes him. Similarly, although on the stratified conception there is a different definition of truth for every language (and metalanguage), what makes them all into definitions of truth is their conformity to Convention T. This suggests that we think of Convention T as, in some sense, the intension of the concept of truth, and of the different languages as so many circumstances of evaluation. Convention T is the unifying element. This is *the intensional reply* to the unity objection.

This leads us to the second problem of the stratified conception. One of our central negative results is the impossibility of a universal metalanguage, a language that mentions all other languages. But then which language is Convention T formulated in, and the other results that applied across the board such as the indefinability theorem itself? This is *the effability problem*. It is a generalization of the revenge paradox, and it effectively undermines the intensional reply to the unity objection.[96]

Finally, we have the following classical problem. If the semantics of a language is given in its metalanguage, and the semantics of the metalanguage in a metametalanguage, etc., then we are never in a position to say what the interpretation of a language *really* is. Since without giving its interpretation we can't be said to have given a language, the upshot is that no language can ever be given. This is the *regress problem*.

I consider these problems to be of great importance. Their solution, correspondingly, will require drastic measures.

[96] See e.g. Black (1948, p.59), Priest (1984), McGee (1991, p.81) for statements of the effability problem. The tradition of the revenge paradox (or *strengthened liar*) actually begins with the attack on stratified theories, in particular Russellian type-theory, as early as Ushenko (1937). See Bennet (1967) for a review of an exchange in 1950s regarding Ushenko's paradox. (See also Rozeboom (1958) and the discussion culminating in Drange (1969).)

PART TWO

CHARACTERIZING TRUTH

Let me recapitulate briefly the dialectic of Part One. Our main task is to analyze the concept of truth, where an analysis was stipulated (§1.1) to take the form of a definition in a precisely defined, i.e. regimented, first-order language. Sentences in regimented languages were also chosen to play the role of truth-bearers, in virtue of the fact that they are defined precisely and in unproblematic terms (§1.2, §1.3). Our first candidate analysis was the naïve conception of truth (§2):

> **The naïve conception**: the concept of truth entails all disquotational T-sentences.

This analysis had the virtue of eliminating the relativity of truth to language by being consistent with the thesis that there is a single language in which everything can be expressed:

> **The strong effability thesis**: there is a language such that anything expressible is expressible in it.[97]

However, the naïve conception was shown to be untenable, and the strong effability principle false, by the liar paradox. Following Tarski, we replaced the naïve conception with a stratified version thereof, called Convention T:

> **Convention T**: A definition in a language M is an adequate truth definition for a language L if it entails all translational T-sentences.

That Convention T is a tenable condition on definitions was shown in §3 by giving a definition of truth that conformed to it. However, it had some negative results, like the refutation of the strong effability thesis. It was shown (§3.1) that a universal metalanguage U would be incoherent, and that the languages were ordered in an unbounded hierarchy. The effability principle can be replaced with a weaker version:

> **The weak effability thesis**: Anything expressible is expressible in some language or other.

This is, in broad outline, the stratified conception of truth as we find it in Tarski. But then, it was pointed out, the stratified conception does not fulfill the task that was set out for it, which was to find a *single* definition of the concept of truth. This was the unity objection. It was replied that, although the stratified conception is committed to an irreducible plurality of truth definitions, they all have to conform to the same Convention T. We can therefore think of Convention T as the intension of truth, and of the various definitions as its extensions. To this reply another objection was raised, to the effect that Convention T, if it is to apply to all definitions for all languages, would have to be formulated in the impossible universal metalanguage U. This was the *effability* objection.

Part Two will develop a solution to the effability objection by showing that Convention T can be formulated, with general applicability, in a language which is not the universal metalanguage U. In

---

[97] In §2.1.2 we called this "the universality thesis".

Chapter 4 we will formalize Convention T into a single first-order sentence. This will reveal its logical form and the expressive resources required for its formulation or, what comes to the same thing, its language. We will show that this language consists of the language of phonology L_{SD} (from §2.2) plus something more. That extra something, however, cannot be regimented in our standard regimentation scheme. Chapter 5 is a digression into the domain of pragmatics, and in particular into indexical theory, in search of the missing expressive resource, which I will call *abstract reference* or *abstract generality*. Chapter 6 will adapt this resource to our own framework and will provide a fully explicit formulation of Convention T. The concluding chapter 7 will spell out our solution to the effability objection and the unity objection, as well as a revision in how we understand concept analysis. Solutions to some other problems will also be given there.

4 The Language of Convention T

4.1. The logical form of Convention T

Our procedure will be to formalize Convention T according to first-order syntax and to ask what the (minimal) language is in which it can be formulated. I will refer to this language as Z. For convenience, let's repeat Convention T here:

> (A) **Convention T (informal):**
>
> A formally correct definition of the predicate "$True(x)$" in M is an adequate truth definition for L if it entails all instances of the *translational T-schema*:
>
> (The translational T-schema): $True(\alpha) \leftrightarrow \varphi$,
>
> where "α" is replaced by the name of an L-sentence and "φ" by the translation of that sentence into M.

4.1.1 The main connective

Superficially, Convention T looks like a conditional:

> A formally correct definition of the predicate "$True(x)$" in M is an adequate truth definition for L if …

Its general form is therefore:

(1) $\phi \rightarrow \psi$.

Here the logical consequent ψ (which comes first in the English formulation) is a predication of a general term ("adequate truth definition for L") on a general term ("a formally correct definition"). Frege habituated us to regiment this form as a quantified conditional:

(2) $\forall d\big(\rho(d) \rightarrow ATD(d)\big)$.[98]

ATD is short for "adequate truth definition" and ρ will be replaced by a definition of "formally correct definition of $True(x)$". The variable d ranges over definitions, which are strings. So we know the first thing about Z: its ontology includes the domain of strings, i.e. phonology. I use the letter d for the

[98] I use Greek letters for the parts I haven't regimented yet and Latin ones for (sometimes abbreviated) strings in the language Z.

variable in order to separate it from the other variables I will use (d is for "definition"). The antecedent ϕ of the main conditional (1) is:

> … it entails all of the instances of the translational T-schema…

The pronoun "it" refers back to "a formally correct definition", i.e. to d of (2). Of it a complex condition is predicated, for which we use the letter σ. The overall logical form of Convention T is therefore this:

(3) $\forall d(\rho(d) \wedge \sigma(d) \rightarrow ATD(d))$.

This reading makes Convention T a statement of a sufficient condition for the adequacy of a truth definition. It is obviously also meant as a necessary condition.[99] Hence, formally, Convention T is a biconditional, having the form of a definition of material adequacy:

Convention T (Step 1):
$\forall d(ATD(d) \leftrightarrow \rho(d) \wedge \sigma(d))$.

What is missing from this formulation is the generality over languages. On our regimentation scheme generality is expressed by quantifiers. Treating "L" and "M" as variables over languages, we get:

(4) $\forall L \forall M \forall d(ATD(d, L, M) \leftrightarrow \rho(d) \wedge \sigma(d, L, M))$.

However, I will suppress the language variables for the time being.

4.1.2 Formal correctness

A formally correct definition of $True(x)$ (in a language M) is a sentence of the form $\ulcorner \forall x(True(x) \leftrightarrow \tau) \urcorner$, where τ is replaced by the definiens, a formula of M with at most "x" as a free variable and not containing the definiendum. This is easily defined as a condition on strings which we abbreviate as "$DForm(x)$". This implies that apart from phonological ontology, Z must have phonological ideology, i.e. structural-descriptive naming resources (§2.3).

---

[99] See Patterson (2006) for an argument against the necessity reading of Convention T. As a counterexample Patterson has in mind definitions of truth of the form "x is true if and only if x corresponds to a fact that obtains" (p. 28), which don't entail the T-sentences but, he would say, are materially adequate. Patterson has to give a satisfactory theory of facts before his definition can be considered formally correct.
Gupta (2011c) argues that conforming to Convention T is not a necessary condition for the extensional adequacy of a truth definition (nor for its intensional adequacy, nor for what he calls "sense" adequacy). He seems to conclude that this implies that it is not necessary for the material adequacy of a truth definition. But this is too quick. The point of Convention T is to give a *criterion* of adequacy, i.e. a decidable test. Extensional adequacy is not decidable (unless we already possess the concept of truth, and then we don't need a definition). The fact that Convention T is weaker than an extensional adequacy criterion is to be expected, and desired. Convention T allows us to assess a truth definition in the terms of the metalanguage.

Another feature of formal correctness is logical consistency. A sentence is consistent if and only if it logically entails no contradictory sentence. The relation of first-order logical entailment is not only well-defined, but has a complete deductive system, from which it follows that we can define entailment in purely phonological terms. The relation of logical entailment will be abbreviated, as a binary relation between (uninterpreted) strings, as follows:

(5) $E(x,y) \leftrightarrow x\ logically\ entails\ y.$[100]

Let "F" abbreviate the name of some logically false sentence (take "$\exists x(x \neq x)$"), then consistency can be defined as follows:

(6) $Cons(x) \leftrightarrow \neg E(x,F).$

The definition of "formally correct definition", or $FCDef$ is then:

(7) $FCDef(x) \leftrightarrow Cons(x) \wedge DForm(x).$

It should be noted that this definition of formal correctness doesn't ensure consistency, since it doesn't, in fact, ensure the absence of circularity. What it ensures is that the definiendum not occur in the definiens, and that the string itself, as uninterpreted, not entail a contradiction. A circularity or a contradiction could still come about via interpretation, for example if one of the terms in the definiens is itself defined in terms of the definiendum, or if the domain contains an object that is somehow linked to the definiendum (e.g., as its extension). It is important therefore to think of the definiendum as a completely uninterpreted string to begin with.

The definition of formal correctness yields the following partial regimentation:

Convention T (Step 2):
$(d)(ATD(d) \leftrightarrow\ (FCDef(d) \wedge \sigma(d))).$

4.1.3 The T-schema

The informal condition encapsulated in σ reads:

> [the definition] entails all instances of the *translational T-schema*:

[100] See Goodman and Quine (1947), p. 120. Maybe you're worried that completeness is established in model-theory, so it might not be general enough for our concerns. For example, the completeness result doesn't meaningfully apply to the model theoretic proof procedure itself, but surely a philosophical account of truth should apply to that as well. I defer to Kreisel (1967) on this point.

(the translational T-schema): $True(\alpha) \leftrightarrow \varphi$,

where "α" is replaced by the name of an L-sentence and "φ" by the translation of that sentence into M.

Since the T-schema is so central, I'll digress in order to get a firmer handle on schemata in general. Schemata are used in several key places in and around mathematical logic. The most important (alongside in Convention T) are in stating first order axiomatizations of fundamental formal theories, namely logic, arithmetic, and set-theory. Let's say that a *theory* is a deductively closed collection of sentences. By *deductively closed* I mean, as usual, a collection T such that if a sentence x is the logical consequence of some conjunction of sentences in T, then x is in T. A subcollection A of T, such that every member of T is a logical consequence of A, is called a *basis* for T. If a basis can be given finitistically (i.e., in a single sentence), I call it an *axiomatization* of T. If it can be given finitistically in the language of T, it is a *finite axiomatization*.

Let's look at the arithmetical case. The straightforward way to regiment the standard Peano-Dedekind axiomatization is in a second-order arithmetical language L^2_{PA}. The axiom that calls for a second-order language is the induction axiom:

(8) $\forall F\big(F0 \wedge \forall x\big(Fx \rightarrow Fs(x)\big) \rightarrow \forall x Fx\big)$.

There is a way to approximate this axiom in a first-order language L^1_{PA} without predicate variables. It consists in taking as axiom every instance of the *induction axiom schema*:

(9) $\phi 0 \wedge \forall x\big(\phi n \rightarrow \phi s(x)\big) \rightarrow \forall x(\phi x)$,

where an instance is had by replacing "ϕ" (along with its argument) with a unary formula of the language (applied to the argument). The class of sentences arrived at in this way is a basis for a first-order theory of arithmetic which we can call PA. This is not a finite axiomatization, since (9) is not a sentence in L^1_{PA}. Although PA consists only of L^1_{PA} sentences, it is not given in L^1_{PA}. In order to see which language it *can* be given in, we have to express it in a formally correct first-order sentence. Let $indInst(x)$ map a unary formula of L^1_{PA} to the instance of (9) that it generates (this is easy to define in phonological terms). In order to assert all of these instances we use the concept of truth (in its capacity as a device of indirect assertion, see §2.1.3):

(10) $\forall x(wff^1(x) \rightarrow True(indInst(x)))$.

In words, for every unary formula x, the instance of the induction schema generated by x is true. This is precisely what we do when we use the schema (9). In this sentence the quantifier ranges over strings, not numbers, and a truth predicate for L^1_{PA} is used. (10) is therefore given in a metalanguage for L^1_{PA}.

Though every sentence of PA is in L^1_{PA}, the *theoretical unity* of PA is provided by the concept of truth, which cannot be expressed in L^1_{PA}.[101]

One uses a schematic first-order theory instead of a second-order theory when one wants to avoid the ontological commitment that a second-order theory makes to a domain of concepts.[102] Instead of quantifying over concepts, in (10) we quantify over strings. This is just what we want for the language Z which needs on the one hand to apply to all languages, but on the other hand cannot have all languages in its domain. If we can formulate Convention T with quantifiers that range only over the phonology of languages, without taking in their semantics, then Z would be just L_{SD} (from §2.2.4), the language of phonology and structural-descriptive naming. It would apply to all languages since all languages have a phonological aspect. As we will see, Z will have to contain something more.

We go back to Convention T. After the previous step of regimentation, we had the following form:

(11) $\forall d(ATD(d) \leftrightarrow (FCDef(d) \wedge \sigma(d)))$.

The letter "σ" stands for the condition expressed by the translational T-schema:

(12) $True(\alpha) \leftrightarrow \varphi$,

where "α" is to be replaced by the name of an L-sentence and "ϕ" by the translation of that sentence into M. We unpack the T-schema in the same way that we did the induction axiom schema. First, we introduce an abbreviation:

(13) $Tsent_{tl}(x) = \underline{True(}\cdot\, x\, \cdot\underline{)} \leftrightarrow \cdot\ tl(x)$.[103]

This function maps a string to the T-sentence that it generates using the translation function tl. This translation function will be scrutinized in the next section. In order to mark its existence, I mention it in the subscript adjoined to the abbreviation.

In the case of the arithmetical induction axiom schema what we wanted is to assert all of its instances. For this we needed a truth predicate, which meant that the theory had to be given in a metalanguage. If it were the case that a truth predicate is needed also in the case of the T-schema, then Convention T would have to be in a truth-defining metalanguage for the metalanguage M, and since it applies to any M, that would make it the universal metalanguage. Fortunately, the T-sentences are not asserted in

[101] See §2.3.3 for a brief discussion of theoretical unity. The first-order theory is, of course, weaker than the second-order one, for reasons related to the need for higher-order quantification in giving an explicit definition of number (see §3.2.4).
[102] See §3.2.5 for discussion.
[103] I will use corner quotes when mixing use and mention. This is less precise and more readable than the machinery used in §2. In this chapter there is less reason to be precise and more reason to be readable.

Convention T. What is asserted (conditionally) is the fact that they follow logically from the proposed definition. We can call this *hypothetical theoretical unity*, and it doesn't require a truth predicate, just a relation of logical entailment, which can be defined in phonological terms.[104]

Here is a full formalization of Convention T, making use of the schematic apparatus with hypothetical theoretical unity:

(B) **Convention T, fully formalized:**

$$\forall d(ATD(d) \leftrightarrow (FCDef(d) \wedge \forall x(wff^0(x) \rightarrow E(d, Tsent_{tl}(x))))).$$

In words: A string d is an adequate truth definition if and only if it is a formally correct consistent definition, and for every sentence x, d entails the T-sentence generated by x using the translation function tl.

Everything in the definiens, except for the translation function tl, is definable in phonological terms. For readability, we can encapsulate it in a single abbreviation $PhCond_{tl}(d)$. More readably, Convention T becomes:

(14) $$\forall d(ATD(d) \leftrightarrow PhCond_{tl}(d)).$$

4.2. Translation

It should not come as a surprise that translation is the problematic term in Convention T, since it is the only notion that makes an appeal to the semantic content of the languages in question. This appeal to content is not reducible to phonological notions, so it is the main obstacle to a formulation of Convention T. The first suspect is the translation function tl.

4.2.1. Translation and interpretation

Distinguish between *translation* and *interpretation*. Informally, translation can be seen as a relation between two languages, interpretation as a relation between a language and the world. Formally we can conceive of a translation as a (partial) function f on expressions. If x and $f(x)$ are of the same syntactic category, we say that f is *syntactically conservative*; if f is syntactically conservative and defined only over sentences, we say it is *holophrastic*; and if f is holophrastic and preserves logical entailment

[104] Gupta (2002, p. 57) rightly notes that strictly speaking, the definition as we've given it in in §3.2.4 will not *logically* entail the T-sentences, but only the definition along with some phonological theory (this was noted in §3.2.2, p.52). Gupta, a little surprisingly, goes on to suggest that the notion of implication in play in Convention T "plainly cannot be that of logical implication". Maybe it's better to read "cannot be that of plain logical implication".

relations (if $E(x,y)$ then $E(f(x),f(y))$), we say it is *coherent*. The notion of coherence can also be defined for non-holophrastic translations. If two translation functions f and f' are such that for every sentence s, $f(s)$ and $f'(s)$ are logically equivalent, we say that f and f' are *equivalent translations*. This is clearly an equivalence relation, and my discussion in what follows should be qualified as *up to equivalence*. Translations are mappings on strings, and the concepts of coherent and equivalent translations are definable in phonological terms. We can safely allow Z to include such mappings in the domain, without risk of being able to define truth generally. The domain of Z is thus a second-order phonological domain.

In contrast with translation, the interpretation of a language is not merely phonological. For instance, the interpretation of the English name "Jerusalem" is a certain stone-clad city in the Middle East, and the interpretation of a sentence, e.g. "Jerusalem is in turmoil" is, or at least should somehow contain, its truth conditions. In any case a statement of the interpretation of a sentence is, or at least should somehow entail, a statement of its truth conditions. The interpretation of a language L is therefore, for our purposes, nothing more nor less than a semantic theory for L. Interpretation, unlike translation, is not a merely phonological notion. To get an intuition of this, observe that a translation function f from German to Polish may tell us that $f(\text{"}Der\ Schnee\ ist\ wei\text{ß"}) = \text{"}\acute{s}nieg\ jest\ bia\l y\text{"}$, but by itself it will not tell us under what conditions either sentence is true, namely, if and only if snow is white. By contrast if, being acquainted with snow and its color, I can't tell whether "śnieg jest biały" is true or not, then I am not in possession of an interpretation of it.

Alongside this essential difference between the notions, translation and interpretation are also connected at the navel. A truth-conditional semantic theory for a language L is nothing but a truth definition in its metalanguage M.[105] Let d be such a definition. Then we say that the translation function f_d *induced* by d is the function that maps each L sentence s into the alphabetically first sentence in M that gives the truth conditions of s according to d (stands on the right-hand side of a T-sentence generated by s):

(15) $$f_d(x) = \iota_1 y.\exists z(z = \underline{True(}\cdot\, x\, \cdot\underline{)} \leftrightarrow \cdot\, y \wedge E(d,z)),$$

where $\ulcorner \iota_1 y.\phi \urcorner$ means $\ulcorner$the alphabetically first y such that $\phi\urcorner$. Conversely, under the right conditions, for a translation function f we say that d_f is the truth definition *induced* by f.[106] The relation between interpretation and translation is related to the one between use and mention: $f(s)$, as used, *expresses* in M the interpretation of s; as mentioned, it *is* the translation of s into M.

---

[105] See §3.2.5. More precisely it is a satisfaction definition. Using a truth definition introduces some (Quinean) indeterminacy which I will not elaborate on.

[106] Here we need to look at non-holophrastic (but still coherent) translations. The base clause of d_f is determined by f's translations of the atomic formulas, and the domain restriction by f's translation of a quantified atomic formula.

4.2.2. Correct translation

The formulation of Convention T we ended up with included the symbol "tl" as an unanalyzed primitive. Since this is our troublemaker, we are called on to analyze it. We make it into a variable f ranging over string mappings, and modify accordingly the places in Convention T in which it occurred (see (13)):

(16) $Tsent(x, f) = \underline{True}(\cdot\, x\, \cdot) \underline{\leftrightarrow} \cdot\, f(x)$,

Convention T now reads:

(17) $\forall d(ATD(d) \leftrightarrow\ PhCond(d, f))$.

The variable f needs to be bound in order for this to be a well-formed sentence, but It is not enough to simply insert a quantifier somewhere. For example, adding an initial universal quantifier yields an obviously wrong result:

(18) $\forall d \forall f(ATD(d) \leftrightarrow\ PhCond(d, f))$.

On this formulation no restriction is put on f, so any (coherent) translation will yield an adequate truth definition. For example, it is easy to construct a coherent mapping that takes, say, "der Himmel ist blau" to "snow is white". But then by (18), there will be a materially adequate truth definition that entails:

"der Himmel ist blau" is true if and only if snow is white,

which is wrong. Nor is there any other position in (17) in which we can put a quantifier over f (not even if we unpack the biconditional). The source of the problem is not hard to find: it is not the unadorned notion of translation that does the job in Convention T, but the notion of *correct* or *adequate translation*. So let us introduce a predicate $ATF(f)$, holding of string mappings if they embody an adequate translation. The correct modification of Convention T is then:

(19) $\forall d(ATD(d) \leftrightarrow \exists f(PhCond(d, f) \wedge ATF(f)))$.

In words: a truth definition is adequate if it stands in the appropriate phonological relation ($PhCond$) to some *correct* translation function.

It is senseless to speak of a correct or adequate translation between uninterpreted strings, so there is no option but to reintroduce the quantifiers over languages that I have been thus far suppressing. A translation is not correct or incorrect in itself, but only as a translation between a particular interpreted language and another. ATF is not a unary predicate over mappings, but a ternary relation between a mapping, a source language and a destination language. The full Convention T becomes:

(C) **Convention T:**

$\forall L \forall M \forall d(ATD(d, L, M) \leftrightarrow \exists f(PhCond(d, f) \wedge ATF(f, L, M)))$.

In words: a string is an adequate truth definition for L in M if and only if it stands in the phonological relation $PhCond$ to an adequate translation from L into M.

The only non-phonological concept here is ATF. The question is whether ATF forces us to assign to Z universal expressive resources if we want Convention T to apply across the board. This, in fact, seems to be the case. A straightforward analysis of ATF will look like this:

(20) $ATF(f, L_1, L_2) \leftrightarrow (x)(|x|^{L_1} = |f(x)|^{L_2})$,

where $\ulcorner |x|^{L} \urcorner$ is the interpretation function for a language L. Although translation is a merely formal notion, *correct* translation depends on the notion of interpretation. And since our variables L and M need to range over *all* languages, Z becomes the impossible universal metalanguage.

We might try to find a different definition of ATF, one that doesn't rely on a full semantic theory of its arguments,[107] but it is important to recognize that the problem does not really lie with the appeal to interpretation that ATF presumably makes. ATF only points the way. The real problem is more fundamental: it is the fact that, to begin with, Z's domain must include all interpreted languages. This is obvious since the very same string can be adequate as a truth definition for one language, but inadequate for another language, even when the languages have the same (uninterpreted) lexicon. It is a feature of the standard semantics of modern quantification that the truth of a quantified statements is understood in terms of the truth of their instances. Therefore, if Z is to apply generally, it must have all languages in its domain, making it a universal metalanguage. By Modus Tollens, Z's generality cannot be that of standard quantification. This shows that there is no way to express Convention T in a standard first-order language. We have to seek a departure from our standard regimentation scheme.

4.2.3. Further reductions

We can improve the situation by reducing the number of variables over interpreted languages in Convention T from two to one. The last formulation of Convention T that we arrived at was:

Convention T:

$\forall L \forall M \forall d(ATD(d, L, M) \leftrightarrow \exists f(PhCond(d, f) \wedge ATF(f, L, M)))$.

[107] $ATF(f)$ relies only on synonymy between x and $f(x)$, and synonymy is, arguably, weaker than the full interpretation of x and $f(x)$.

What does it mean to quantify over interpreted languages? On standard regimentation, this means that all languages are distinct objects in the domain of the quantifying language. But what kind of objects are they? In §1.3.4 and §3.2.5 we identified a language L with its interpretation function, or with the extension of its satisfaction predicate. Referring to all languages means having all of these functions or extensions in the domain. But this is not the only way to refer to languages. We have mentioned already that the interpretation of a language is identifiable with a truth or satisfaction predicate for it, or, on the model-theoretic approach, with something a little stronger; in any case, the interpretation of a language is *given* by a single sentence in the metalanguage.[108] We can therefore identify a language L with its truth definition d, which is just a string. In this case I write $L = \mathcal{L}_d$. Of course, the string d, if uninterpreted, defines nothing; rather, we need to take d as interpreted in an appropriate metalanguage. If M is L's metalanguage, we write $L = \mathcal{L}_d^M$ for the language defined by the string d when interpreted as a sentence of a language M. I call this way of referring to languages *mediated notation*, since it refers to an object by referring to a representation of it. The mediated language name $\mathcal{L}_d^M$ contains unmediated reference to the metalanguage M. If $\overline{d}$ is the definition of M in a metametalanguage $\overline{M}$, then we can dispose of this, and write: $L = \mathcal{L}_d^M = \mathcal{L}_{\overline{d}}^{\mathcal{L}^{\overline{M}}}$. The unmediated reference to $\overline{M}$ could be eliminated in th same way, but it becomes clear that can never achieve purely mediated notation for languages. These towers of mediated notations represent segments of the language hierarchy. The fact that no purely mediated notation is possible is simply the regress problem, exhibited phonologically (see §3.3).

Because of the regress problem, we cannot get rid of all variables over languages. But we can get rid of all of them but one. There are two ways to do this. In the first, we replace the variable L by a term in mediated notation, based on the variable over strings d:

(21) $\forall M \forall d(ATD(d, \mathcal{L}_d^M, M) \leftrightarrow \exists f(PhCond(d, f) \wedge ATF(f, \mathcal{L}_d^M, M))).$

We can make this even simpler by considering, for a definition d, its induced translation function f_d. By the definition in (15), in conjunction with the characterization of $\mathcal{L}_d^M$, f_d will automatically be an adequate translation function from $\mathcal{L}_d^M$ into M. So (21) is equivalent to:

(22) $\forall M \forall d(ATD(d, \mathcal{L}_d^M, M) \leftrightarrow PhCond(d, f_d)).$

Moreover, from the definition of f_d it follows that d will entail all T-sentences that have f_d translations on their right-hand sides. In other words, that

[108] This follows directly from the finitude constraint of §1.3.3B.

(23) $(x)(wff^0(x) \rightarrow E(d, Tsent(x, f_d))).$

The phonological condition $PhCond(d, f_d)$ is composed of two parts (see §4.1.3): the statement that the definition is formally correct $FCDef(d)$ and the schematic condition spelled out in (23). (22) is therefore equivalent to:

(D) **Convention T (almost final formulation 1):**

$\forall M \forall d(ATD(d, \mathcal{L}_d^M, M) \leftrightarrow FCDef(d)).$

In words: a formally correct truth definition is materially adequate for the language that it defines. This formulation contains, aside from phonology, only a single variable over interpreted languages.

(D) is a little surprising in view of how we are used to thinking about the task of finding a truth definition (e.g. in §3.2). Usually we are given a language L and then we ask what conditions a metalanguage M has to satisfy in order to formulate a truth definition for L, and how this definition is to be formulated. But on the formulation in (D), we start from the metalanguage M and give the conditions under which we can define a new language. I call this formulation the *language-synthetic* formulation of Convention T.

The second way to reduce the number of quantifiers over languages to one is the *language-analytic* formulation. The task in mind here is to say, for a given language pair L and M, under what conditions a definition d is adequate in M for L. Now in order for such a task to be stated precisely, the languages L and M have to be given precisely, i.e. by definitions d_1, d_2 in a metalanguage for both of them, $\overline{M}$. The language-analytic formulation of Convention T is then this:

(E) **Convention T (final formulation 2):**

$\forall \overline{M} \forall d_L \forall d_M \forall d(ATD(d, \mathcal{L}_{d_L}^{\overline{M}}, \mathcal{L}_{d_M}^{\overline{M}}) \leftrightarrow \exists f(PhCond(d, f) \wedge ATF(f, \mathcal{L}_{d_L}^{\overline{M}}, \mathcal{L}_{d_M}^{\overline{M}}))).$

In words: a formally correct definition d in a language M is an adequate truth definition for a language L (where L and M are given in a metalanguage $\overline{M}$) if and only if there is some adequate translation function f such that d entails all T-sentences that depend on f.

We therefore have two precisely regimented formulations of Convention T, both of which have only a single initial universal quantifier over interpreted languages. This universal quantifier, it was argued, cannot be the standard universal quantifier. It remains to be explicated in the next chapter.

5 Abstract Generality and Indexicality

5.1. Setting the task

With the usual first-order quantifier, a universal statement is interpreted in terms of the satisfaction of the statement by every single object in the domain. Let's call this *real quantification*. We have seen that there can be no domain which contains all languages. Therefore, the initial quantifier over languages in (both formulations of) Convention T cannot be a real quantifier. The problem of this chapter is to define a new universal quantifier to serve as the initial quantifier over languages in Convention T. This quantifier needs to allow Convention T to apply to languages generally, without thereby being committed to a domain of languages. We will call this *abstract generality* or *abstract quantification*. To gain some intuition of the difference between the two quantifiers, consider a universal statement:

(1) ⌜for every object x, $\phi(x)$⌝.

We can interpret the quantifier (1) either as real or as abstract. Both interpretations can be glossed as saying that given some particular object a, it is the case that $\phi(a)$. The difference is in the significance we attach to the word "given" in the gloss. If (1) is understood as a real quantification, i.e. as ⌜$\forall x \phi x$⌝, then the word "given" is vacuous: any object a (in the domain) is ϕ. There is no sense in which a needs to be *given* over and above simply being in the domain. For example, one obvious sense of an object a being given in a language L is when L has a name referring to a. In this sense of given, there is no (usable) language in which the real numbers are *given* to us all at once, since they outnumber the names. Yet we can say things about them using the real quantifiers, and in this sense the quantifiers transcend what can be given: they apply to what exists independently of any name for it. Abstract quantification, by contrast, should be understood otherwise than as assuming a standing domain of independently existing objects. If (1) is understood abstractly, then $\phi(a)$ can be inferred from it only once a is really given, in some sense of the word. Saying more precisely what this sense is will be part of our burden.

Let's refer to our usual regimentation scheme, the one set up in §1.3.1, as *standard regimentation*, or *standard semantics*. There is no standard expressive device that does the work of abstract quantification. The language Z has therefore to constitute a departure from standard regimentation. But it is important to distinguish between essential and merely notational departures. Briefly, a non-standard language L constitutes a *merely notational* departure from standard regimentation if there is some standard language in which every sentence of L has a synonym; otherwise L is an *essential* departure. The trouble is that standard regimentation is a very powerful regimentation scheme. In the literature we find a great many departures from standard regimentation and its classical logic: from higher-order logic, to intensional semantics, to intuitionistic logic, etc. etc. But in most cases these departures are translatable

without remainder into a standard language; in fact, these languages are almost always given in a standard metalanguage. A language which can be thus brought back into the fold of standard regimentation is a merely notational departure. This is the "sticky hierarchy effect".

The expressive device of abstract quantification will have to be an essential departure from standard semantics. The fundamental philosophical notion on which the theory of this device will be based is the notion of the *use* of language. This chapter will consequently consist in a substantial digression into the theory of language use, i.e. pragmatics. I will recover a notion of abstract reference and abstract quantification from David Kaplan's philosophical thesis of the direct referentiality of indexical expressions. Although I think Kaplan's philosophical view of indexicals implies the desired notion, this is not the case with his formal theory. In §5.2 I will argue that Kaplan's formal theory is in fact inadequate as a regimentation scheme meant to reflect the philosophical view. §5.3 will be a search after the missing piece, and this will provide us with a precise theory of abstract reference and abstract generality. In the next chapter we will adapt the results of this one to languages without indexicals.

5.2. Conceptual reference, direct reference, and the failure of indexical semantics

The term "indexical" comes to us from Peirce's classification of signs. There have been several tentative theories of something like indexical expressions, most notably Russell's "ego-centric particulars" and Reichenbach's "token-reflexives", but the theory's accepted modern form is due in large part to Kaplan's *Demonstratives* (1989a).[109] Kaplan's essay presents a detailed (though perhaps not very organized) exposition of a philosophical thesis concerning indexicals, namely the thesis that they possess a special mode of reference called *direct reference*. It also contains a formal system, a model theory, meant to reflect the philosophical thesis. My claim in this section is that Kaplan's formal system fails to reflect his philosophical thesis. The reason, in brief, is that it constitutes a merely notational departure from standard regimentation, whereas reflecting direct reference requires an essential departure.

Indexicality is to be understood against the background of intensionality. This holds both for the philosophical doctrine and for the formal semantics. In §5.2.1 I will present the basic idea of intensional semantics, and in §5.2.2 I will argue that intensional semantics is a merely notational departure from standard regimentation. The rest of the section will be devoted to indexicality: §5.2.3 will introduce pragmatics and indexicals in general, in §5.2.4 I will state the direct reference thesis, and in §5.2.5 the

[109] *Demonstratives* was published only in 1989, but was written in the 1970s and was well-known long before it was published.

basic idea of Kaplan's indexical semantics will be exhibited and it will be argued that it fails to capture the direct reference thesis.

A brief methodological comment before we begin. Pragmatics and the theory of indexicality have been developed, and are usually practiced, as theories about natural language competence. The same (though to a lesser extent) goes for intensionality. In this essay I am not interested in modelling natural language competence, but in seeing what it takes (which expressive resources are required) in order to regiment various kinds of conceptual content. "Data" from ordinary language use is not, for me, empirical phenomena that my theory has to explain or predict, but clues about kinds of content that we need to know how to regiment. For example, indexical terms such as "now" and intensional operators such as "always" have intuitively clear and unquestionable meanings, yet on the face of it they elude regimentation in the standard extensional regimentation scheme. My problem is not to model the grammar that actually exists in the minds of speakers, but to see what kind of extensions to the standard regimentation scheme are needed in order to be able to express such notions precisely and responsibly. In practice, the difference between this problem and the one of modelling natural language competence amounts to two things. First, the data I rely on doesn't need to be completely natural. It is not native speaker judgments that I am interested in procuring, but the agreement of my fellow theorists, in particular you, my reader. This means that I can give partial regimentations of my data with the intention of isolating the concept that seems to elude regimentation in the standard scheme. Second, the regimentations I will propose are not to be evaluated according to constraints of learnability or cognitive plausibility, which play an important role in linguistics. In particular, since I make no claim that the regimented form cognitively underlies the naturally given one, I don't need to show how to derive the latter from the former by a series of computable transformations or movements.

5.2.1. Intensional semantics

The intensional framework comes to answer to perceived lacunae in standard regimentation. The first is that on standard regimentation we don't feel we have a good representation of meaning, or linguistic content. The second is that certain notions, unproblematically expressible in natural language, are not captured. The most salient of these, perhaps, is the existence of so-called *intensional operators*.[110] Let L_{pr} be a standard extensional language containing "a" and "b" as sole individual constants and "P" a unary predicate. The domain contains Obama (referred to by "a") and Trump ("b"); "P" means "president". In L_{pr} sentences such as

(2) Trump is president

[110] I don't treat propositional attitudes in this work.

are regimented in a straightforward way:

(3) Pb.[111]

This sentence happens to be true (at the time of writing). Another sentence that happens to be true is:

(4) The president is president.

In L_{pr} this would be regimented as:

(5) $P(\iota x.Px)$.[112]

The following two (partially regimented) sentences have clear and unproblematic meaning:

(6) It is always the case that Trump is president,
(7) It is always the case that the president is president.[113]

The first sentence of this pair is false, the second true. These sentences can't be regimented in L_{pr} because L_{pr} doesn't have a term for "always", and such a term, apparently, can't be added to it, for two reasons. The first, minor, problem, is that we don't have the syntactic category of an adverb (a non-logical operator on sentences) in our syntax. This can be remedied without philosophical consequence by modifying the definition of wff. Let "A" regiment "always". Then (6) and (7) should be regimented, respectively, as:

(8) $A(Pb)$,
(9) $A(P(\iota x.Px))$.

However, this regimentation is faulty, and this is the major problem. "Pb" and "$P(\iota x.Px)$", the respective arguments of "A" in (8) and (9), have the same extension (they're both true). By the principle of compositionality (§1.3.4D), (8) and (9) therefore have the same truth value. But since (6) and (7) differ in truth value, (8) and (9) are not adequate regimentations of them.[114]

Our standard extensional regimentation scheme does not, on the face of it, allow us to regiment the adverb "always". This is good reason to modify it, since terms like "always" are both useful and intuitively unproblematic. The clue to which revision is needed is given in the observation that on closer inspection, it is not accurate to say that sentence (2) is true. For although it is true at the time of writing

---

[111] Straightforward except, of course, for the fact that individual constants like functional expressions are contextual abbreviations, see §1.3.2. See also §5.2.4 below.
[112] Definite descriptions are to be understood, à la Russell, as abbreviations. See below. Notice that (5) (and also (3)) is ambiguous (see §1.3.2), but not in a way that will bother us.
[113] The partial regimentation consists in eliminating tense (the copula is to be read as tenseless) and in using the philosophical "it is the case" construction in order to make the logical form unambiguous.
[114] Compositionality makes it the case that every sentential operator will be a truth-function. This is why it's pointless to add the syntactic category of adverb.

this chapter, 2017, it was not true, e.g., in 2016. Sentences are therefore not true or false *simpliciter*, but true or false *at a time*. Our semantic theory should be such as to assign truth-values to sentences relative to a time. Accordingly, let us define a language of a new kind. Let L^I_{pr} have the same lexicon and syntax as L_{pr}. The difference is that the interpretation function of L^I_{pr} gives semantic values relative to a temporal parameter. For example, the semantic clause for the string "Pb" will be:

(10) $|Pb|^t = 1 \; iff \; Trump \; is \; president \; at \; t.$

As desired, "Pb" is now not true or false in itself but true at $t_2 = 2017$ and false at $t_1 = 2016$. Predicates have their extensions relative to time:

(11) $|P|^t = \{x: x \; is \; president \; at \; t\}.$

The points against which expressions are evaluated are generally called *indices*, and the metalinguistic variable that ranges over them is the *index parameter*. L^I_{pr} has only a temporal index parameter, but there may be many *index types*, such as place (consider "it's raining"), subject ("turnip is tasty"), and so forth. Kaplan glosses indices as *circumstances of evaluation*.[115] "Pb" is true if evaluated with respect to t_2 and false if evaluated with respect to t_1. Since no actual act of evaluation is required to take place, maybe it is better to say, more simply, that "Pb" is true with respect to t_2 (and not "if evaluated").

It is straightforward to define an operator that will regiment "always" in L^I_{pr}. Intuitively, a sentence is always true if it is true at all times. Let $Time$, the collection of all times, be the set $\{t_1, t_2\}$. Then we can define:

(12) $|\ulcorner A(\phi) \urcorner|^t = 1 \leftrightarrow \forall t'_{\in Time}(|\phi|^{t'} = 1).$[116]

These operators are called *index-shifting* operators. For every t we have:

(13) $|A(Pb)|^t = 0,$

(14) $\left|A\left(P(\iota x. Px)\right)\right|^t = 1,$

as desired (though see below). We can place relations on the temporal domain (e.g. precedence), and define further temporal notions in L^I_{pr} (e.g. "before"). In this way intensional languages allow us to express in a formally correct way a wide range of contents that elude standard regimentation.

---

[115] Kaplan (1989a, p. 494). It might be a little confusing at first that the term "index" is used in the theory of intensionality and emphatically not in the theory of indexicality. Montague (1968) introduced the term "index" in his attempt to formalize a theory of indexicals, but his theory was not ripe for it. Montague's theories were very influential in intensional semantics, and "index" came to belong to the terminology of that field.

[116] This is a contextual definition, see §1.3.2.

Intensionality also provides us, ostensibly, with an improved notion of linguistic meaning. In extensional semantics the only object we can associate with an expression is its extension. Sentences strictly speaking don't have extensions, but for uniformity we posited objects that stand for their truth values and thought of these objects as extensions. Except for their strings, on an extensional framework there is no difference between expressions with the same extension, or sentences with the same truth value. Intuitively, the fact that two different sentences happen to be both true is not a fact of great semantic importance, and this is an intuition that extensional semantics apparently fails to capture. In intensional semantics we can isolate an object which more closely resembles the intuitive meaning, or *content*, of an expression. We saw that in L^I_{pr} extensions are evaluated relative to a time. We can define the unrelativized semantic value of an expression, called its *intension* or content, to be the function that takes an index t to the extension of the expression at t:

(15) $$|P| = \lambda t.|P|^t.$$

The intension of a sentence "Pb" is a function from times to truth values, which can loosely be identified with the collection of times relative to which the sentence is true:

(16) $$|Pb| = \lambda t.\left(\begin{matrix}1\ if\ |Pb|^t = 1\\ 0\ \ otherwise\end{matrix}\right) \cong \{t: |Pb|^t = 1\}.$$

The intension of a sentence is sometimes called a *proposition*. Above I gave "A" a contextual definition, but it can also be defined as a function on the intension of ϕ, which is why it is also called an *intensional* operator. That function is the intension of "A":

(17) $$|A| = \lambda i \lambda t.\left(\begin{matrix}1\ if\ \forall t'(i(t') = 1)\\ 0\ \ otherwise\end{matrix}\right) \cong \{Time\}.$$

In this way intensional semantics provides us with an explication of the intuitive idea of linguistic content. Intensional systems in the literature are often much more complicated, with many interacting index types, various intensional operators, and other sophistications. The sketch above captures, however, its fundamental feature, that of explicating intensional operators using quantifiers over special entities.[117]

---

[117] There are other ways to approach intensionality, but this is the most popular and the one at the basis of Kaplan's indexical semantics. See von Fintel and Heim (2011) for a rich variety of developments and applications of the basic principle. My focus is on semantics, but there is also much to say about intensional *logic*. See Fitting (2015) (*SEP* entry).
The kind of special entity used most often in the literature is not times, but possible worlds. This makes no difference as far as the formal system is concerned. Notice that either way, propositions and the sentences that express them turn out not to be truth bearers at all. Even a sentence which has the same truth value at all times, e.g. "$A(Pb)$", are not strictly speaking true or false, since they are not of the right kind. See §5.2.5 for more on this.

5.2.2. Internalization and mediated content

Intensional semantics, as developed above, is a merely notational departure from standard regimentation. Let L_{Ipr} be a standard extensional language, such that its domain is the union of the domain of L^I_{pr} with the collection $Time$, and its vocabulary is like that of L^I_{pr} except that to the atomic predicates we add an argument place for time, and we add a predicate $Time(x)$ which holds only of times. The interpretation of the extra argument is time of evaluation, so that, for example, we have:

(18) $|P(x,y)| = 1 \leftrightarrow x\ is\ president\ at\ time\ y.$

It is easy to see that every expression of L^I_{pr} is correctly translatable into L_{Ipr}. The sentences:

(19) $Pb,$

(20) $A(Pb),$

have the following translations:

(21) $P(b,x),$

(22) $\forall x(Time(x) \rightarrow P(b,x)).$

The observation that sentences such as "Obama is president" are neither true nor false as they stand is now captured by the fact that their translations into L_{Ipr}, here (21), are not sentences at all but open formulas. Intensional semantics is therefore only a notational variant on standard regimentation, and not an essential departure. More generally, if L^* is a language that purports to be a departure from standard regimentation, but is given in a standard metalanguage M, then we can often define a standard object language L_* which has a translation for every sentence of L^*. We simply take the expressive resources used in M to give the semantics of L^*, and include them in L_*. This strategy I call *internalization (of the metalinguistic resources)*.[118] Internalization is important since it shows clearly, in the object-language, what expressive resources are required for the regimentation of a certain notion. Take the three languages of this section: L_{pr}, L^I_{pr} and L_{Ipr}. Comparing L_{pr} and L_{Ipr} shows immediately that regimenting intensional operators involves ontological commitment to indices. In L^I_{pr} this commitment is not avoided but only hidden away in the metalanguage.[119] If we want to preserve the convenience of L^I_{pr}, we can define abbreviations in L_{Ipr} which emulate it:

---

[118] The procedure is a little more involved than how I sketched it, but still pretty straightforward. See Cresswell (1990) for an extended treatment of the relation between intensional and internalized languages ("internalized" is my term, not his). See also Schlenker (2003, §3, Appendix II). Schlenker's target in that paper is indexicality, not intensionality. The internalized language is usually more expressive than the notational departure.

[119] In §3.2.4 we preferred a first-order definition of truth with more sets in the first-order domain to a second-order definition. This can be viewed as a case of internalization.

(23) $P_i x \leftrightarrow P(x, y)$,

(24) $A(P_i x) \leftrightarrow \forall y P(x, y)$.

This lets us use the expressions of L_{pr}^{I} within L_{Ipr} with the same meanings.

This procedure takes care of intensional operators, but leaves us without a satisfactory general notion of content. In L_{pr}^{I} we could speak of the intensions of certain expressions (see (15), (16)), and this explicated for us the philosophical notion of linguistic content. These objects can be defined in the same way in L_{Ipr}. For example, there is no problem in defining the set of times x for which the wff "$P(b, x)$" is true. As before, this will give us the set of times in which Trump is president. But now this set doesn't appear to explicate a notion of linguistic content more informative than extension.

Since linguistic content will be important for us in the sequel, it is important to come up with some way to approach it. The content of sentences and formulas in standard languages are, respectively, truth and satisfaction conditions. But conditions (though said in the plural) are not a kind of object in any obvious sense. If we want to speak of *the* content of an expression, we find ourselves at a loss for objects. This is a difficult problem, and I suggest that we settle for less than an object that really captures *the* content of an expression. Instead of content proper, I propose that we make do with statement of content. In other words, instead of reifying truth conditions into objects, we take statements of truth conditions as a substitute (an ersatz, in Lewis's term). These statements are sentences, or strings, and so unproblematic objects. For example, the content (substitutes) of "$P(x, y)$" and "$P(a, t_1)$" will be the right-hand sides of their respective semantic clauses:

(25) $|P(x, y)| = 1 \; iff \; x \; is \; president \; at \; time \; y$,

(26) $|P(a, t_1)| = 1 \; iff \; Obama \; is \; president \; in \; 2016$.

The content of an expression in an object-language L is therefore a string, interpreted in its metalanguage M. This is the *mediated approach* to content.[120]

5.2.3. Pragmatics

In L_{pr}^{I}, sentences such as Pb are not true or false. But this is strange, since sentences such as "Trump is president" are asserted all the time. Assuming that the assertion aims at truth, this calls for explanation. The explanation, as I am presenting it, does not belong to the theory of content but to the theory of

[120] Compare the mediated notation for languages introduced in §4.2.3. Note that we don't take any string which is the right-hand side of a *true* equivalence, but a string which is the right-hand side of an equivalence which *follows* from the semantic theory. For example, unless "Hesperus = Phosphorus" follows from the semantic theory, the content of "Hesperus" and of "Phosphorus" will turn out different.

assertion. This is a lacuna in intensional semantics that calls for another departure from standard regimentation, pragmatics. A further linguistic phenomenon that intensional semantics fails to deal with is the behavior of indexical expressions such as "I" and "now". Sentences such as "I am president" are not true or false as they stand, but with respect to whoever utters them. This relativity is, or should be, of another kind from the relativity to circumstance of evaluation that is the defining feature of intensional semantics, as we will see.

Both of these facts – the assertability of propositions and the behavior of indexicals – are facts about language *use*. The science of language use is called *pragmatics*. This is a basket term that includes under it many different theories and approaches. I will focus on one strand of it, Kaplan's theory of indexicals. The guiding intuition in pragmatics is that not only the extension of expressions, but also their very content, can depend on external factors. Consequently, it is postulated that words have another kind of meaning alongside extension and content: a character. Strawson, in a paper that helped shape pragmatics as a science, makes the case that linguistic meaning cannot be mere reference. He writes:

> To give the meaning of an expression… is to give *general directions* for its use to refer to or mention particular objects or persons; to give the meaning of a sentence is to give *general directions* for its use in making true or false assertions. It is not to talk about any particular occasion of the use of the sentence or expression. The meaning of an expression cannot be identified with the object it is used, on a particular occasion, to refer to. The meaning of a sentence cannot be identified with the assertion it is used, on a particular occasion, to make. For to talk about the meaning of an expression is [to talk]… about the rules, habits, conventions governing its correct use… (1950, p.327).[121]

According to this view the truth-bearer in pragmatics cannot be the sentence, but has to be some object which is sensitive to the occasion of use. Language is used in a variety of ways, but in order to fix ideas let's take the conversation as the paradigm case. Again, conversations are many and varied, but for simplicity we can restrict attention to situations in which two competent speakers of a shared language take turns asserting declarative sentences to one another. Call this simplification the *conversation heuristic*. On this heuristic we can identify an instance of language use with the act of voicing of a sentence in the shared language as part of a conversation. Such an act we call an *utterance*, and this will now be our truth-bearer in this chapter.

[121] Ironically, that very same paper ends with a rather discouraging statement for the development of pragmatics. Strawson there says about standard semantics that it doesn't "give the exact logic of any expression of ordinary language; for ordinary language has no exact logic".

A pragmatic theory will take the form of a truth or satisfaction theory for utterances. Sometimes I will call this a *pragmatic semantics*.[122] As in the case of standard semantics, we want the theory to be finitely statable, and since there are infinitely many possible utterances, its statement should rely on the way utterances are composed from basic elements. We want to rely as much as possible on the already existing semantic theory for sentences, so we analyze utterances into pairs consisting of the sentence uttered and the circumstances of utterance, encapsulated in an object called the *context (of use)*:

(27) $Utterance(x)\ iff\ \exists y \exists z(wff^0(y) \wedge Context(z) \wedge x = \langle y, z \rangle)$.[123]

A pragmatic truth predicate is therefore a binary relation between sentences and contexts. More generally, and in continuity with our previous habits, we can think of a pragmatic theory as a function from strings and contexts, which we write $|\alpha|^c$, to extensions. As before, true utterances are stipulated to have the number 1 as their extension. My interest here is not with pragmatics in general but only with one branch of it: the theory of indexical expressions, especially as conceived by David Kaplan.[124] If a lexeme α is such that for some contexts c_1, c_2, $|\alpha|^{c_1} \neq |\alpha|^{c_2}$, I say that α is an *atomic indexical*. Any expression containing an atomic indexical is an *indexical expression*. I will disregard demonstratives, and focus on what Kaplan calls "pure indexicals", which don't depend on an act of demonstration for their content.[125]

5.2.4. Direct reference

It is not Kaplan's formal theory of indexical semantics that I'm after, but his (I think underdeveloped) philosophical thesis about the special indexical mode of reference. Kaplan's philosophical thesis regarding indexicals is that they are *directly referential*. Other expressions we can call *descriptively* or *conceptually referential*.[126]

---

[122] Sometimes *indexical semantics*, or *theory of indexicals*, or simply *semantic theory* when the object-language has indexicals.

[123] By analyzing utterances into sentence-context pairs I aim to sidestep a debate between Kaplan and (later followers of) Reichenbach concerning the nature of the pragmatic truth-bearer. Briefly, Kaplan thinks that logical validity can be defined only for sentence-context pairs and not for utterances, because utterances are concrete dated particulars and therefore cannot figure in two positions in a logical argument (see Kaplan (1989a, p.522), more recently Perry (2017, §3)). I take sentence-context pairs to be idealizations of the concrete dated utterance events, for use in semantic theory, in the way that physicists often work with points having mass, even though points cannot have mass. For this line, see Garcia-Carpintero (1998, pp.539ff).

[124] The source for this theory is Kaplan (1989a) (which was written in the 1970s). Henceforth references to Kaplan (1989a) will be given without the publication year. A relatively recent, short and reasonably precise introduction to indexicals is Schlenker (2011).

[125] Kaplan p.491.

[126] Kaplan contrasts direct reference with *sense-mediated* reference, in Frege's sense of "sense". I'm not sure what that sense is, but Kaplan thinks of it as conceptual in nature (though this might not be Frege's concept of "concept"). See p. 505fn31.

Conceptual reference is the kind of reference that characterizes standard languages. A good place to observe this is the case of proper names or individual constants. Recall that, strictly speaking, in standard languages as I have defined them there are no individual constants (see §1.3.2). The proper regimentation of, say, a proper name such as "Trump", would use a one-place predicate "*T*" which holds only of Trump.[127] The sentence

(28) Trump is president,

would then be regimented, using Russell's technique, as a complex sentence saying that there is a single thing satisfying T, and that thing is president:

(29) $\exists x(Tx \wedge (y)(Ty \rightarrow x = y) \wedge Px)$.

If we wanted to obscure things by emulating the surface grammar of (28), we could introduce a contextual abbreviation which would make (29) look like a simple predication:

(30) $P(\iota x.Tx)$.

And if we wanted to conceal the logical form of our regimented sentences even more, we could abbreviate the inner expression by a pseudo-individual-constant "b":

(31) Pb.

We can say, somewhat loosely, that b *conceptually refers* to Trump:

(32) $|b| \cong Trump$,

meaning that the systematic contribution that the contextual abbreviation "b" makes to semantics consists in the individual Trump. Conceptual reference, on this way of construing it, is achieved by a combination of quantifiers and predicates.

At this point you might want to object as follows. The regimentation (29) of the unregimented (28) required a considerable change in form. In particular, a unique existence claim is added which is nowhere to be seen in (28). We resort to this regimentation procedure because we've willfully banished individual constants from our regimentation scheme. But this choice was not forced upon us. Many presentations of first-order logic, equivalent to our own, include individual constants as undefined primitives in the language. On such presentations (31) is not an abbreviation but an *bona fide* sentence which closely mirrors the syntax of (28). Why make things more complicated than they are?

The answer to this is that if we do include individual constants in our object-language, (32) stops being a loose way of speaking and becomes a regular semantic clause in the metalanguage. In this case it is,

[127] This is Quine's procedure in (1981, p.149ff).

in a sense, presupposed that there is a unique object in the domain referred to in the object-language by "*b*" and in the metalanguage by "Trump". Otherwise we could not use the equality sign. But this is precisely what (29) adds to the ostensibly simple predication in (28). The difference is that in (29), the unique existence claim is explicitly stated in the object-language, whereas if we construe (31) as unabbreviated, it is implicit. Worse, if "Trump" is construed as an individual constant in the metalanguage, then a unique existence claim is implicit in the metalanguage as well. Allowing genuine individual constants would only hide their conceptual nature, not do away with it.[128]

The key feature of conceptual reference is its dependence on an ontology and ideology which are given in advance of the particular use of the language. Conceptual reference is thereby made independent of the circumstances of use. By contrast, a term *refers directly* if it refers not in virtue of its conceptual content, but through some kind of special relation between the act of uttering it and its referent. The precise nature of this relation is a matter of dispute. On Peirce's threefold classification of signs, an indexical sign is one that denotes by virtue of some kind of existential relation, such as physical contiguity or causal connection. The usual example is the smoke signaling the fire, but this example is of no great use to us.[129] Russell also had a theory of something like indexical expressions which he called "egocentric particulars", reference to which is grounded by the epistemological relation of acquaintance. Logically, Russell held that only objects with which we are directly acquainted can be referred to through genuine individual constants ("logically proper names"); all other objects were to be referred to through descriptions, which is to say by combining quantifiers and predicates. Russell's strictures on what counts as direct acquaintance eventually precluded him from acknowledging most things as objects of acquaintance, making his theory unattractive, but later writers use a milder notion of acquaintance to ground direct reference.[130] I don't wish to commit to either Peirce's or Russell's view of the relation between indexicals and their referents. I will refer to this relation, neutrally, as *contact*.[131]

---

[128] This is another case of internalization.

[129] See Burks (1949), also Atkin (2015), for discussion of Peirce's theory of indexicals. Perry (2017) seems to think that if I use "you" to refer to the person I'm addressing, (in this case you, dear reader,) then this reference is (partly) in virtue of a real relation between us, involving causation and perception. Unfortunately, he doesn't elaborate (as this isn't his topic), but it is unclear in which way causation is involved, and whether perception is really needed. For example, I am not perceiving you at the moment, and any causal relation between you and occurrences of "you" in this footnote is in the wrong direction. Nonetheless, I'm sure I've succeeded in referring to you. So it isn't clear whether and how perception and causation are really involved in referring to an addressee.

[130] See Russell (1910-11), (1972, pp. 28ff). See Evans (1982) for a later defense of acquaintance theory.

[131] This term I take from Posy (2017), who uses it in order to attribute something like a direct reference view to Kant. There is a substantial literature on direct vs. descriptive reference which I will not refer to in a systematic way. Most of it is concerned with reference in natural language, and says little to my own concerns (see §5.2). For a recent comprehensive treatment of reference with many, well, references, see Hawthorne and Manley (2012).

The difference in mode of reference between indexicals and non-indexicals implies a difference in the kind of meaning. The kind of meaning that governs direct reference is called *character*, whereas the reference of non-indexicals is determined by their *content*. This difference should not, on the face of it, affect the objects referred to. It is not the case that indexicals refer to special objects, to which conceptual reference is impossible. Rather, we can assume that any object referred to directly can be referred to conceptually (though not necessarily in the same language). This fact we call *the reality of indexicals*.

Direct reference is a philosophical thesis. But if it is to be more than just an epiphenomenon, it should take part in turning the wheels by placing constraints on a formal semantic theory for indexicals. In Kaplan I find two features of indexicals that can be made into such formal constraints: the definability of the notion of *pragmatic validity*; and *the prohibition on monsters*.

A sentence is *pragmatically valid* (*PV*) if true whenever uttered.[132] The typical example is:

(33) I am here now.[133]

Pragmatic validity is, according to Kaplan, to be distinguished from the superficially similar phenomena of necessity and eternity. Though true in every utterance, pragmatically valid sentences are not always or necessarily true. To see this, consider that a sentence $\ulcorner\phi\urcorner$ is always (necessarily) true if and only if the sentence $\ulcorner always\ \phi\urcorner$ ($\ulcorner necessarily\ \phi\urcorner$) is true simpliciter. But we can certainly think up false utterances of:

(34) I am always here (now),

(35) I am necessarily here now.

Therefore, pragmatic validity is distinct both from necessity and from eternity.

The pragmatic validity of a sentence is due to the indexicals in it.[134] When uttered, the indexicals in (33) refer (either through existential relation, or acquaintance, or what have you) to the speaker of the utterance, the time of the utterance and the place of the utterance, respectively. By anatomy and physics, the speaker will have to be situated in precisely the place of the utterances at the time of utterance. Therefore the truth conditions of (33) will always be fulfilled. This is a feature of indexicality that a

[132] "Pragmatically valid" is not Kaplan's term. He thinks of such sentences as logically true, relative to his special model theory for indexicals. See p. 509.

[133] Although this is the typical example, it is not really pragmatically valid, since its negation can be used truly, for example in written notes or recorded messages. On the conversation heuristic it does come out valid. More on this below.

[134] Actually, there are pragmatically valid sentences that do not involve indexicals essentially, for example "there have been utterances of English sentences". Another kind is expressible in French: "Je ne te tutoie jamais" (See Zimmerman (1997)). I won't address them here.

formal semantic theory should express. It is therefore a requirement on such a theory that it be able to define a predicate PV applying to all and only pragmatically valid sentences.

The second formal constraint that the thesis of direct reference places on formal theories of indexicals is the prohibition on monsters, as follows. Above we saw that pragmatic validity is distinct from necessity or eternity. But can't it be just a further kind of intensional notion? After all, there are sentences that are always but not necessarily true and vice versa, but we model this within intensional semantics by adding another index type, not by calling for a new kind of semantic system. Likewise, we might explicate pragmatic validity by a new index type – contexts. However, according to Kaplan there is an important formal difference between intensional and pragmatic notions: there can be no pragmatic operators in the object-language. If pragmatic validity were an intensional notion, we could define a pragmatic validity operator analogous to "always" and "necessarily", which when applied to a sentence $\ulcorner\phi\urcorner$ yields another sentence which is equivalent to $PV(\ulcorner\phi\urcorner)$. The thesis of direct reference requires that such operators be impossible. Kaplan isn't very clear on the reason, but something along the lines of the following seems to be in play. Pragmatic operators would make the truth of an utterance they occur in independent of the things that stand in contact with that utterance. For example, imagine that such a pragmatic validity operator $\ulcorner\Pi\phi\urcorner$ exists. Then

(36) $\Pi(I\ am\ here\ now)$,

is invariably true. In particular, the objects with which an utterance of (36) comes in contact play no role in this. The indexicals in it therefore do not refer directly. Such operators on characters Kaplan calls *monsters*,[135] and although he doesn't treat them in this way, I will make the indefinability of monsters the second criterion for the adequacy of a theory of indexicals.

Let me note that in Kaplan's own writings the prohibition on monsters is never given as an organized piece of philosophy. It is not even clear whether his claim is empirical or a-priori.[136] Empirical evidence has been adduced against it for the case of natural language.[137] The empirical question will not concern me: if there are monsters in natural language, then to that extent the indexicals in natural language are not directly referential. My intention is to isolate direct reference as a philosophical notion, and in this I concur with Kaplan's intuition that pragmatic operators should not be definable. However, as we shall

[135] Kaplan, p.510.
[136] The confusion is visible in quotations such as this: "I am not saying we could not construct a language with such operators, just that English is not one. And such operators *could not be added to it*." (p.510, italics in the original)
[137] See especially Schlenker (2003) and the references in Schlenker (2011).

see, the prohibition on monsters will be the undoing of Kaplan's own formal semantics for indexicals, and the clue to my revision of it that will eventually lead us to a notion of abstract reference.[138]

To sum up, the philosophical thesis of direct reference makes the following demands on formal semantic theories that presume to embody it: (a) They need to be able to define pragmatic predicates, in particular the predicate of pragmatic validity; (b) They shouldn't be able to define operators in the object-language that express these predicates, in particular there should not be a pragmatic validity operator. We can crystallize these requirements into a two-step test:

(G) **The direct reference test:** given a pragmatic semantic theory,

a. Define the pragmatic validity predicate PV, i.e. find a formula δ in the metalanguage such that, for example, $\delta(\text{"}I\ am\ here\ now\text{"})$ holds, whereas $\delta(\text{"}Obama\ is\ here\ now\text{"})$ doesn't;
b. Show that a pragmatic validity operator is not definable in the object-language, i.e. an operator Π such that $\ulcorner\Pi(\phi)\urcorner$ is true whenever $PV(\ulcorner\phi\urcorner)$. Such an operator is successfully defined if, for example, $\ulcorner\Pi(I\ am\ here\ now)\urcorner$ turns out true (in every context), while $\ulcorner\Pi(Obama\ is\ here\ now)\urcorner$ turns out false. The theory passes the test if it is shown that it is impossible to define it successfully.

I submit that a formal theory expresses the philosophical thesis of direct reference only if it passes this test. Unfortunately, Kaplan's own system, which is also the standard in the industry, fails the test.

5.2.5. Kaplan's indexical semantics

Kaplanian indexical semantics amounts, formally, to adding another parameter, on top of the index parameter, to the semantic function. This parameter represents the context of use. Let L^{IP}_{pr} be like L^{I}_{pr} but such that the semantics is now a three-place function$|\phi|^{c,t}$.[139] Alternatively, we can internalize intensionality along the lines of §5.2.2 to get a language L^{P}_{Ipr} with only a context parameter in the semantic function. We add to this language, on both versions, the indexicals "I", "now" and "$here$". To the metalanguage we add the functional expressions "$speaker(c)$", "$time(c)$" and "$place(c)$" that refer to the features of the context, which serve as the semantic values of the indexicals:

(37) $\quad |now|^{c} = time(c),$

[138] See Israel and Perry (1996) for a defense of monsters, which is however tightly linked to Perry's conception of linguistic content. Rabern (2013) makes a case that the usual objectual quantifiers should strictly speaking be thought of as monsters. This observation is significant, and will to a certain extent be borne out by my own considerations, especially of the next chapter. An illuminating early discussion is Thomasson (1975). Predelli (2014) discerns a three-way equivocation of "monster" in Kaplan. This distinction makes sense for English, but is arguably not relevant to the formal system developed here.
[139] It may have more than one index parameter.

(38) $|I|^c = speaker(c)$,

(39) $|here|^c = place(c)$.

The assertability problem mentioned above can be easily solved. The problem was that on the one hand sentences such as "Trump is president" were not true or false in themselves, but true or false only as evaluated with respect to time; and on the other hand they get to be asserted all the time. The answer is that when asserted, they are evaluated with respect to the time of assertion. This answer can be represented in two ways. We may stipulate an assertion principle, which states that the semantics of an asserted sentences is $|\phi|^{c,time(c)}$. This is a rule for interpreting assertions, and it is more suitable for the uninternalized L_{pr}^{IP}. Alternatively, we may prescribe a rule of regimentation which adds the indexical "now" to every sentence upon assertion. The proper way to regiment "Trump is president" as an assertion is:

(40) $P(b, now)$.[140]

The unrelativized semantic values of indexicals are called *characters*. In the formal system, they are represented as functions from contexts to intensions (for L_{pr}^{PI}) or from contexts to extensions (for L_{Ipr}^{P}):

(41) $|I| = \lambda c.\, |I|^c = \lambda c.\, speaker(c)$.

For uniformity let's say that all expressions, even non-indexicals, have characters:

(42) $|P| = \lambda c.\, |P|^c = \lambda c.\, \{\langle x, y\rangle : x\ is\ president\ at\ time\ y\}$.

If the context parameter in the statement of the character is vacuous, as in (42), we say it is a *stable character*; otherwise, as in (41), a *proper character*. Sentences too have characters:

(43) $|P(I, now)| = (\lambda c.\, 1\ iff\ |I|^c \in |P|^c) \cong \{c : speaker(c)\ is\ president\ at\ time(c)\}$.

This is, in a nutshell, the basic idea in Kaplan's formal system of indexical semantics. Unfortunately, it doesn't pass the direct reference test. Recall, the test has two clauses. Clause (a) demands that we define the pragmatic validity predicate PV in our system. Informally, a sentence is pragmatically valid if every utterance of it is true. The definition for L_{Ipr}^{P} is obvious:

(44) $PV(s)\ iff\ \forall c(|s|^c = 1)$.

We can show that the sentence:

---

[140] This is why propositions are not truth bearer. Truth for them is an inherently pragmatic notion. If, as Kaplan (e.g., p.489, p.499) and Lewis (1986, pp. 92ff) maintain, "the actual world" is an indexical expression, then this applies also to propositions as sets of possible worlds.

(45) I am here now

turns out pragmatically valid, and

(46) Obama is here now

doesn't. That (45) is PV follows at once from our conversation heuristic: taking utterances as acts of voicing sentences ensures that the utterer will always be present at the place of utterance. Therefore in every context c, the speaker will be in the place at the time of c, which is precisely what it takes to make (45) true. An utterance of (46) will be false in any context c such that Obama is not present in the place at the time of c; for example, if Trump were to utter (46) in the White House in 2017 (and Obama would be elsewhere), then his utterance would be false. The fact that this context is a possible one, and therefore a member of the domain of contexts, entails that (46) is not pragmatically valid. This is as desired.

Clause (b) of the test is to show that a pragmatic validity operator is not definable for the object-language, i.e. no operator Π such that ⌜Π(ϕ)⌝ is true just in case $PV(\ulcorner\phi\urcorner)$. Unfortunately, on our system it is quite straightforward to define such an operator, based on the definition of PV:

(47) $|\ulcorner\Pi(\phi)\urcorner|^c = 1\ iff\ \forall c'(|\ulcorner\phi\urcorner|^{c'} = 1)$.

Regardless of when and where and by whom they are uttered, the sentence

(48) Π($I\ am\ here\ now$)

will be true, and

(49) Π($Obama\ is\ here\ now$)

will be false. "Π" is a pragmatic validity operator in the language L^P_{Ipr}, i.e. a monster. Kaplan's system therefore fails the second clause of the direct reference test.

What is the significance of this failure? The prohibition on monstrous operators is a consequence of the philosophical doctrine that indexicals refer directly (§5.2.4). Consequently, a formal system on which such operators can be defined, like L^P_{Ipr}, does not capture the philosophical doctrine. It doesn't distinguish, as Kaplan says with regards to other systems, between "the distinct conceptual roles played by contexts of use and circumstances of evaluation".[141] Kaplan himself did not define a pragmatic validity operator in his system, but such an operator is readily defined in it, in a completely analogous

[141] See Kaplan, p.512. Kaplan says this about two dimensional logics which he contrasts with his own system. But the contrast he's after is to be found only in the informal presentation.

way to how the corresponding intensional operators are defined.[142] The reason is that the collection of contexts is exactly on a par with the collections of times, places and worlds, so contexts are real objects in the domain of the metalanguage, in the range of the usual quantifiers. What, short of an ad-hoc restriction, should stop us from using such quantifiers to define context-shifting operators in the object-language? The distinction between indexicality and intensionality, character and content, is just a difference in label. If we want to uphold the direct reference thesis in our formal semantics, we need to represent contexts differently than as standard (real) objects in the domain.

5.3. Abstract reference

Our task is to mount a new construal of the context parameter c, one which will do justice to the direct reference thesis. This means that our interest lies no longer in the pragmatic language L^P_{Ipr} (henceforth simply L^P), but in a metalanguage for it M^P. We saw that conceiving M^P in a straightforward way as a standard language, with the context parameter c as a standard variable, is not the right way to go. What we are looking for is not a new language, but a new kind of language, a new regimentation scheme.

Our standard scheme, first-order extensional logic, is very general. There are many different first-order languages, differing with respect to the objects that they refer to and the concepts that they express. The regimentation scheme doesn't tell the difference between kinds of objects: frogs, numbers, times and possible worlds, from a logical point of view they are all just objects. Whatever kind of object you can find, and whatever you wish to predicate of it, there is a standard language that will let you do it. This is a sense in which our standard language hierarchy is unbounded. Nevertheless, it has its limits. Standard regimentation countenances only objects and predicates, quantification over those objects, and truth-functional connection of formulas. Intensionality seems at first sight to transcend these limits, but if intensional languages are defined in terms of intensional objects such as times and possible worlds, then we see that they constitute only a notational departure from the standard. Our hierarchy is not only unbounded, it is sticky.

However, if the direct reference thesis is true, then the previous section shows that indexicality resists being tamed in this way. In showing that Kaplan's system fails the direct reference test we made no assumption about what *kind* of object contexts are. The only thing that was assumed is that they are objects in the full first-order logical sense: that which is referred to by the variables and quantified over by the quantifiers of standard languages. This is the assumption that we need to reconsider. What is required is a new *logical* category, and a conception of contexts that makes them part of that category.

---

[142] To get a definition of the pragmatic validity operator in Kaplan's system, look at his own definition of ⌜$\Box\phi$⌝ in p.545, and replace the set of worlds by the set of contexts (adjust the variables accordingly).

5.3.1. Three questions

Such a new conception will have to answer three guiding questions:

(H) Three Questions:

a. **A metaphysical question**: What kind of thing is a context, if it isn't an object in the standard sense of the word? In particular, what is the essential difference between a context and an index? The tension here is that a satisfactory answer to a "what kind of thing" question would normally specify the constitution of the thing in terms of previously understood objects using logical or set-theoretic construction. For example, the concept of (uninterpreted) *sentence* is defined in terms of strings of phonemes; and a proposition, often, in terms of sets of possible worlds, etc. But if such a definition is available for contexts, why aren't they just another kind of object?

b. **A (meta-)semantic question**: What is the relation between the parameter *c* and contexts? A standard variable refers individually to objects in a domain (relative to an assignment function). Given a domain of contexts, what prevents the context parameter from referring to them individually in the standard way?

c. **A formal question**: In view of the answers to the previous questions, what are the formal properties of *c*? How should it be allowed to combine with other expressions to form sentences, and what are the logical properties of the sentences it occurs in? What is the semantics of sentences containing *c*? Our answer to this question should pass the direct reference test from §5.2.4.

Our philosophical guiding thread will be the thesis of direct reference. Here is part of Kaplan's summary to have before our eyes:

> In the case of [the indexicals], the linguistic conventions which constitute *meaning* consist of rules specifying the referent of a given *occurrence* of the word ([less abstractly, an utterance]), in terms of various features of the context of the occurrence. Although these rules fix the referent and, in a very special sense, might be said to define the indexical, the way in which the rules are given does not provide a synonym for the indexical. The rules tell us for any possible occurrence of the indexical what the referent would be, but they do *not* constitute the content of the occurrence. Indexicals are directly referential. The rules tell us what it is that is referred to. Thus, they *determine* the content (the propositional constituent) for a particular occurrence

> of an indexical. But they are not a *part* of the content (they constitute no part of the propositional constituent).[143]

Kaplan is struggling here to clarify the distinction between content and character. He makes in this passage three contrasts: first he says that the characters of the indexicals "[specify] the referent", "fix the referent", and "define the indexical", but in a way that "does not provide a synonym for the indexical";[144] second, that they are rules that "tell us for any possible [utterance] of the indexical what the referent would be," though "they do *not* constitute the content of the [utterance]"; finally, that "they *determine* the content… but they are not a *part* of the content" of an utterance. However, these contrasts are not part of his technical apparatus. The distinction between content and character, index and context, belongs only in the informal gloss.

The purpose of the present section is to offer a framework which formally captures Kaplan's contrasts in the quote above. The procedure will be this. First (§5.3.2) I will consider Kaplan's own understanding of context, on which contexts are not essentially dissimilar to indices. I will argue (§5.3.3) that this understanding misses the mark, but will extract from it its intuitive core, which is the necessary presence of a linguistic agent in a context of use. This leads us to consider in depth the notion of linguistic agency. This notion is explained using the idea of content generation, though the specific sense of this idea has to be made precise (§5.3.4). Briefly, content generation and linguistic agency will be explicated in terms of the division of ontological commitment between the metalanguage, in which characters are stated generally, and the object-language in context, in which indexicals are used. The agent of a context takes upon himself the ontological commitment to the references of the indexicals, thereby relieving the metalanguage of it. In this way characters have a special mode of reference which doesn't imply ontological commitment, which I dub *abstract reference*. This mode of reference lets us take apart the very concept of object into its conceptual constituents, and put them together again in a different way. The result is what I shall call *abstract objects* (§5.3.5). Contexts will be such things, but not only. Finally (§5.3.6), I will lay out a syntax, semantics and logic for abstract objects and show that the resulting system passes the direct reference test.

5.3.2. Contexts as speaker-time pairs

In order to see what contexts are, let's observe what contexts do. They were introduced in §5.2.3, when we recognized that pragmatics mandates a change of truth-bearer, from the sentence to the instance of use of a sentence. Adopting the conversation heuristic, we narrowed this down to the event of utterance

---

[143] Kaplan, p.523; italics in the original. I change "occurrence" to "utterance" in what follows.
[144] See p. 518 for "providing a synonym".

of a sentence. An utterance we agreed to analyze into the pair consisting of the sentence uttered and the circumstances of utterance, which we collectively called context:

(50) $Utterance(x) \leftrightarrow \exists s \exists c(Sentence(s) \wedge Context(c) \wedge x = \langle s, c \rangle)$.

By the vague term "circumstance of utterance" what is meant is any fact about the occasion of utterance which might affect the truth value of the uttered sentence. In the case of indexical theory, these are facts about the reference of indexicals.

On the conversation heuristic, the speaker and the time determine completely all the other contextual facts about the utterance: the place of the utterance ("here") is the place of the speaker at the time of utterance; the addressee ("you") is whomever the speaker is addressing at that time; etc. Consequently, it seems we can represent contexts simply as pairs consisting of speakers and times:

(a) **The ordered-pair definition:**
$Context(x) \leftrightarrow \exists s \exists t(Speaker(s) \wedge Time(t) \wedge x = \langle s, t \rangle)$.

As before, if $c = \langle s, t \rangle$, we put $speaker(c) = s$ and $time(c) = t$.[145] All other atomic indexicals such as "here" and "you", as well as indexical expressions such as "my wife" and "the town I was born in", are then definable in terms of speaker and time:

(51) $|here|^c = \iota x.in(speaker(c), x, time(c))$,

(52) $|my\ wife|^c = \iota x.Wife(x, speaker(c), time(c))$, etc.

This approach, which we can call *the ordered-pair approach* to contexts, is essentially Kaplan's. Actually, Kaplan's contexts include also a place and a possible world component, so they constitute *centered worlds*.[146] I will stick to speaker-time pairs for simplicity.

What makes the speaker and the time of utterance the center of a context, in terms of which all the other indexicals can be defined? The answer is that an utterance is an act, and the speaker-time pair constitute the *agent (in the narrow sense)* of the act. I say "in the narrow sense" since usually we would identify the agent with the speaker only. But this would not determine the context, as the same speaker can be associated with several utterances. An agent in the narrow sense is the speaker *qua* agent of this very utterance, i.e. a speaker at a time, or a time-slice, or a stage, of a speaker (on the assumption that a speaker makes at most one utterance at a time). The speaker, which is often associated with several utterances, we call the *agent in the proper sense*.[147]

---

[145] "$Speaker(x)$" is a predicate, and the uncapitalized "$speaker(c)$" is a function; and likewise for "$Time(x)$" and "$time(c)$".
[146] See, e.g., Liao (2012) on centered worlds and their problems.
[147] The notion of speaker in the proper sense will not play a part in what follows, but it's good to have it in mind to compare with agent in the narrow sense.

The relation between a context and an agent can be brought out by comparing contexts to indices. Indices are glossed as circumstances of evaluation, contexts as circumstances of use. Take, for example, the sentence "it is raining". This English sentence can be *evaluated* against a time and place in which no English speakers, or even any speakers, exist. But it cannot be *used* in such a time and place. A context, unlike an index, has to have an agent, a linguistic subject *on the scene*. Let's formulate this insight as our first discovery about contexts:

(b) **The agent on the scene insight:** A context is centered around a linguistic agent.

5.3.3. From speaker-time pairs to linguistic agency

As it stands, the ordered-pair approach cannot serve as a theory of contexts even for the Kaplanian semantics of §5.2.5. The reason is that if we understand contexts as speaker-time pairs, we lose the power to define pragmatic validity. For let $c^* = \langle Obama, 1900 \rangle$, and accept that Obama was born in 1961. Our test case for pragmatic validity is:

(53) I am here now.

But now:

(54) $|(53)|^{c^*} = 0$.

From which it follows that:

(55) $\neg PV((53))$.

So on the ordered-pair approach, L^P_{Ipr} fails clause (a) of the direct reference test.

Even if the idea of a speaker at a time captures what it is to be a linguistic agent, a speaker-time pair doesn't capture the idea of a speaker at a time. This was noticed by Kaplan.[148] His answer, which as far as I can tell has not been improved on, is to limit the contexts to those speaker-time pairs in which the speaker occupies some (and therefore one) spatial position at the time of the context. This means a revision of our definition of context:

(c) **The naïve situatedness definition:**
$Context(x) \leftrightarrow \exists s, t(Speaker(s) \wedge Time(t) \wedge x = \langle s, t \rangle \wedge \exists y(in(\mathrm{s}, y, t)))$.

The condition that was added corresponds to Kaplan's Clause no. 10 in his definition of an interpretation (model) of a pragmatic object-language.[149] Pairs that conform to (c) are called *proper contexts* in the

[148] Kaplan, p. 509.
[149] p.544. His formulation is superficially different because his contexts include place and world parameters.

literature, [150] and pragmatic validity is defined relative to the collection of proper contexts. With this modification, "I am here now" turns out pragmatically valid, as desired. Let's call this refinement *the proper context approach*.

The situatedness definition (c) is an improvement on the original pair definition (a) since it accounts for the example of the first clause of the direct reference test. But as proper contexts are still standard objects, by the argument in §5.2.5, it fails the second clause. The reason that it does, I submit, is that a speaker-time pair doesn't capture the philosophical notion of linguistic agency, even if we can be sure that the specified speaker occupies a place at the specified time. To see this, observe that it doesn't even pass the first clause, though it does account for the example. The added condition in (c), the situatedness condition, is blatantly ad-hoc. In order for (53) to come out true in all contexts, we added a condition which amounts to saying that contexts are such that (53) should come out true in all of them. Other kinds of pragmatically valid sentence would require adding further conditions to the definition. To wit, Kaplan adds another such clause (no. 9) for the sentence "I exist". And if there are further pragmatically valid sentences, for example "I am speaking now", they would need further special conditions.[151] Effectively, Kaplan's procedure amounts to including all of the pragmatically valid sentences, more precisely their translations in the metalanguage, into the definition of context. It sheds no light on what it is that makes them pragmatically valid in the first place. Such a definition offers no substantial elucidation of the notion of linguistic agency.

What is it that makes (53) pragmatically valid in the first place? Why accept that the speaker always occupies the place of utterance at the time of utterance? This is not a particularly deep or incontrovertible fact; it is a consequence of our agreement to limit our investigation to spoken conversations between humans, what we called "the conversation heuristic" in §5.2.3. But if this is a mere heuristic, then we shouldn't include it in our *definition* of context, (c). Rather, we should think of it as a (strictly speaking false) *theory* of contexts, one that we accept only as a simplifying assumption, a heuristic:

(d) **The naïve situatedness heuristic**: $in(speaker(c), place(c), time(c))$.

You might want to object that (d) is no less ad-hoc than Kaplan's clause no. 10. I agree. But now it is clear that the ad-hocness belongs to the conversation heuristic and not to indexical semantics in principle.

The conversation heuristic is just one possible simplifying theory that we can adopt. We can also imagine other sets of axioms expressing different simplifications that help us study indexicality from

[150] E.g. in Schlenker (2011, p.1565) and Predelli (2011, p.291).
[151] Since proper contexts, on the present definition, include contexts in which the speaker is silent, "I am speaking now" does not come out pragmatically valid, though on the conversation heuristic it should.

other angles. What is crucial is not to mistake these heuristics for genuine analyses of the concept of context. Indeed, it has long been recognized that the negation of (53) sometimes gets to be used truly, if we include non-conversational instances of language use such as recorded messages and written notes. (53), it turns out, is not pragmatically valid after all.[152] Generalizing even further we see that even assuming a speaker and a time is but a simplifying heuristic, at least if by a speaker we mean a human individual. After all, we shouldn't on philosophical grounds rule out the possibility of language use by collectives, angels, machines, or what have you, in time or outside of it.[153] What seemed like a definition of context in (a) turns out too to be a mere heuristic:

(e) **The ordered-pair heuristic:** $\exists a \exists t(Speaker(a) \wedge Time(t) \wedge c = \langle a, t \rangle)$.

Mistaking the heuristics (d) and (e) for metaphysical truths is what obscured the source of the failure of Kaplan's system to be an adequate theory of direct reference. Once we drop the heuristics, definitions (a) and (c) cease to be applicable, and all that we are left with is the informal insight in (b): that contexts are centered around linguistic agents. This is the insight that should guide us in looking for a new conception of context. But we know by now that the search should not be for another logically standard kind of object to represent the linguistic *agent*. We are looking for a basic understanding of linguistic *agency*, or subjectivity.

5.3.4. Linguistic agency and the abstractness of characters

Linguistic agency is the capacity for language use. Languages, according to the definition in §1.3, consist of phonology and semantics. A phonological string ceases to be noise and becomes linguistic once it is endowed with meaning, or content. Linguistic agency is the capacity to endow phonological strings with meanings, or in other words, the capacity for *content generation*. To say that a context contains an agent essentially is to say that it is a site for content generation.[154]

The word "generation" denotes, literally, an action or process that takes place at or over a certain time. To say that something is generated is to say, among other things, that it does not exist prior to the time of generation. But this is only metaphor. Generation for us can't mean, say, some psychological or

[152] Kaplan mentions this at p.491fn but waves the problem away. See Predelli (2011) for a treatment and a good literature survey. Predelli, as well as many of the authors he cites, basically accepts the pragmatic validity of (53) and Kaplan's definition of proper context. In order to account for the occasional truth of "I am not here now", he modifies the semantics so as to include improper contexts. On my view, once we look at language use more generally, the notions of pragmatic validity and proper contexts shift accordingly and there is no place for improper contexts.

[153] Kaplan (p. 553), as a "note on possible refinement" of his framework, considers disembodied (though not non-existent) linguistic agents.

[154] See Kaplan (1989b), esp. ch.3, for contexts as sites for content generation.

neurological process carried out in the heads of individual speakers. Making pronouncements on psychological processes requires empirical observation, and we are here working from the armchair.[155] The object of the present investigation is semantics, so we need to cash out the temporal metaphor in semantic terms. To say that content is generated, I repeat, is to say that it does not exist prior to the moment of generation. Our task in this section is to de-metaphorize "prior".

Let's deal first with intensional languages (as developed in §5.2.1). In intensional languages, with every well-formed expression there is associated an object, an intension, as its content. Intensions are mathematical functions from a domain of indices (times, possible worlds, etc.) to extensions. In intensional semantics, however, content is not generated. It is there, so to speak, to begin with. We can say that the content of a non-indexical expression is a *standing fact*. The notion of content generation applies specifically to indexical expressions. It is their contents that are "generated" by the agent in the context; in other words, it is their contents that do not exist "prior" to the context. For example, the pragmatic-intensional language L_{pr}^{PI} (§5.2.5) contains the term "I". If Trump is the speaker in some context c_2, then the content of "I" in c_2 will be a constant function from times to Trump, or simply Trump:

(56) $$|I|^{c_2} = \lambda t.speaker(c_2) \cong Trump.$$

To say that contents are generated in context is to say that the content $|I|^{c_2}$ does not exist outside c_2. The idea here is obviously not that Trump himself does not exist in the time preceding $time(c_2)$. The sense of "prior" that we are seeking is logical or semantic, and has nothing to do with the context feature $time(c)$. The point is that *the possibility of referring* to Trump using "I" in c_2 depends only on properties of c_2, and in particular on the identity of $speaker(c_2)$. It therefore does not depend on the standing domain of L^P: for "I" to refer to Trump, the domain of L^P doesn't have to have Trump as member.[156] This is what it means for the content of "I" to be generated in c_2.

But here is the sting. Let's make two observations. First, indexicals have characters, and the characters of indexicals, unlike their contents, do not depend on the context of utterance. Like the contents of non-indexicals, they are ungenerated standing facts about L_{pr}^{PI}. Second, on the standard Kaplanian framework presented in §5.2.5, the character of an indexical expression is a mathematical function from contexts to contents:

(57) $$|I| = \lambda c.\,|I|^c = \lambda c.speaker(c).$$

---

[155] Working from the armchair is something different from the (perhaps empirical) method of psychological introspection. We are engaged in the former.

[156] In our case, since L^P also happens to have an individual constant "b" for Trump, then Trump is also in its standing domain. But this isn't necessary.

Mathematical functions are such that their existence implies the existence of their arguments and values. This means that where the character $|I|$ exists, so do its values. One of its values is the content $|I|^{c_2}$. But then, putting these two observations together, $|I|^{c_2}$ exists independently of ("prior" to) c_2. In other words, it is *not* generated in c_2, contrary to what was said in the previous paragraph. This points to a basic incongruity between the philosophical idea of content generated at a context, and the technical apparatus of representing characters in terms of standard mathematical functions. Kaplan thinks of this apparatus as a matter of convenience,[157] but in this case the tool is not apt to the task. Using a mathematical function introduces a subtle fallacy of reification into the framework, and subtle though it may be, it is opens a crack wide enough for monsters to crawl in.

If we want to do justice to the idea of linguistic agency as content generation, we need to find a different way to think of characters. Such a way is close at hand. Recall the difficult passage from Kaplan, quoted in §5.3.1. Kaplan there describes characters as "rules specifying the referent" of an indexicals, without "providing a synonym". Recall also Strawson's words quoted in §5.2.3, about an expression's meaning (in the sense of character) being "*general directions* for its use to refer". Following these two leads, I suggest that we think of characters as belonging to the domain of the practical: as practical directives, or imperatives, or instructions for an agent how to generate content. The assumption here is that practical directives do not by themselves involve theoretical, and in particular ontological, commitments (except by virtue of heuristics of the kind sketched in §5.3.3). It is only when in fact carried out by an agent that such directives serve to generate ontological commitment.[158]

Our starting point (§5.3.2) was the contrast between circumstances of evaluation (indices) and contexts of use. The basic insight (b) was that contexts involve an agent in some essential way that indices don't. This contrast is reflected in the difference between intensions as mathematical functions and characters as practical directives. A directive, unlike a theoretical statement, requires an agent on site to carry it out. The existence of a directive doesn't automatically imply the existence of an agent which can implement it, or the existence of the product of such an implementation. In systems in which this property of directives is important, a mathematical function is not a good model for them. If we take characters (e.g. (57)) to be practical directives and not mathematical functions, then it is not the case that their existence implies the existence of the contents (e.g. (56)) that agents generate by following them. This fact, that characters do not imply the existence of the objects that they determine, we can call *the abstractness of characters*.

---

[157] Kaplan: "it is convenient to represent characters by functions from possible contexts to contents" (p.505).
[158] The fact that practical directives don't carry ontological commitment is not often explicitly recognized in philosophy, but it does surface occasionally. For example, the practical nature of Euclid's geometrical postulates is sometimes interpreted in this way when compared to Hilbert's corresponding axioms. See Mueller (1981, p. 14f); compare Chihara (2003).

In making sense of content generation I relied on the intensional approach to content (§5.2.1). In §5.2.2 I made some effort to convince you that this approach is inadequate. What does content generation mean for extensional languages? On such languages we can no longer say that with every expression there is associated an object as its content, and so there is no obvious candidate for being the "generated" content in the case of indexicals. But we can still make sense of content generation, and in particular of the abstractness of characters, in terms of direct ontological commitment to the arguments and values of intensions. In these terms, to say that characters are abstract and that content is generated in context is to absolve the metalanguage M^P of ontological commitment to the arguments and the values of the intensions that would have served as contents on an intensional counterpart. Formally, with every context c we associate a domain D^c of objects, reference to which is generated in c. To say that characters are abstract is to say that we can't assume these objects to be part of the standing domain of M^P. This yields the same result as before: "I" in L^P at c_2 can refer to Trump even if Trump is not a member of the domain of M^P.

This implies the following, perhaps surprising, change in the logic. if the referent of an indexical α does not exist D^{M^P}, then a fortiori it does not exist in D^{L^P}. The quantifiers of L^P, however, range over its standing domain. This means that the law of existential generalization $\ulcorner \phi(\alpha) \rightarrow \exists x \phi x \urcorner$ does not hold in L^P for α an indexical. Indexicals call for a kind of free logic, though not for the reasons for which free logic was originally devised.

The abstractness of M^P has one consequence which is particularly important to our on-going concern (the concept of truth and the language of Convention T): Truth for L^P cannot in general be defined in M^P. Consider the T-sentence in M^P for the L^P sentence "I am president now":

(58) $\quad |P(I, now)|^c = 1 \; iff \; speaker(c) \; is \; president \; at \; time(c).$

The significance of the fact that characters are abstract is that the expressions "$speaker(c)$" and "$time(c)$" do not refer in the usual sense to the speaker and the time of c. In order for the indexicals "I" and "now" to refer in c_2 to Trump and to 2017, it is not necessary for Trump and 2017 to be objects in the domain of the metalanguage M^P. The expressions "$speaker(c)$" and "$time(c)$" are unsituated instructions directed at a prospective situated agent, telling that agent how to generate content, in terms of the feature of the agent's situation; in Kaplan's terms, they are instructions on how to "fix the referent" of "I" to the speaker of the context. From this it follows that the sentential character "$speaker(c) \; is \; president \; at \; time(c)$" is not a judgment, but only an instruction for making a judgment in context; in Strawson's words, it is a "general direction for the use" of the sentence "$P(I, now)$" in making an assertion. A pragmatic semantics in M^P for an object-language L^P does not imply a truth predicate. We say that M^P is an *abstract metalanguage* for L^P.

5.3.5. Abstract objects

The aim of our investigation is to answer the three questions stated in §5.3.1B. The metaphysical question (a) was:

> What kind of thing is a context, if it isn't an object in the standard sense of the word?

The (meta-)semantical question (b):

> What is the relation between the parameter c and contexts?

With the notion of abstractness (of M^P) in hand, we are ready to provide answers.

The concepts to be developed as answers are those of an abstract object and of abstract reference. Abstract objects are to be distinguished from *real objects*, which are the usual objects found in domains of standard languages and in context domains D^c. Before we look into abstract objects, we need to have some characterization of real objects before our eyes. A comprehensive theory of the concept of object is well beyond the scope of this work. I'll say just what I need to say to make my meaning clear. Briefly, I take it to be constitutive of real objects that they can be individuated, and individuation depends on what I will call *unity* and *constitution*, or *constitutional distinctness*. An object can be individuated relative to a language or a context only if both its unity and its constitution are given in that language or context. Abstract objects are such that their unity and constitution come apart.[159]

As I've stressed a while back, I'm here speaking of real objects in a logical, not a metaphysical, sense. Many metaphysically different kinds of things are real objects for some language or another: planets, wars, numbers, concepts, possible worlds, etc. A real object is something that is referred to by a singular term, i.e. by a variable or by an indexical. Singular terms appear in two kinds of atomic formulas: predications $\ulcorner P(\alpha, \beta, \ldots)\urcorner$ and identity statements $\ulcorner \alpha = \beta \urcorner$. These linguistic forms provide a clue to the notions of unity and constitution.[160] They are slippery notions and I will not explicate them in any generality. Let me limit the discussion to composite objects, and in particular to *container objects* such as tuples, since other composite objects can usually be reduced to them.[161] The simplest container objects are collections, and the difference between unity and difference is best displayed with relation to them.[162] The *unity* of a container object is the fact that a concept can be predicated of it (or withheld

---

[159] The issue is intimately related to, but not the same as, the issue of *identity*; also the notion of a *sortal concept*. See Lowe (2017) for a brief but lucid and (relatively) comprehensive introduction, and references therein.
[160] Lowe (2017) speaks of a *principle of individuation* of objects as a combination of a principle of unity and a criterion of identity (p. 993). He doesn't link unity to predication as I do.
[161] If there are objects which are not composite, it might be that unity and difference cannot come apart for them.
[162] The unity of collections is sometimes referred to as *totality*.

from it) as of a single thing, disregarding the plurality of its components. For example, the collection $A = \{Kennedy, De\ Gaulle, Obama, Berlusconi\}$ has never been a president, though every one of its members have been. In this way the predicate "president" applies to A's members, but not to A (and vice versa for its negation). Collection unity is the simplest, and consists merely in taking the plurality of a collection's members as a single object, i.e. in denoting it using a singular term. The *constitution* of an object generally is that by which it can be distinguished from and counted with other objects. For collections, constitutional distinctness amounts to a difference in their memberships and nothing else. For container objects more complicated than collections, such as tuples, we would include *organization* as well, i.e. the order of the members.

To say that unity and constitutional distinctness, or *individuation* for short, constitute objecthood, is to say that they ground the existence of the individuated objects in a domain of discourse. In other words, if for some principled reason a language L cannot express unity and constitution for some objects, then these objects are not members of D^L. For example, proper classes with respect to the language of ZF set theory L_{ZF} lack the appropriate unity in L_{ZF}, and therefore are excluded from membership in the domain of L_{ZF}. This is in spite of the fact that the constitutional distinctness of proper classes *is* given in L_{ZF}: every one of the members of a proper class is a member of L_{ZF}'s domain.[163]

Initially, we would like to say that if the members of a collection are not objects in a domain D^L, then neither should the collection be. But we can stipulate cases to the contrary. For example, let $a = \{0,1\}$, $b = \{0,2\}$, and let L be a language with $\in$ as the sole non-logical predicate (with the usual interpretation), and let $D^L = \{0, a, b\}$. Then a and b are distinct in L, though the difference in their membership is not given in L. This seems to show that a difference in membership is not sufficient for constitutional distinctness, or that constitutional distinctness is not sufficient for individuation. But this is not the case. All that this example shows is that being able to *express* a difference in membership between two sets in a language L is not necessary for expressing the distinctness of the sets in L. The distinctness of a and b relative to L was a fact stipulated by us, and the stipulation *was* grounded in a difference of membership, although this difference was expressed not in L. A distinction between two collections must always be grounded by some difference in membership, be it expressible in the language under discussion or not. By contrast, if a distinction can *in principle* not be made between two objects of a given kind, *only then* should we say that they cannot be counted as objects in a given domain. If you can't be counted, you can't be counted as an object.[164]

---

[163] Cantor famously distinguishes between consistent and inconsistent multiplicities. The latter are such that it is impossible to think of them "as a unity, as 'one finished thing'" (Cantor's letter to Dedekind, in van Heijenoort 1967, p. 114).

[164] Urelemente in set theory apparently exhibit such a distinction without a difference. But this is only apparent. Urelemente are needed in set theory if set theory is used, not as foundations for mathematics (in which case Urelemente are disposable), but as a theory of collections of arbitrary (real) objects. The theory of collections

To sum up, objecthood for container objects amounts to membership and unity, and, sometimes, also organization. If for some principled reason either membership or unity for a certain collection a is not given in a language L or a context c, then a is not a member of, respectively, D^L or D^c.

Let's apply these remarks to the case at hand. Objecthood, we say, depends on unity and constitutional distinctness. Consequently, the absence of unity and constitutional distinctness implies failure of objecthood. Usually, we would leave it at that. But our recent notion of an abstract pragmatic metalanguage, coupled with contexts as sites for content generation, allows us to conceive of a special situation in which unity and constitution for a certain kind of object are both available, though in different languages. This gives rise to the notion of an abstract object, as follows.

Contexts can be regarded as containers. To see this, consider that the only role of contexts for us is to provide reference for the indexicals. Consequently, we can think of a context c as a container for all of the objects that stand in contact with the agent, i.e. as a long tuple of context features.[165] Does the tuple need to be so long? Can't we define all its members in terms of speaker and time, as we did in §5.3.2?:

(59) $place(c) = \iota x.\, in(speaker(c), x, time(c))$,

(60) $addressee(c) = \iota x.\, (speaker(c)\ is\ addressing\ x\ at\ time(c))$, etc.

According to the direct reference thesis, we shouldn't be able to. The problem with these definitions is that, since they are formulated in M^P, they imply that the referents of the indexicals exist in the domain of M^P. But this contradicts the abstractness of M^P, according to which the members of D^c are not (necessarily) members of the domain of M^P. Therefore, the context tuples have to include all context features and not just speaker and time.

The objecthood of tuples depends on their membership (in order) and their unity. However, in the case of contexts these two things come apart. Let c_1 be some context. Then c_1 is the tuple consisting of all the things that can be referred to (indexically) in c_1. In other words, c_1 consists of all the things that have both unity and constitutional distinctness relative to c_1. But then, if we accept the principle according to which a container cannot contain itself, then c_1 itself is not one of those objects. Consequently, c_1 does not have unity and constitutional distinctness relative to c_1. There will therefore not be an indexical α such that an utterance of α in c_1 refers to c_1. Now trivially the constitution (ordered

---

abstracts from the particular individuals to be collected, but it doesn't want to abstract from their individuation, since its main use is in counting them in advanced way, and individuation is required for counting. Once we turn to apply set theory to a specific domain as a calculus of collections, say when we want to count the real numbers, then we have to bring their differences into account. See Potter (2004, §3.8) for this point of view.

[165] "Contact" is the term I used non-commitally for the relation between objects and an agent that underlies the direct reference of indexicals, see §5.2.4.

membership) of c_1 is available in c_1: we can refer in c_1 to every member of c_1 using the right indexical. It is therefore the unity of c_1 which is missing in c_1.

In M^P we get the converse. M^P is used to state characters, expressions such as $speaker(c)$, which refer to the speaker of the context. The context parameter c is a singular term that refers to contexts in M^P. So the unity of contexts *is* given in M^P. However, as we saw in §5.3.4, the members of the different contexts, for example their speakers, are not (in general) members of the domain of M^P. In other words, the constitution of contexts is unavailable in M^P.

The picture we have is that contexts do have both unity and difference, but not together in the same language or context. The membership of a context c_1 is given in c_1, while its unity is given in the abstract M^P. I call this kind of object an *abstract object*. The context parameter c can be called an *abstract variable*, and the relation between c and the contexts *abstract reference*. This goes also for characters: the expression "$speaker(c)$" refers abstractly to speakers of contexts. The concept of reference, like the concept of object, undergoes a kind of fission: abstract reference pertains to the character of an indexical as the rule of reference; direct reference pertains to indexicals in context and yields the actual reference. Abstract reference, on this view, is the complement of direct reference.

The notions of abstract objects and abstract reference are our respective answers to the metaphysical and the metasemantic questions from above. In the next section I will lay out the formal properties of the abstract variable, and they will provide an answer to the third question. Before we proceed, let me make two more points about abstract objects.

That a context c_1 is an abstract object means two things: that it is not an object in its associated domain D^{c_1}; and that it is referred to abstractly in M^P, through the abstract variable "c". But then it is only an abstract object *relative* to the context c_1 (itself) and the abstract variable "c". c_1 might, *per accidens*, also be a *real* object relative to some other domain. For example, nothing tells against the existence of the context c_1 in the standing (conceptual) domain of L^P (and consequently also in the domain of M^P), referred to by some non-indexical individual constant. Or there might be a context c_2 that contains c_1 as a member, so that c_1 can be referred to directly, through an indexical, in c_2. To say that contexts are abstract objects is not to say that they are a new kind of object, but that they are an object in a new way. Maybe it is better to say that they are *objects abstractly*. The fact that contexts can also be referred to in the standard way we can call *the reality of contexts principle*.[166]

[166] Recall the *reality of indexicals* principle from §5.2.3.

We need some way to speak about the content that is "generated" in a context. In §5.2.2 the mediated approach to content was proposed as a way to represent content for extensional languages. The idea was that the content of an expression α in a language L will be referred to by the statement, in a metalanguage for L, of the semantic value of α. But in the case of indexicals we can't use such statements from the abstract metalanguage M^P, since real reference and truth-conditions for L^P expressions are not available in it (see the closing of §5.3.4). However, by our recent reality of contexts principle we may assume that for every context c there is some non-abstract metalanguage M^c, the domain of which will include both D^c and the standing domain of L^P. In M^c we can formulate real content statements for L^P at a context c. I call M^c a *surveying metalanguage* (for c).

5.3.6. The abstract variable

The third guiding question in §5.3.1B was the formal question:

> What are the formal properties of the abstract variable c?

The task now is to give the syntax and the semantics of M^P in light of our recent discussion, and especially in light of the developments of the concepts of *character* and *context* from the two previous sections. In addition, we want to see whether it passes the direct reference test from §5.2.4 (spoiler: it does).

The usual variable in standard languages expresses real objecthood, which by the previous section implies unity and constitutional distinction. Unity is expressed by the fact that the variable is a singular term, a potential subject of predication. Constitutional distinctness is made available by the existence of several different variables x_1, x_2, etc. For instance, in a binary predication $\ulcorner\phi(x_1, x_1)\urcorner$ the two relata have to be the same individual, while in $\ulcorner\phi(x_1, x_2)\urcorner$ they may be distinct. In this way the expression of the possibility of distinctness (though not of its actuality) is built into the syntax, and even the notation, of the variable.[167] By contrast, the notion of abstract object is such as to allow unity but, in principle, no distinction. There should be no syntactic possibility of using two different abstract variables in the same sentence, e.g. to refer to two distinct contexts c_1 and c_2. There is only one abstract variable. Accordingly, I make a notational modification and, instead of the Latin lowercase "c" evocative of the usual variable, use the symbol "@".[168]

---

[167] See Fine (2003) for a discussion of this point which, however, leads Fine to a quite different place (to Fine (2007)).

[168] This symbol is sometimes used to denote the actual world (e.g. Menzel (2006)). Lewis (1970) argues that "actually" is an indexical term referring to the world of the context (more or less; in 1970 Kaplan's distinction between index and context was still brewing); see also Lewis (1986), Kaplan (1989b, p.594f). Notice that if "@" indeed refers to the actual world, and if actuality is indeed indexical, then "@" can only refer to the actual world abstractly (in my sense). This is because the metalanguage, both in Lewis and in Kaplan, contains no indexicals.

Apart from the abstract variable itself, M^P has terms that refer to the context features. Which features we can say a context has depends on our full metaphysical theory of contexts and linguistic agents. We don't have a full theory. We do have a basic characterization of contexts according which contexts are centered around linguistic agents; and a characterization of linguistic agents according to which what they do is endow strings with content. Minimally then, M^P should have a term for the function of endowing strings with content, i.e. for a semantic interpretation function $|s|^{@}$ with one standard variable ranging over strings and one abstract variable over contexts. This expression represents the use of the string s by the agent of the context @. Beyond the basic characterization of contexts, we also indulged in certain heuristic assumptions about contexts, in particular their possession of certain features such as speaker, time, etc. With this in mind we can let M^P contain the singular terms $speaker(@)$, $place(@)$, $time(@)$, etc. These, together with the real variables, are the only genuine singular terms of M^P (ignoring the term "1" in locutions such as $\ulcorner|s|^{@} = 1\urcorner$). The syntactic theory of M^P (the definition of wff) is exactly like that of standard languages (see §1.3.3), except that atomic formulas are now defined to include abstract singular terms. (There is no quantifier over @.)

So much for the syntax of M^P. The semantics is a little more complicated, since in M^P, unlike in standard languages, the norm of the language doesn't converge with the central concept of its theory of meaning. I'll explain. Standard regimentation is inherently assertoric in force, and the constitutive norm of assertion for it is truth. Note that by talking about assertion I don't mean a speech act. The main purpose of standard regimented languages is to express theories, and this is done in the assertoric or indicative mood. A theory is successful if it is true. This is the sense in which the norm of standard languages is truth. It so happens, as we've seen, that the concept of truth (or satisfaction) is also the central concept of the theory of meaning for standard languages. This is, of course, no mere accident. The concept of truth is constitutive of standard languages.[169]

But with M^P things are less comfortable. First, sentences in M^P are not truth-bearers. This is because M^P is not assertoric in force. Sentences in M^P in general are rules for generating (assertoric) content, and rules are not assertions. Therefore truth is not the norm of M^P. Second, since M^P sentences don't have truth-conditions, there's no sense in thinking of truth conditions as their meanings. Some other concept will have to fulfill that role. Finally, there is no reason to assume at the outset that there will be a single concept that fulfills both duties, of being a norm and of constituting meaning, in the way that truth does for standard languages.

[169] Dummett (1959) is the classical place for the connection between truth as the norm of assertion (its success condition) and truth as constitutive of meaning. See also §2.3.2.

The meaning theory for M^P is a statement of what happens when a sentence of M^P is applied to a context. I call it *application semantics*.[170] An M^P sentence (a rule) is applied to a real, non-abstract, context (is acted on by the context's agent), to yield an item of content. I'll present application semantics by showing it for a particular context instead of by giving general definitions. Let c_1 be a context in which the speaker is Obama ($= a$), the place is the White House ($= h$), and the time is 2016 ($= t_1$). I use square brackets to denote the "action" of applying an M^P expression, and a paired arrow subscripted with the context's name to signal the result of the application.

(I) **Application semantics for M^P in c_1**

(a) $[@] \rightrightarrows_{c_1} c_1$;

(b) $[speaker(@)] \rightrightarrows_{c_1} a$, and likewise for all other context features;

(c) For any string s, $\left[|s|^{@}\right] \rightrightarrows_{c_1} |s|$;

(d) If $\ulcorner \phi(\alpha_1, \ldots \alpha_n) \urcorner$ is a sentence containing abstract singular terms $\alpha_1, \ldots, \alpha_n$, then $[\phi(\alpha_1, \ldots, \alpha_n)] \rightrightarrows_{c_1} \phi([\alpha_1], \ldots, [\alpha_n])$.

Think of the paired arrow as symbolizing the generation of content. On its left-hand side, in square brackets, is written an abstract expression (a character). Subscripted to the right of the paired arrow, the context in which the character is applied is designated. On the right-hand side is a statement of the generated content. The statement of L^P content at c_1 is expressible neither in M^P nor in L^P, but only in a surveying language M^{c_1} (see the last paragraph of §5.3.5). The right-hand sides of the application clauses are therefore expressions in M^{c_1}. In words, clause (b) says that the abstract expression "$speaker(@)$", applied to the context c_1 in which Obama is the speaker, generates reference to Obama; clause (c) says that the abstract expression $\ulcorner |s|^{@} \urcorner$, applied to c_1, generates reference to the reference of s; clause (d) implies, for example, that the abstract expression "$|I|^{@} = speaker(@)$", applied to c_1, results in the identification of the reference of "*I*" with Obama, or in other words in the endowment of the string "*I*" with its meaning: $|I| = a$.

This is the meaning theory for M^P. What about the norm of M^P? We can tell that the norm of standard languages is truth by the fact that truth is what characterizes successful theories, i.e. sentences that we endorse. Likewise, we will get a clue about the norm of M^P if we inspect the sentences of M^P that we endorse. Sentences that we endorse I call, for now, *correct*. I write $\ulcorner M^P \Vdash \phi \urcorner$ to say that $\ulcorner \phi \urcorner$ is correct in M^P. The first group of correct sentences constitute the pragmatic theory for L^P:

(61) $M^P \Vdash |I|^{@} = speaker(@)$,

[170] Strictly speaking, it is neither a semantics nor a theory.

(62) $M^P \Vdash |a|^{@} = a,$

(63) $M^P \Vdash |P(I, now)|^{@} = 1 \leftrightarrow P(speaker(@), time(@)).$

These are the sentences that ascribe characters to indexicals. They are the analogues to sentences such as "$|Ws| = 1 \leftrightarrow snow\ is\ white$" in standard semantics. To say that (61) is correct (and not, say, "$|I|^{@} = time(@)$") is to say that "I" refers to the speaker of the context (and not to the time of utterance). The pragmatic theory consists of all sentences of the form:

(a) $\boldsymbol{M^P}$ **semantics:** $M^P \Vdash \ulcorner |\phi| = 1 \leftrightarrow \psi \urcorner$.

where ϕ is well-formed in L^P and ψ is like ϕ except that all indexicals are replaced by the corresponding context features, i.e. "I" by "$speaker(@)$", etc.[171]

I use the symbol $\vDash$ to denote truth in non-abstract languages. For example, I write $\ulcorner M^{c_1} \vDash \phi \urcorner$ to mean that $\ulcorner \phi \urcorner$ is true in the surveying language M^{c_1} ($\ulcorner \phi \urcorner$ is true *about* the context c_1). For example, the following sentences are the results of applying (61)-(63) to c_1 using (I):

(64) $M^{c_1} \vDash |I| = a,$

(65) $M^{c_1} \vDash |a| = a,$

(66) $M^{c_1} \vDash |P(I, now)| = 1 \leftrightarrow P(a, t_1).$

When we apply an M^P expression to a context, the result is expressed by a sentence in the metalanguage surveying the context.[172] We can abbreviate uses of (I) by the following principle:

(b) **The Commitment Rule:** $M^P \Vdash \phi \Rightarrow M^{c_1} \vDash [\phi]$.

In words: if a sentence is correct in M^P, then the result of its application to a context is a true sentence (in the surveying metalanguage, about the context). In particular, the contrapositive tells us that if an application results in a falsity, then the applied sentence is incorrect. Consequently, the application of a correct M^P sentence will always result in true sentences. We therefore say of correct sentences that they are *a-priori*. This is the norm of M^P.

This is one group of a-priori sentences: the pragmatic theory. Two further groups are the sentences belonging to phonological theory and to logical theory. Since contexts are essentially centered around agents, and agents essentially are that which endows strings with meanings, phonology is an essential part of a context. It's not that the domain of the context contains strings necessarily. The domain of the context contains the objects that the agent of the context can refer to directly. Strings are not necessarily such objects. Strings are the objects *through* which the agent refers to whatever he or she refers to.

---

[171] This is, if you like, the T-schema relativized to indexicals.
[172] We are adopting the mediated approach to content of §5.2.2.

Phonological facts are therefore facts *about* the context, not facts expressible *within* the context; as such, they are facts expressible in the surveying language. To say that these facts are a-priori is in effect to assume that all agents share a phonological constitution.[173]

That logical theory is a-priori is obvious and I will not elaborate on it. For later use it is important to put down the principle of substitutivity of identicals. In M^{c_1} we get, as a special case, the substitutivity of coreferring terms:

(c) **Substitutivity of coreferring terms:** $M^{c_1} \vDash |\alpha| = |\beta| \Rightarrow M^{c_1} \vDash |\phi(\alpha)| = |\phi(\beta)|$.

The form that this principle takes here is of importance for what follows.

Finally, if we accept certain methodological heuristics (see §5.3.3), then the following too is a priori:

(d) **Naïve Situatedness Heuristic**: $M^P \Vdash in(speaker(@), place(@), time(@))$.

An application of this yields:

(67) $M^{c_1} \vDash in(a, h, t_1)$.

Summing up, the a-priori sentences are: the pragmatic semantic theory for L^P; phonological theory; logical theory; and the methodological heuristics.[174] They are a gerrymandered group, and to that extent the norm of M^P is less a coherent unity than truth is for standard languages. I leave discussion of that to a later date and proceed to show that this semantics passes the direct reference test.

Clause (a) of the test requires to define the pragmatic validity predicate. We say that a sentence s of L^P is pragmatically valid if it is true a-priori, i.e. if $\ulcorner M^P \Vdash |s|^@ = 1 \urcorner$. To see that this works, let's show that "I am here now" is pragmatically valid and "Obama is here now" isn't:

1.	$PV(s)\ iff\ M^P \Vdash |s|^@ = 1$	Definition
2.	$M^P \Vdash in(speaker(@), place(@), time(@))$	(d)
3.	$M^P \Vdash |in(I, here, now)|^@ = 1 \leftrightarrow in(speaker(@), place(@), time(@))$	(a)
4.	$M^p \Vdash |in(I, here, now)|^@ = 1$	2,3
5.	$\boldsymbol{PV}(\text{"}\boldsymbol{in}(\boldsymbol{I}, \boldsymbol{here}, \boldsymbol{now})\text{"})$	1,4

[173] A sensus communis?

[174] In L^P we also had non-indexical expressions (P, a, b), and this means that in M^P we have their non-abstract translations. Since true sentences involving them will generate sentences true in every context, they are also a-priori sentences.

To see that "Obama is here now" is not PV, let's imagine a context c_2 in which Trump ($= b$) is the speaker, the White House ($= h$) is the place, and 2017 ($= t_2$) is the time. (We also assume that the White House houses at most one person at a time):

6. $M^{c_2} \not\models in(a, h, t_2)$ — Fact about c_2
7. $M^{c_2} \models |in(a, here, now)| = 1 \leftrightarrow in(a, h, t_2)$ — (a),(b)
8. $M^{c_2} \not\models |in(a, here, now)| = 1$ — 6,7
9. $M^P \not\Vdash |in(a, here, now)|^@ = 1$ — 8, (b)
10. $\neg \boldsymbol{PV}(\text{"}\boldsymbol{in(a, here, now)}\text{"})$ — 1, 9

⊠

Clause (b) of the test requires to show that a pragmatic validity operator for L^P is impossible. A pragmatic validity operator Π is such that ⌜$\Pi(\phi)$⌝, when used, is true if and only if ⌜ϕ⌝ $\in PV$. We show that the assumption that such an operator is used in c_1 contradicts our previous result.

11. $M^{c_1} \models |\ulcorner \Pi\phi \urcorner| = 1 \; iff \; PV(\ulcorner \phi \urcorner)$ — Assumption
12. $M^{c_1} \models |\Pi in(I, here, now)| = 1 \; iff \; PV(\text{"}in(I, here, now)\text{"})$ — 11
13. $M^{c_1} \models |\Pi in(I, here, now)| = 1$ — 5, 12
14. $M^{c_1} \models |I| = |a|$ — (64),(65)
15. $M^{c_1} \models |\Pi in(I, here, now)| = |\Pi in(a, here, now)|$ — 14, (c)
16. $M^{c_1} \models |\Pi in(a, here, now)| = 1$ — 13, 15
17. $PV(\text{"}in(a, here, now)\text{"})$ — 11, 16
18. But 17 contradicts 10. Therefore there is no operator $\boldsymbol{AP}$ fulfilling 11.

⊠

This completes the demonstration that the pragmatic system M^P is adequate to the thesis of direct reference.

5.4. Summary and discussion

In §5.2 we set up the notion of direct reference and argued that an objectual indexical semantics (one that treats contexts as real objects) cannot do justice to it. §5.3 proposed an alternative. The thought was to let the concept of context of use inform a departure from the standard formal framework. An inspection of this concept revealed its dependence on the concept of a linguistic agent, which in turn we equated with the capacity for content generation. This latter we cashed out in terms of ontological commitment. The fact that contexts of use are centered around linguistic agents was implemented in the formal system using the notion of reference devoid of ontological commitment: abstract reference.

This gave rise to the notion of an abstract object: an object referred to from a perspective in which it has no reality. Finally, we gave the rules for the operation of this formal system and saw that it is passes the direct reference test, hence adequate to the philosophical doctrine of direct reference.

Before we move on to adapt this system to the problem of the concept of truth in regimented languages, let me make some remarks and suggestions for further work. First, it is important to understand exactly how the test was won. The crucial difference between Kaplan's objectual pragmatics from §5.2.5 and our abstract system is that in Kaplan's system the law of substitutivity of identicals (c) fails for contents. Its failure on Kaplan's system is analogous to its failure for extensions on intensional semantics (see (6) and (7) in §5.2.1). In the latter case the source of the failure is the implicit relativization of the extension to a non-extensional parameter, the index. In Kaplan's case the source is the relativization to a non-intensional parameter, the context. This parameter (in both cases) can be quantified over, and this is what lets monsters in in Kaplan's case. Kaplan was alive to the fact that direct reference hates monsters, but since he was unable or unwilling to depart from objectual semantics, there was nothing he could do to stop them from coming in. By contrast, on our system we have autonomy of content: in a context, language is oblivious to character and we have full substitutivity of identicals. A monstrous operator is one the content of which depends on facts about character. Where content is autonomous such operators cannot be defined.

There are questions concerning indexicality that are central in the literature and which I have not addressed. One of the central philosophical questions in this area concerns the essentiality of indexicals, the question whether what they express is expressible by non-indexical means. Some writers argue that indexicals express perspectival contents that are inexpressible otherwise, and, of course, others object. I believe the present understanding of indexicality can contribute to this debate, but I will leave this to another occasion.[175] A different issue is raised by empirical linguistics. Some linguists claim to find empirical evidence for the existence of monsters in natural language. Whatever impact this evidence has on the foregoing is indirect, and certainly deserves exploration (on another occasion).[176]

---

[175] The loci classici for the essential indexicality thesis are Perry (1979) and Lewis (1979). See Cappelen and Dever (2013) for a survey and a sustained critique, also Magidor (2015). Cappelen and Dever expressly dissociate the idea of essential indexicality from the question of monsters in Kaplan-style semantics. I think my system can be used to mount an essential indexicality thesis that is impervious to their attacks. Such a thesis might not, however, do all that Perry and others might want from it. (Incidentally, Perry does not accept the prohibition on monsters. See Israel and Perry (1997).)

[176] See Schlenker (2011) for presentation and references. Notice that the existence of monsters in natural language is not incompatible with the thesis of direct reference as understood above, but only with natural language being directly referential. In any case, the reported monsters usually (if not only) come up in the analysis of propositional attitudes. Propositional attitudes are usually analyzed in intensional terms, but there are reasons to think that they too require a different treatment, one that digs even deeper into the concept of a linguistic agent. In view of Schlenker's evidence, we could make it a condition on such a treatment that it explain the presence (or the appearance) of monsters in propositional attitudes.

It might be illuminating to compare the present system with other systems that do similar things. I am especially thinking of Kit Fine's logic of arbitrary objects, and Bas van Fraassen's supervaluationist semantics.[177] In both cases a semantics is put forth to support a logic for objects which fall short of full-fledged existence. But in both cases the semantics is defined in terms of really existing objects, and the language is a merely notational variant on standard languages.[178] Abstract objects, on my account, do not in themselves fall short of full-fledged existence. Their abstractness inheres in the mode of referring to them.

My use of "abstract object" is different from the dominant one. In the literature, abstract objects are most often defined as being non-spatiotemporal and causally inefficacious.[179] The central examples are mathematical objects such as numbers and sets, and content objects such as concepts and propositions. Theories of these objects have the same logical status as theories of any other object. The fact that an object is abstract, on the usual view, is unrelated to its being an object in the logical sense. On my reading, the abstractness of an object is a disruption in its very objecthood.

Let me end this digression with two difficult issues about M^P which I will leave open. We saw that sentences of M^P are a-priori, but that the a-priori was a gerrymandered bunch. Phonological theory was a-priori because of how we defined language (in §1.3.1); logical theory is a-priori since it abstracts from content; the heuristics were a-priori by stipulation. But what grounds the pragmatic clauses, the assignments of characters to expressions? In what sense is it a-priori that the character of "I" is "$speaker(@)$"? Kaplan (and others) often speak in terms of conventions of language. I don't know what the relevant notion of convention is here, but the idea of convention fits in well with the practical-normative nature of characters.

A related question is: what is the metalanguage of M^P? Which language can we formulate the application semantics (I) in, seeing that we can't express it, as we are used to, in terms of a mathematical function on contexts? There is something inherently practical, as opposed to theoretical, about M^P. This is why it constitutes an essential departure from standard semantics (though this will only be proven once we formulate Convention T in it). By the weak effability thesis of §4, this makes it the language of the ineffable.[180]

---

[177] See Fine (1985), van Fraassen (1966).
[178] Fine says as much himself, e.g. (1983, p.57).
[179] See Rosen (2017) (*SEP* entry).
[180] Maybe this explains Strawson's extravagant remark, at the end of his influential paper on meanings as rules of use, to the effect that "natural language has no exact logic" (1950, p.344). Strawson's attitude is probably what leads Jason Stanley to say that "[w]hereas the notion of rule of use is vague and mystical, Kaplan's notion of the character of an expression is not only clear but set theoretically explicable in terms of fundamental semantic notions" (2012, p.893). It is undoubtedly the proliferation of set-theoretical explications in the philosophy of language that tempt Stanley to make his own extravagant remark, that "advances in [the philosophy of language] make even the most unaccomplished of its practitioners vastly more sophisticated than Kant".

6 Abstract Generality and Convention T

In this essay we are only peripherally concerned with indexical terms such as "I" and "now". The discussion of indexicality was for us a means, a ladder leading to the notion of abstract reference. It is this notion that we are interested in, since it will enable us to formulate Convention T coherently. But first we need to see how to extend its reach beyond the indexical words. Some philosophers and linguists acknowledge a form of *covert indexicality* exhibited by various kinds of expressions that are not classically considered indexical. In §6.1, I will briefly present covert indexicality. In §6.2 the idea of covert indexicality will be used in order construe standard regimented languages as thoroughly indexical, and the language of Convention T, Z, as an abstract metalanguage in the sense of §5.3. In §6.3 Convention T will be formulated.

6.1. Covert indexicals

Proper names are not indexicals in the usual sense, but a popular view in philosophy is that their mode of reference is similar to that of indexicals in being non-descriptional. This view is associated with Kripke, who argues that proper names are *rigid designators (RD)*.[181] In terms of intensional semantics, a rigid designator is a singular term that has the same reference in all indices:

(1) $RD(s)\ iff\ \exists x \forall i(|s|^{c,i} = x)$.[182] [183]

But this definition doesn't really capture non-descriptionality, since it might happen that the definiens holds of some string in virtue of its descriptive characteristics.[184] Kripke distinguishes between *de jure* rigid designators, for which (1) holds "by stipulation", and *de facto* rigid designators, for which (1) might hold in virtue of their descriptive content.[185] Since he is probably not concerned with regimented languages, it is not entirely clear what he means by "stipulation". One way to read him is to say, with Stanley, that a singular term is a rigid designator *de jure* if it fulfills (1) in virtue of "the semantical rules of the language unmediately [linking] it to [its] object".[186] Semantic rules, on the usual picture,

---

If my foregoing procedure was correct, then Kaplan's set-theoretic explication of character is in fact inadequate to the philosophy that motivates it. There is, after all, a certain amount of (philosophical) sophistication left in Kant which is not made obsolete by the (technical) sophistication of contemporary philosophy.

[181] Kripke (1980, p. 48).

[182] This is a definition of "obstinately rigid designators", since on it s denotes x even if x doesn't exist in i. See Salmon (1982, p.4).

[183] In order to keep in harmony with the literature, I revert provisionally to using the usual indices and contexts.

[184] For example, the phrase "the president in 2016" turns out rigid, though it is descriptional. See Kaplan p.494f for an example with possible worlds.

[185] Kripke (1980, p.21fn).

[186] Stanley (2017, p.922).

state characters and not contents; an unmediated linking would then be one in which the context and index parameters are vacuous, for example:

(2) $|a|^{c,i} = Obama.$

From (2) and (1) we get that "a" is a rigid designator. Let's call this the *stipulatory approach to nondescriptionality*. In terms of the abstract metalanguage M^P, a stipulatory meaning clause for the proper name "Obama" would read:

(3) $|a|^{@} = Obama.$

The stipulatory approach doesn't capture another important component of Kripke's philosophy of proper names, which he calls the *causal* or *historical chain* picture of reference.[187] According to this picture, an utterance u of a proper name refers to an object o if and only if one of two conditions hold: either u stands in a certain relation to a preceding utterance which refers to o; or u refers to o directly, through some kind of contact. We can abbreviate this and say that there is a naming tradition going from o to u. Where $u = \langle \phi(\alpha), c \rangle$, such that $\phi(\alpha)$ is a sentence containing a proper name α, we say that there is a naming tradition for α going from o to the context c. We call o the *origin* of α in c, in symbols: $lto(\alpha, c) = o$. The causal-historical chain picture of reference can be seen as a theory of meaning for proper names:

(4) $|a|^{c,i} = lto("a", c).$

In words: the reference of "a" in context c (relative to index i) is the origin of the naming tradition that reaches up to the use of the string "a" in the context c.

Both on the stipulatory approach to proper name reference and on the linguistic tradition approach, proper names are rigid designators. On the linguistic tradition approach they are also indexical (i.e. context-dependent). Since they are not among the expressions classically counted as indexicals, they are sometimes referred to as *covert* or *hidden indexicals*.[188] The function $lto(s, c)$ is then a context feature like $speaker(c)$. This is the *covert indexicality approach* to non-descriptionality.

There are two important differences between classical ("overt") and covert indexicality. First, the classical indexicals such as "now" usually vary in content across contexts, even when the speaker stays the same. The content of covert indexicals, on the other hand, though context-dependent, exhibits much less variability. A covert indexical will usually have the same content, not only for the same speaker,

[187] Kripke (1980, p. 139).
[188] See Haas-Spohn (1995) for an extended treatment of hidden indexicality in predicates and proper names. Haas-Spohn's book is a meticulous and far-reaching analysis of the philosophical significance of hidden indexicality.

but for all speakers of a speech community for many generations. If proper names are indeed indexicals, then their indexicality is much more coarse-grained than that of the classic indexicals. The second difference concerns the context feature that they designate. The overt indexicals correspond to fixed features that are part of the structure of the context: speaker, time, etc. They are given by unary functions on the context: $speaker(@)$, $time(@)$, etc. By contrast, the denotations of the proper names are all given by the same binary function $lto(\alpha, @)$, which takes a string, the proper name itself, as argument. Perhaps it is less natural to call this function a *feature* of the context, but in a technical sense that's what it is.[189]

Beside proper names, natural kind predicates are also sometimes taken to be covertly indexical. The usual examples are substance terms such as "gold" and "water" and biological species terms such as "tiger". Putnam (1975) put forth the thesis that the extensions of natural kind predicates depend on the history of their use rather than on any descriptive content. On this thesis, the meaning of a predicate such as "gold" ("G") depends on a linguistic tradition formed in the presence of an object or of stuff belonging to a certain natural kind of substance, gold. This object or stuff is the origin of the tradition. The predicate "G" applies to some new object or stuff whenever the latter belongs to the same kind as the origin. If ϕ is a natural-kind predicate, we write $nklto(\ulcorner\phi\urcorner, c)$ for the origin of the linguistic tradition for ϕ that leads up to the context c. Then on the covert indexicality picture of natural kinds, the abstract semantic clause for "G" is:

(5) $|G(x)|^c = 1 \; iff \; x \; is \; of \; the \; same \; kind \; as \; nklto("G", c)$.[190]

This is the covert indexicality approach to natural kind terms.

The question now arises whether we cannot extend the idea of covert indexicality to predicates that are not natural kind terms, for example "president", and eventually to the whole (non-logical part) of language, including the quantifier domains.[191] The thesis that says that this is possible is the *ubiquitous indexicality thesis*. Putnam seems to suggest as much at some points,[192] but his arguments are not sufficiently explicit or general, and he never gives a criterion for deciding when a word is amenable to such a treatment (much less an analysis of why). There are good reasons to think that natural kind terms

[189] The idea of hidden indexicality is traced back to Putnam (1975), who focuses on natural kind terms (see below). Putnam speaks of an "unnoticed indexical component" (p.234), and apparently means only that certain words are indexical that weren't traditionally thought to be such. Haas-Spohn has a more substantive distinction: covert indexicals are those that depend for their content on the world of the context, whereas overt indexicals depend on the other features (see §3.2 in her book). (Contexts are for her, like for Kaplan, tuples containing possible world.)
[190] See Haas-Spohn (1995), §3.3 for a similar definition.
[191] See Stanley and Szabo (2000) for a discussion of context-based quantifier domain determination.
[192] E.g. in (1975, p.244): "[W]e might doubt that there *are* any true one-criterion [descriptional] words in natural language…". And a little later: "Not only does the [covert indexicality account] apply to most nouns, but it also applies to other parts of speech".

are especially fit for a covert indexicality treatment (to the exclusion of other terms), but since natural language is not my object of investigation, I will leave the question open.[193]

6.2. Bringing it all back home

Which is the right approach to non-descriptional content – the stipulatory or the covert indexicality one? Let's look at an argument by Kaplan against the covert indexicality approach for proper names in natural language. Recall the two differences between classical and covert indexicality. First, the context-dependency of a covert indexical is coarse-grained: among the utterances of a single speaker, and even of a whole linguistic community, a difference in context will usually not make for a difference in denotation. Second, classical context features are various unary functions on the context, while the context feature for covert indexicals is a single binary function taking a context and a string. This suggests that the context dependence in the case of covert indexicals is of a different kind than in the case of the classical indexicals. It resembles the kind of context dependence that language exhibits regardless of indexicality, just from the fact that I rely on context in order to determine which language a certain sound or visual mark is to be interpreted in (and that the sound or mark are linguistic in the first place). The context dependence of indexicals Kaplan calls *semantic*, and the kind pertaining to language in general *presemantic*.[194]

When I hear the string "I" uttered, I depend on the context in order to determine that what I've heard is the English first-person singular pronoun. This is a presemantic context dependence. I then look a second time into the context to see who the utterer is, in order to assign him or her to the uttered string as its reference. This second context dependence is brought about by the semantics of the first-person singular pronoun. By contrast, upon hearing the string "Obama", once I've determined which lexicon it belongs to, no further inspection of the context is needed in order to determine the reference. Context is only appealed to presemantically. The term "Obama", according to Kaplan's claim, is no more indexical than, say, "president". If Kaplan is right, then Kripke's causal-historical chain picture is not there to explicate the lexical meaning of a proper name, but only the way in which it received its meaning. It is a thesis of diachronic semantics, or presemantics. It is therefore by stipulation that proper names are non-descriptional (and a similar argument would go through for natural kind terms).

[193] See, e.g., Abbott (1989). Haas-Spohn (1995) seems to endorse a ubiquitous indexicality thesis (see esp. §3.9), but her concept of indexicality might not be fully directly referential, since it depends on a fallible epistemological relation to the referents (§4.1).
[194] See Kaplan, p. 562. Later, Kaplan uses the term *metasemantic* (1989b, p. 574).

This argument, if sound, refutes the covert indexicality approach and with it the ubiquitous indexicality thesis for natural language. But it opens the way to an analogous thesis for regimented language. The success of the argument depends on how we draw the line between the semantic and the presemantic. Thinkers such as Kaplan, Putnam and Kripke take themselves to be modeling natural language competence. The assumption here is that a language user has an underlying mental representation of the meaning of the expression used, and that it is empirically discoverable what this representation is, at least to a certain approximation. The ways to discover what the underlying representation is are the tests and diagnostics of the trade, the trade here being, I guess, empirical semantics. Kaplan, Putnam and Kripke write at a time in which these methods are not yet very developed, so their procedure is lax by modern standards. Still, if they do take themselves to be engaged in empirical semantics, then Kaplan's claim that proper names are context-dependent only in a presemantic way can be interpreted as a prediction that the relevant tests and diagnostics would make it implausible that the underlying representation of a proper name include a representation of its linguistic tradition (and *a fortiori* for natural kind predicates, and *a multo fortiori* for other predicates).[195]

Since our own project is not the modeling of natural language competence, but the investigation of regimented languages, we may draw the boundary between the semantic and the presemantic along another line. Regimented languages are artificial devices for the precise and accountable expression of conceptual content. A regimented language is not used in order to communicate content between two speakers, but in order to express content *tout court*, regardless of who the user is, regardless even of whether there is any user. One of the central aims of language regimentation is to disconnect discourse as much as possible from subjective circumstances such as user and time of use. This is why regimented languages usually contain no indexicals. The only context-dependence of regimented languages is of the presemantic kind. This is a context-dependence that we can't shake off. We can, however, exploit it.

We want to adapt the abstract semantic framework of §5.3 to our standard regimentation scheme. This framework was developed with indexicals and semantic context dependence in mind, but since there are no indexicals in regimented languages, we can decide to treat their presemantic context dependence as if it were a semantic context dependence. Presemantic context dependence, in the case of regimented languages, is not a dependence on a historical linguistic tradition, but on the "act" or "event" of regimentation. By "act" of regimentation I don't mean any real act taking place at a particular time by

[195] How plausible is it to say that the New Reference Theorists take themselves to be practicing empirical semantics? For them it is still less "natural language" and more "ordinary language", less a cognitive faculty and more a guide to our ordinary intuitions, which are the best clue we have to metaphysics. Forty years later, Stanley (2017, p.920) writes that "[t]he fact that natural-language proper names are rigid designators is an empirical discovery about natural language". This might be an overstatement (see, e.g., Geurts (1997), Elbourne (2005) for dissenting views), but it is a telling one. It tells the story of the empiricization that the philosophy of language has undergone in the past 30 years.

particular people. These details are irrelevant for regimented languages. The presemantic context dependence of regimented languages is a dependence on the definition of the language. And it is this definition that is now proposed to be construed as semantically, and not presemantically, context dependent. On this proposal the entire non-logical content of a regimented language, including its domain, is (covertly) indexical. Only the logical features of the language will be considered presemantic. The line between the presemantic and the semantic is drawn along the line between regimentation scheme and regimented language.[196]

The notion of context relevant for regimented languages is not one that contains speakers, times or places as features. These belong to the conversation heuristic, and regimented languages are not designed for conversation. What remains of the concept of context once we have purified it of everything inessential is the bare notion of linguistic agency, and this notion, once we've distilled it into its conceptual core, is just the idea of language use. Once we've made these purifications and distillations, there remains no difference between context, agent, and language. The abstract variable "@" will here refer, abstractly, to the regimented languages themselves.

There are no longer any non-indexical terms in our languages. In L_{pr}, for example, the terms "a","b" and "P" become indexical. If they are indexical then the abstract metalanguage does not have to express interpretations (contents) for them, and can accordingly be very thin. The only vocabulary it will have will include the logical and phonological (structural-descriptive) lexicon, and the abstract variable @ with its context feature expressions. This is precisely the language Z that we have been after since Chapter 4. The variable @ in Z ranges abstractly over all standard languages, i.e. over the entire language hierarchy; since it is abstract, it does so without in principle being able to distinguish between languages, and therefore without having them as individuals in its domain. In this way Z applies to all languages without being able to define truth for them.

Since it can't distinguish between its object-languages, it has a uniform pragmatic-semantic theory for all of them, which takes the form of a generic character statement for each predicate. The interpretations of terms in a regimented language are not fixed by a historical linguistic tradition, but by the regimentation conventions in force in that language. The character of a string s is given by the expression $\ulcorner rgm(s, @)\urcorner$. A predicate "$P$", for example, will now have a single uniform character, and whether it expresses the concept of being a president, or of being a philosopher, or anything else, is determined in context by the regimentation decisions that hold for that context. The general form of a character statement for a predicate is:

[196] Notice how, since predicates and quantifiers are now indexical, the difference between direct and conceptual reference (recall §5.2) is obliterated.

(6) $Z \Vdash |Px|^{@} = 1 \leftrightarrow x \in rgm("P", @)$.[197]

Since rgm is the only context feature that we will use, we can abbreviate it and write $\ulcorner @s \urcorner$ for $\ulcorner rgm(s, @) \urcorner$. This notation suggests that we can conceive of the abstract variable not as a singular term, but as an operator on strings. Its philosophical interpretation is as a *use* operator: it takes a string and "returns" a use of it. I scare-quote "returns" because in some cases, most notably when s is a sentence string, using it doesn't result in any object. This leads to the following anomalous feature of Z: the syntactic category of an expression of the form $\ulcorner @s \urcorner$, for s a string, depends on the syntactic category of the string s. For example, the following are well-formed (and can serve as generic character statements for their syntactic categories):

(A) **Generic character statement:**

(a) $Z \Vdash |s|^{@} = @s$ (if s is a singular term),

(b) $Z \Vdash |s(x)|^{@} = 1 \leftrightarrow x \in @s$ (if s is a predicate),

(c) $Z \Vdash |s|^{@} = 1 \leftrightarrow @s$ (if s is a sentence).

In words, (A)(a) says that the singular term "s", in the language @, refers to the object that "s" refers to according to how the language @ is defined. This might look trivial or circular.[198] But this is only an illusion of circularity, due to a "flattening" of levels in abstract discourse. (A), after all, is not a declarative sentence but a rule to be followed in a particular context. To see it in action, let's specify a context. Specifying a context does not anymore consist for us in naming speakers and times, but in defining a regimented language. We can take one of the languages we've already used earlier in the essay (§5.2.1):

Regimentation statement for L_{pr}:

Let L_{pr} be a standard language, containing "a" and "b" as sole individual constants and "P" a sole unary predicate. The universe contains Obama (referred to by "a") and Trump ("b"); "P" expresses presidenthood. More precisely, we can say that "$P(x)$" is true of all and only presidents.

We can now apply the generic character (A) to the context L_{pr}. See §5.3.6(I) for the rules of application. Recall that the results of an application of a character to a context, the content generated, is stated in a surveying metalanguage. Here this is the language in which the regimentation statement for L_{pr} is given. The application clause of the generic character for the string "a", for the context L_{pr}, is:

---

[197] The main result of Haas-Spohn (1995) is a similar definition of *formal character* (§3.9). She is concerned with natural language and has a non-abstract pragmatics.

[198] Recall Kripke's (1980, p. 68f) critique of Kneale's (1962, p.629f) suggestion that the meaning of a proper name, say "Socrates", is "the individual called 'Socrates'".

(7) $\left[|a|^{@} = @a\right] \rightrightarrows_{L_{pr}} |a| = Obama$.

Here we see that the circularity that we worried about is only apparent. The character statement seems circular in its abstract statement since on both sides of the equality sign we find the string a. Abstractly, a character statement says that a string means in context what it means in context. Once applied to a concrete context, however, the appearance of circularity disappears when the second occurrence of "a" is replaced by a name in the metalanguage for what "a" means in the object language.

We can start to connect the pieces. The task we set ourselves in chapter 5 was to find or develop a device of generality that would not carry ontological commitment. This device is the abstract variable @. With it, the language Z can refer to all languages of the standard language hierarchy in abstraction from their contents. We don't need to assume a universal domain of languages. On this picture languages, as contexts were in §5.3.5, are abstract objects. The unity of a language *L* is given abstractly in Z, and its constitutional distinctness (ontology and ideology) is given in *L* itself, though without unity. In the case of contexts, we also made use of the notion of a surveying metalanguage, one from which a real (non-abstract) semantics can be expressed for a context. In the case of languages, this coincides with the usual notion of metalanguage. This is, then, our answer to the question of the language Z.

6.3. Convention T regimented

The abstract variable @ doesn't need a quantifier to bind it. Its semantics of application provide universal applicability without assuming a standing domain of objects. Whenever situated in a concrete linguistic situation, whether a context of utterance or a regimented language, and only when so situated, the application of sentences of Z can be carried out. This is the sense in which an abstractly general statement applies to an object only once that object is in fact *given*, and not, as in real universal quantification, regardless of whether it is given or not (see §5.1).

A language is defined by stating its domain and the meanings of its terms. This is called a regimentation statement (see §6.2(a) for an example). If we were to formalize a regimentation statement, it would look exactly like a definition of satisfaction (as in §3.2.4G), except that the purpose of a regimentation statement is not to analyze some language given in advance, but to set up a new one by stipulation. Let d be a regimentation statement (a definition of satisfaction). By §3.2.4G, d has the following form:

(8) $sat(x, \sigma) \leftrightarrow \forall r(\phi_d r \rightarrow r(x, \sigma))$,

where ϕ_d is the complex condition on the relation r modeled after the inductive definition of satisfaction (see §3.2.4). The language that is thereby defined, when considered as an object, is the smallest relation r that satisfies ϕ_d. The string that picks this object out we abbreviate as $\mathcal{L}_d$:

(9) $\mathcal{L}_d = \ulcorner \imath r. \forall x \forall \sigma (r(x, \sigma) \leftrightarrow \forall r' (\phi_d r' \rightarrow r'(x, \sigma))) \urcorner$.

A string can only pick out an object when interpreted. We let the symbol "$\mathcal{L}_d^M$" designate the object that the string $\mathcal{L}_d$ picks out when interpreted in a language M. $\mathcal{L}_d^M$ is simply the object-language defined in M by the string d. This is the mediated notation for languages that we've used in §4.2.3.

The expression $\mathcal{L}_d^M$ picks out an object depending on M. Therefore M has to be given in order for the expression to be used. But if given, then it is given by a further regimentation statement, formulated in a metametalanguage $\overline{M}$. The same reasoning can be carried out for $\overline{M}$, and we find ourselves in an infinite regress, so that no language can ever be given. We can put an end to the regress by referring to the metalanguage using the abstract variable @.[199] The symbol $\mathcal{L}_d^@$ refers to the object referred to by the string in (9), except that the language in which the string is to be interpreted is @, i.e. it is not yet given. Only when $\mathcal{L}_d^@$ is applied to some particular language will it yield an object. This object will be an object-language for the language to which $\mathcal{L}_d^@$ is applied. And the regimentation statement d will of itself be an adequate truth definition for it. Thus, for a language that we have regimented, the regimentation statement yields of itself an adequate truth formulation. This is the first formulation of Convention T:

(B) Convention T (abstract version)

(a) Language-synthetic formulation:

$\forall d (ATD(d, \mathcal{L}_d^@, @) \leftrightarrow FCDef(d))$.[200]

For an example, take the regimentation statement for L_{pr} in §6.2(a). It was given informally, but we could formalize it into a formally correct sentence d_{pr} in some suitable metalanguage M_{pr}. Then by (B)(a) we would have:

(10) $ATD(d_{pr}, \mathcal{L}_{d_{pr}}^{M_{pr}}, M_{pr})$.

In words: the regimentation statement for the object-language is of itself an adequate truth definition for that language. The synthetic formulation is thus a statement of the philosophical point that the concept of truth is constitutive of language.[201]

A point to notice here is that an application of the abstract Convention T to the case of some particular object-language results, not in a statement in the metalanguage, but in a statement *about* the metalanguage. This is because ATD is a predicate that applies to a truth definition, and the truth

[199] See below, §7.3.
[200] $FCDef(d)$, I remind you, means that d is formally correct (see §4.1.2).
[201] See §2.3.2, §5.3.6.

definition is what applies to sentences in the object-language. Convention T, applied, results in a statement in the metametalanguage.[202]

Often we are interested, not in the definition of an object-language in a metalanguage, but in whether one language M is in a position to define truth for another language L, such that both languages are given independently in advance. The language-synthetic formulation doesn't apply to this case since it applies only when the object-language is given in terms of the metalanguage. Now if L and M are given, then they are given by regimentation statements d_M and d_L in a common metalanguage. The second formulation of Convention T gives the conditions under which a definition d is an adequate truth definition in M for L:

(b) Language-analytic formulation:

$$\forall d_L \forall d_M \forall d(ATD\left(d, \mathcal{L}^{@}_{d_L}, \mathcal{L}^{@}_{d_M}\right) \leftrightarrow \exists f \left(ATF\left(f, \mathcal{L}^{@}_{d_L}, \mathcal{L}^{@}_{d_M}\right) \wedge PhCond(d, f)\right)).$$

Let's see this formulation in action by considering the (special) case in which M_{pr} is both the object-language and the candidate metalanguage.[203] In this case we put M_{pr} in ATD and in ATF in both argument places. Applying the abstract character (B)(b) to the context M_{pr}, we get:

(11) $$\forall d(ATD\left(d, M_{pr}, M_{pr}\right) \leftrightarrow \exists f \left(ATF\left(f, M_{pr}, M_{pr}\right) \wedge PhCond(d, f)\right)).$$

The identity mapping id is trivially an adequate translation in this case, in symbols: $ATF(id, M_{pr}, M_{pr})$. Since all adequate translations are equivalent (see §4.2.1), we can put id for f in the definiens:

(12) $$\forall d(ATD\left(d, M_{pr}, M_{pr}\right) \leftrightarrow PhCond(d, id)).$$

In words: a truth definition in M_{pr} for M_{pr} is adequate if it stands in the required phonological condition to the identity translation. Unpacking $PhCond$ (see §4.1.3) and assuming d is formally correct, we have:

(13) $$\forall d(ATD\left(d, M_{pr}, M_{pr}\right) \leftrightarrow Cons(d) \wedge \forall x \left(E\left(d, Tsent(x, id)\right)\right))$$

[202] Tarski says precisely this in *CTFL*, p.188f, in the footnote appended to Convention T. Unfortunately, his English translator Woodger missed this point, and puts "metatheory" instead of "metametatheory". The Polish and the German versions have two *meta*s (each). See also *CTFL* p.175 (which Woodger translates correctly).
[203] For readability I stop using mediated notation in favor of the language names directly.

In words: a truth definition in M_{pr} for M_{pr} is adequate just in case it is consistent and it entails all disquotational T-sentences (disquotational T-sentences is what you get when you take the identity mapping as your translation function).

From the reasoning of the liar paradox (§2.3.1) we get:

(14) $\forall d(\forall x(E(d, Tsent(x, id))) \rightarrow \neg Cons(d))$.

In words: any definition that entails all disquotational T-sentences is inconsistent. Since all adequate translation functions are equivalent, this implies:

(15) $\neg\exists d(ATD(d, M_{pr}, M_{pr}))$.

This is a statement to the effect that M_{pr} cannot define truth for itself. Abstracting, we get the general result known as Tarski's indefinability theorem:

(16) $\neg\exists d(ATD(d, @, @))$.

In this way the indefinability theorem follows from Convention T (see §3.1).

One last comment before we close off this chapter. Chapter 5 was presented as the search for an essential departure from standard regimentation. We had good reason to believe that Convention T required such a departure (because of the effability problem), but saw that it was hard to come by (because the sticky hierarchy principle). Indeed, intensionality was argued to be a merely notational departure (§5.2.2), and on the usual treatment, so was indexicality (§5.2.5). However, indexicality on the usual treatment was not even adequate to its own purpose, that of capturing direct reference. It is the thesis of direct reference that led us to take apart some of the fundamental notions of standard regimentation and put them back together in a different way, which we called abstract semantics. But is abstract semantics an essential departure from standard regimentation, or will we be able to find a standard language that expresses the same content? The direct reference thesis does not answer this question.

In this section we have found the answer. Since Convention T can be formulated in abstract semantics in full generality, and since we know that Convention T cannot be formulated in standard semantics in full generality, abstract semantics is an essential departure from standard regimentation. The difference between Z and the standard languages is not a difference in mere ontology or ideology, but a difference in mode of discourse. I call the standard languages *theoretical languages*, since their primary function is to state theories, linguistic representations of the non-linguistic world. The abstract language Z I call the *methodological language* (in homage to Tarski's "methodology of the deductive sciences"). It is used to represent theoretical languages, but with prescriptive rather than descriptive force.

7 Tarski's Revenge

The device of abstract reference and the distinction between theoretical and methodological discourse lead to the successful resolution of the unity problem from §3.3. I will spell this out in §7.1. The rest of this chapter will be devoted to sketching further developments and answers to other objections. The way in which the unity problem is resolved suggests a revision in how we conceive of the task of concept analysis. §7.2 sketches this revision, which I call the *two-pronged approach*. §7.3 answers the regress objection (from §3.3) and §7.4 resolves the issue of the diallele which haunted us since the beginning of the work (see §1.3.4, §3.2.1). §7.5 lists some further issues that will not be addressed in this work.

7.1. The indexical reply to the unity objection

Let's review once more the dialectic of Part One. Our task was to analyze the content of the concept of truth, and the problem was that our basic assumptions led to there being many incompatible analyses. The first assumption was that the content of a concept is the same as its conditions of application to an arbitrary object, and that therefore that an analysis should take the form of a definition. The second assumption was that the basic truth-bearers are sentences in regimented languages. This choice had an important consequence: since sentences are relative to a language, a truth definition would also, at least at first sight, be relative to a language. The concept of truth would thereby receive many different definitions, which by the first assumption would mean many incompatible analyses. In chapter 2 we hoped to avoid this relativity with the aid of the naïve conception of truth. The naïve conception is compatible with the existence of a universal language. Limiting our discussion to universal languages would have been a well-motivated move which makes language relativity benign. Unfortunately, the naïve conception, along with the possibility of universal languages, was found to be untenable.

This is where the stratified conception came in (chapter 3). Dropping the requirement of semantic closure from the naïve conception yielded Convention T, the material adequacy criterion for definitions of truth. Following Tarski, we then described a procedure for coming up with an adequate definition of truth for a given language *L*. The task of Part One was thereby achieved: a definition of truth was found. But since a universal metalanguage was not possible, there was no way to get away from the relativity of truth to language. There was no one definition that we could say gives the content of the concept of truth. This was the unity objection to the stratified conception.

To this we replied that it is Convention T, the criterion for material adequacy of truth definitions, that provides the unity of the concept: a sentence is a definition of truth insofar as it conforms to Convention

T. One way to understand the relation between Convention T and the definitions is on the model of the relation between intension and extension in semantics. Intensions are functions from circumstances of evaluation (indices) to extensions, and they explain how the same expression can have different extensions. They provide the unity of the meaning of the expression, as it were, across its different extensions. Applying this model to the case of truth, we can think of the different languages as so many circumstances of evaluation, and of Convention T as a kind of informal function from a language to a truth definition.[204] This was the intensional reply to the unity objection. However, the intensional reply ran up against the effability objection: since there is no universal metalanguage, there is no language in which to formulate Convention T so that it will apply across the board to all languages. This was the effability objection to the intensional reply, or *Tarski's revenge paradox*, and with it we closed Part One.

Part Two was a search for a reply to the unity objection that would sidestep the effability objection. The strategy was to come up with a new expressive device that will let Convention T range over all languages without the ontological commitment that attends standard universal quantification. The first step (chapter 4) was to formalize Convention T in order to isolate the place in which the new device is needed. There were two such places: a (single, initial) quantifier over interpreted languages, and a ternary predicate $ATF(f, L_1, L_2)$ that says of a string mapping f that it is an adequate translation function from a language L_1 into a language L_2. The next step (chapter 5), and the heart of Part Two, was the development of abstract indexical semantics. The guiding thread was to establish a logical difference between the concept of an index, or a circumstance of evaluation, as used in intensional semantics, and a context of use as needed for indexicality. The properties of contexts gave rise to a new abstract semantics, which expresses generality without ontological commitment. This allows us to mount an *indexical reply* to the unity objection, on which languages are considered as contexts of use, and not, as on the intensional reply, as circumstances of evaluation. Convention T can then safely be seen as that which provides the unity of the concept of truth. The principal task of Part Two, and with it the central goal of this essay, I therefore consider fulfilled. There remains the issue of the predicate ATF of adequate translations. I will discuss it below (7.2.3).[205]

---

[204] This applies more directly to the general strategy for constructing truth definitions, but Convention T is the more central item of Tarski's account.

[205] This essay, I've stressed in the introduction, should not be seen as an exegetical piece on Tarski. Nonetheless, it might shed light on Tarski exegesis. See David (2008) for an interpretation of *CTFL* based on a very similar idea to the main claim of this essay, though with less detail. Another interpretation of Tarski consonant with the present approach is Hodges (1986).

There are many solutions to the liar paradox that presume to avoid or mitigate the revenge paradox. Among these there are several that make use of the notions of indexicality and context-dependence. The most notable, perhaps, are Parsons (1974), Burge (1979), Barwise and Etchemendy (1987), Simmons (1993), Juhl (1997) and Glanzberg (2004). See Gauker (2006) for a critique of their common ground. There are some deep differences between these approaches and the present essay. First, these writers are all concerned with paradox in natural language (though without a precise statement of what that is). Second, though they all appeal to context dependence, they don't

7.2. The two-pronged approach to concept analysis

The indexical reply calls for a revision of how we understand concept analysis. Convention T, we agreed, conferred conceptual unity on the myriad definitions of truth. In order to complete the picture we have to say something more about *how* it provides this unity, or *what it is in general* to provide the unity of a concept.

The purpose of conceptual analysis is to lay bare the content of the concept analyzed. According to the view we started out with, this means providing a clear statement of the conditions under which the concept applies to an arbitrary individual (or individuals, in the case of relational concepts). In order for the statement to be clear, i.e. admit of no ambiguity or indeterminacy, it has to be given in a regimented language L. An analysis therefore takes the form of a definition, a sentence of this form:

(1) $Px \leftrightarrow \phi(x)$,

where "P" is replaced by a name of the concept (a string not in L) and ϕ by a formula of L which holds of all and only the objects falling under P. Since the string replacing "P" is not in L, neither is the definition. Let L^+ be a language like L with the addition of P to its ideology. We think of P as, to begin with, uninterpreted. We define the *import* of a definition to be the set of its logical consequences:

(2) The import of a definition $d = \{x \in L^+ : E(d, x)\}$.

The members of the import of d are the sentences *analytic* to the concept P (relative to d). We can equate the *content* of a concept with its import. For example, if we define:

(3) $Bachelor(x) \leftrightarrow Male(x) \wedge age(x) \geq 18 \wedge \neg\exists y Wife(y, x)$,

then (4) is analytic to the concept *bachelor* relative to (3), and (5) isn't (for a some individual constant):

(4) $Bachelor(a) \rightarrow Male(a)$,
(5) $Bachelor(a) \wedge Male(a)$.

Notice that if we take the import of (3) and replace each occurrence of "$Bachelor(x)$" with its definiens, what we get is the set of logical truths of L. This is reassuring in view of the deep connection between logic and analyticity that writers on the subject have always emphasized. But it is also surprising: if the content of a concept is no more than the set of logical truths, then what new information is uncovered by an analysis? The sentences analytic to a concept (relative to a definition d) are, if we

---

offer an analysis of the concept of context. Nonetheless, there are certain interesting and not always obvious affinities between the approach presented here and, especially, Parsons, Glanzberg and Gauker. I leave it to a later date to expose these affinities.

accept d, exactly the logical truths. Analyticity is therefore nothing more than logical truth, conditional on the acceptance of a definition. It is this condition, the acceptance of the definition, that makes concept analysis informative. But if this is the case, then the definition cannot be all there is to concept analysis. In addition, we have to accept it *as* an analysis of the right concept. This acceptance is itself a judgment to the effect that the proposed definition is materially adequate. And such judgment needs to be performed against a previously specified criterion. What is missing in our view of concept analysis is an account of this criterion.

Let's look at a simple case first. Consider the concept of an ordered pair, or more precisely, the relation of x being the ordered pair of y and z, in symbols: $x = \langle y, z \rangle$.[206] A successful definition will provide a necessary and sufficient condition under which an arbitrary triad stands in this relation. For number theory, the following definitions both provide an adequate application condition:

(6) $\langle x, y \rangle = (x + y)^2 + x$,

(7) $\langle x, y \rangle = 2^x \cdot 3^y$.[207]

But they will differ in import. Sentence (8), for example, is analytic of the concept of ordered pair relative to definition (6) but not to (7):

(8) $\langle 4,4 \rangle = 68$.

Can it be that two incompatible definitions are both adequate to same concept? The reason that we consider both (6) and (7) to be successful is that on both of them the *characteristic property of the ordered pair* holds, which is that the individuation of pairs depends on the order of their members. Formally, they all entail the following sentence cp:

(9) $\forall x \forall y \forall z \forall w (\langle x, y \rangle = \langle z, w \rangle \leftrightarrow x = z \wedge y = w)$.

Any definition d which entails cp is an adequate definition. If a definition entails, in addition, other sentences independent of cp, no harm is incurred. Such sentences are what Quine calls the *don't cares*.[208] However, we can't use cp as a definition, since it doesn't have the right form. It doesn't state necessary and sufficient conditions for an arbitrary object to be an ordered pair, but only for two ordered pairs to be identical. cp is not a definition, but it can be used to state an *adequacy criterion* for definitions:

(10) $APD(d) \leftrightarrow E(d, cp)$.

---

[206] The following is informed by Quine (1960, §53).
[207] More strictly: $z = \langle x, y \rangle \leftrightarrow z = (x + y)^2 + x$, etc.
[208] (1960, p.238).

Here "APD" abbreviates "adequate definition of ordered-pair". In words: a definition is adequate if (the sentence expressing) the characteristic property of ordered pair is included in its import. Since cp is formulated only in terms of logic and the definiendum, the criterion is applicable to any proposed definition of ordered pair, in any language. It does not presuppose set theory, or arithmetic, or anything else. It is in virtue of their conformance to (10) that the particular definitions (6) and (7), and any of the many others that are in use, are definitions of the concept of ordered pair. We call (10) the *character* of the concept of ordered-pair, and the various (adequate) definitions various explications of its *content*. We saw above that different definitions of ordered pair have different imports, different sets of analytic sentences. The character picks out the part of the import that will be common to all adequate definitions, the *essential import*:

(11) Essential import of the concept of ordered-pair: $\{x: E(cp, x)\}$.

The sentences that belong to the essential import of a concept are called *properly analytic* of the concept. Quine's don't-cares are those analytic sentences (relative to a definition) that are not properly analytic.

Both definitions (6) and (7) are materially adequate to the concept of ordered-pair, and there are many others, but many definitions aren't, for example:

(12) $\langle x, y\rangle = x^2 + y^2 + x.$

Conceptual analysis consists, not in the definition by itself, but also in the judgment that the definition is adequate. This step is what sanctions the identification of the analytic sentences with the logical truths. Definition (6), say, is a successful analysis of the concept of ordered pair inasmuch as it goes along with the recognition that, e.g., (12) is not a successful analysis. In general, neither a definition on its own nor a character on its own can be called an analysis of a concept. We need both. The definition gives necessary and sufficient conditions for the application of a concept. It tells us how to use the concept. The role of the character is to connect the definitions with that which is to be analyzed, to present them as analyses of *this* concept. The definitions work from the bottom up, so to speak, and the character works from the top down. This is the *two-pronged approach* to concept analysis.

The character, we say, is that which presents the definition as materially adequate, or as the definition of the right concept. But who says that the character itself captures the right concept? What promises *its* material adequacy?

When we say that we are out to define a concept, for instance the concept of truth, we are assuming that the concept is in some way *given* to us prior to the investigation, and that the definition we are looking for seeks to capture this previously given concept. But if a definition is called for, then this is because the concept is given in a way that is deficient – vague, or confused, or somehow unreliable. Let's call

this deficient presentation of the concept *the intuitive grasp*. The intuitive grasp of a concept is like unregimented discourse: it contains ambiguities and indeterminacies of both form and substance, but it enjoys a certain epistemological primacy in that it is given and not made. However, if the concept is given in this deficient way, how can it be used in the assessment of a definition's adequacy? I take it to be one of Tarski's main achievements to realize that there is an intermediate station between the intuitive grasp and the regimented definition: the character of a concept.[209]

We can think of the character of a concept as a regimentation of its intuitive grasp. This makes the question of whether a definition conforms to it or not well-defined. However, how do we determine the faithfulness of the character itself to the intuitive grasp? How should we decide between two competing characters? There is no question of looking for yet another criterion of material adequacy as an intermediate station between the intuitive grasp and the character, for we will find ourselves with the same problem for the new criterion. In the case of truth, the two candidates to decide between are the naïve conception and Convention T. Here the decision is easy. Although the naïve conception is perceived by many to be closer to the intuitive grasp than Convention T, it is simply untenable. And since Convention T is nothing but a straightforward weakening of the naïve conception, it can contain nothing counter-intuitive that the naïve conception lacked. If theorists object to Convention T, it is only because they think they can do better.

The grounds to accept Convention T is, as its name suggests, conventional. We, who are engaged together in the enterprise of truth, agree for it to serve as the character of the concept of truth. Whether we thus agree only for lack of a better alternative or for a more positive reason is irrelevant. For the validity of Convention T, an agreement between us is enough.[210]

---

[209] Tarski of course didn't use the term 'character'. Sher (1999) comments that in regimenting Convention T "the material task itself is construed as a formal task" (p.150). This is an apt description.

[210] See Patterson (2012, pp.46ff,p.110) for the history of "convention" in Convention T. Also illuminating is Tarski's description of his task in his 1966 definition of the logical notions. There he says that his approach "has a normative character: we make a suggestion that the term be used in a certain way". This suggestion is "independent of the way in which [the term] is actually used", but also "in agreement… at least with one usage which actually is encountered in practice" (in Tarski and Corcoran (1986, p. 145)).
I mentioned that the character of a concept to its intuitive grasp is like that of regimented to unregimented discourse. But this is no more than an analogy. We should not identify the intuitive grasp of a concept with the unregimented use of the concept's term. An intuitive grasp can have many sources, for example historical tradition. Reading Tarski we see quite clearly that the intuitive grasp of truth for him is firmly rooted in the tradition beginning with Brentano's engagement with Aristotle's conception of truth and reaching Tarski through Twardowski and Kotarbiński. See Tarski (1944, §3); Woleński and Simons (1989); Murawski and Woleński (2008).

7.3. Reply to the regress objection

We are in a position to solve the regress problem.[211] First, let's notice an elaboration that our framework introduces into the use-mention distinction. The defining feature of the methodological metalanguage Z is the abstract variable @. It refers abstractly to the theoretical language being used, in the way that the character $time(@)$ refers abstractly to the time at which the indexical "now" is being used. This is a special mode of reference to language. The distinction between use and mention of language is famous, indeed a mainstay of analytic philosophy of language, but it is not usually noticed that it is not a two-way, but actually a four-way distinction. First and second, there is the obvious distinction between *mentioning* an expression *as a mere string* and mentioning it *as interpreted*. To say that the word "cats" ends with an "s" is an instance of the first kind, and to say that it denotes **cats** is an instance of the second. Third, the word "cats" is something we can *use*. I did so just now, in the bold-faced occurrence. Thus the word "cats" in the bold-faced occurrence is used, not mentioned. But in the recent underlined occurrence, the word is *mentioned* again, and described *as being used*. This is the fourth item of the distinction. To sum up, the four modes are: we can either *use* a word, or *mention* it; and we can mention it either *as a mere string*, *as interpreted*, or *as being used*. Doing the latter is the function of the abstract variable @. Between these four modes, the modes of using an expression and of mentioning it as interpreted are *real*, in the sense that they carry semantic commitment to whatever is being expressed (e.g. cats); and the modes of mentioning as a mere string and mentioning as being used are *abstract*, in not carrying such commitment. Although I need to know a language in order to *use* a word; and I need to have a semantic theory in order to *mention* a word *as interpreted*; I need neither in order to *mention* a word *as being used*. I can overhear a conversation in which the sentence "koty jedzą myszy" is uttered and presume that the word "koty" is therein being used without having an inkling as to what it means. Not so if I wanted to use the word, or refer to it as interpreted (for example, by commenting "that's true"). The abstract variable allows us to mention an expression with regard to its meaning (not as a mere string) without taking on the related expressive commitments. It does this by relaying those commitments to the speaker.[212]

How does this relate to the regress objection? We can see the connection by looking at Quine's (1969b) notion of ontological relativity, which is explicitly based on Tarski's stratified conception of truth. Quine there says that the ontology of a language is not given absolutely, but only in terms of some further language.[213] Since the same will apply to this further language, we are in for a regress – precisely the regress objection raised against the stratified conception of truth. Quine isn't worried. "In practice", he says, "we end the regress… by acquiescing in our mother tongue and taking its words at face value"

---

[211] See §3.3. The point of the solution was already given in §6.3.
[212] Mentioning a word as being used depends on the possibility of referring to an uninterpreted string. This is why it was crucial to isolate a semantically inert phonological medium. See §1.3.1, §2.2.4, §4.1.2.
[213] Quine actually speaks of theories. This is important, but not for our present concerns.

(p.49). Quine doesn't elaborate on the idea of "acquiescing in our mother tongue", but we can now give a more precise account of it. For consider the phrase "our mother tongue" (equivalently: "the home language"). This is an indexical expression (because of "our" or "home"). But Quine shouldn't be read as referring specifically to his own mother tongue, English. Even French speakers can end the regress. What Quine means is that any theorist can end the regress by acquiescing in *their* home language. Quine is not *using* the indexical "the home language", he is referring to its use, or equivalently, to its character. What's more, the languages in question shouldn't be thought of as being French or English, or any of what we've called "world languages" in chapter 2. For us (and also for Quine), it only makes sense to ask after the semantic commitments of a regimented language. So the phrase "the home language" refers to a context in which we are using a regimented language, and specifically to the language being used. Since we identify a context with a language (see §6.2), the phrase "the home language" is none other than the abstract variable @.

What does it then mean to say that "in practice we end the regress by acquiescing in our mother tongue"? Recall our mediated notation for referring to languages (§4.2.3, §6.3). The mechanism there exploited the fact that a language must be given by a formula, and substituted reference to the defining formula (the regimentation statement) for reference to the language as object: "$\mathcal{L}_d$" referred to the language defined by the formula d. But since formulas are meaningless except in relation to a language, we had to specify the metalanguage in which the defining formula is given: "$\mathcal{L}_d^M$" refers to the language defined by the string d as interpreted in the language M. When we wanted to use mediated notation for M as well, we stumbled on a good visualization of the regress problem, in the form of towers of superscripts such as "$\mathcal{L}_{d_1}^{\mathcal{L}_{d_2}^{M}}$", etc. In practice, we ended the regress by plugging the abstract variable into the metalanguage position: "$\mathcal{L}_d^{@}$". What we are doing when we do this is we mention the metalanguage *as used*, and in this way avoiding ontological commitment to it and avoiding the regress. To say that "in practice we end the regress by acquiescing" is just to say that we are always using some theoretical language or other, and that that language cannot mention itself and thus can't give rise to the regress. This follows from the impossibility of a universal metalanguage. In other words, we have made the effability objection work for us: there is no language in which to formulate the regress objection.[214]

7.4. Truth and meaning

The concepts of truth and meaning were involved in a diallele: they were understood in terms of one another, in seeming violation of the requirement of non-circularity of definitions. Our way out was to

[214] This, I think, answers Field's (1974) worries about Quine's ontological relativity thesis. Field ignores Quine's remark about acquiescing, maybe because Quine doesn't elaborate it.

devise a definition independent of either notion, doing the work of both (see §3.2.1, §3.2.5). Above in §7.2 we gave a sketch of an account of how it is that definitions give the contents of concepts: they do it by conforming to material adequacy conditions, or characters. Definitions of truth are such inasmuch as they conform to Convention T, the criterion of material adequacy for the concept of truth. One term in Convention T was not yet accounted for: the predicate ATF, expressing the concept of a correct translation function. In this section we will connect this last missing piece.

Instead of going there directly, however, let us in passing take care of another objection made against Tarski's stratified conception, which is sometimes called the *modal objection*. This objection is made clearly and explicitly in Pap (1954), and then more famously in Putnam (1985), and also in other places.[215] The objection notes that the T-sentences for a language L, on the stratified conception, are theorems of the theory of truth for L, and therefore necessarily true. However, it is held, the interpretation of a language is a contingent matter. The problem can be viewed clearly in the following contrast:

(13) The sentence "snow is white or snow isn't white" is logically true.

(14) If "or" had meant what "and" means, then the sentence "snow is white or snow isn't white" would have been false.

How can it be that a logically true sentence might have been false? This is a patent absurdity, and since it seems to be a consequence of the stratified conception, serves as a refutation of the latter. But the refutation turns on an equivocation with respect to the term "sentence". For us a sentence is individuated by its interpretation, but a phonological string can be interpreted in different ways. In (13) logical truth is predicated of one sentence, and in (14) it is denied of another. It makes no sense to say of an interpreted sentence that it could have meant something else than it actually means, than it makes sense to say of the empty set that it could have had more members than it actually has. This is our reply to the modal objection.[216]

The confusion behind the modal objection is important to expose. The objection takes it for granted that the interpretation of a language is a contingent matter, and therefore that statements of truth conditions are never necessary. In some salient sense, this assumption is undeniably correct. Not only is the semantics of, say, English, not necessary, it is not even fixed in time, since words change their meanings all the time. But if this is so, then how can we accept necessary meaning statements? The answer is that there is another, deeper, equivocation here, this time with respect to the term "language". Language is said in a great many ways: as a historical object, a cognitive object, a syntactic object, etc. etc.

[215] For instance Heck (1997). See Patterson (2008c), Raatikainen (2008) for answers which resemble my own, and for more references.
[216] Compare Gupta and Belnap's (1993) approach to the same problem, p.21.

Philosophers of language, lamentably, are hardly ever explicit about which concept of language they have in mind. In the present case the confusion is between an empirical object and a mathematical object that is used to model it. The concept of language in play in Tarski-style definitions is that of a mathematical object which can be used to model empirically given languages. To complain that statements about a language in this sense follow necessarily from its definition is like complaining that a mathematical model of a physical phenomenon entails its empirical consequences necessarily.

Something like the intuition behind the modal objection does play a role, not in a critique of Tarski's theory of truth, but of its use in empirical semantic theories. The thought that meaning can be explicated in terms of truth conditions suggests that Tarski-style truth definitions can be put to use as meaning theories for natural languages. Etchemendy (1988, §1.2) argues that a definition d_1 can only be used as a meaning theory for an (empirical) language L if we assume at the outset that it is an adequate truth definition for L. But then we can't think of it as telling us anything about truth.[217]

We can frame Etchemendy's position in terms of the two-pronged approach to concept analysis. The question concerns the interaction between the predicates ATD (adequate truth definition) and ATF (adequate translation function). The definition of ATD (Convention T) was:

(15) $\quad ATD(d, L, M) \leftrightarrow \exists f(ATF(f, L, M) \wedge PhCond(d, f)).$

The relevant component of "$PhCond$" was the fact that d entails T-sentences that use the translation function f. The use of truth definitions in empirical theories of truth reverses the roles of truth and translation. For this purpose what we need is a definition of ATF, along the lines of:

(16) $\quad ATF(f, L, M) \leftrightarrow \exists d(ATD(d, L, M) \wedge PhCond(d, f)).$[218]

Etchemendy's point is that the project of discovering the semantics of an empirical language is the project of asserting ATF of some translation function, whereas the project of finding a truth definition is that of asserting ATD of some definition. We can't engage in both projects at the same time since in order to judge ATF, we already have to assume ATD, and vice versa.

The stratified conception does, after all, imply a reciprocal relation between the concepts of truth and meaning (a diallele). This relation is not, however, the relation of reciprocal *definition*. That would render both concepts unusable. On the present analysis the relation between truth and meaning is that of reciprocal *characterization*. It is a diallele in character and not in content, and therefore a pragmatic

[217] This is essentially Dummett's (1959) point.
[218] I have made some simplifying assumptions. In particular, this predicate applies only to translations from languages to their metalanguages.

and not a logical circumstance. We can still use a truth definition, either for truth or for meaning, but not for both.

7.5. In closing

There is clearly much more to say about the stratified conception, but I will leave it to another occasion. In closing I will do no more than mention some open issues and directions for further development.

(a)

The first thing to look for when presented with an account of truth that presumes to do away with the paradoxes is a revenge paradox. This is a sound imperative, but it should be nuanced. The value of the semantic paradoxes is that they are like being struck with a hammer: they cannot be ignored. But just as the hammer should not be blamed, it is necessary to look for the deep problem, of which the paradoxes are just the symptom. Accordingly, we should look for a more general way to state the imperative. A revenge liar paradox arise for solutions to the liar paradox that make use of expressive resources beyond those of the object-language. The more general imperative is therefore to state the language of the solution explicitly and see whether that language is included in the range of application of the solution. Anything less would make a theory nothing but a superficial approximation to a solution. In the terms of this essay, any solution should answer its own *effability question*.

Formulated in this generality, we see that the predicament is even deeper and more far-reaching than we suspected at first. The effability objection to the stratified conception shows that even if we completely renounce the hope for semantically closed languages and embrace a full-out hierarchy of metalanguages, even then do we face a revenge problem. The challenge is then to formulate a theory that accounts also for the language of the theory, and the question is whether Z can represent its own application semantics, or whether we need to look for some further language. I confess that I don't have a good answer at this point, but I will advance the following perhaps abstruse consideration. First, there is a sense in which Z does represent its own semantics. Let's formulate the desideratum:

> Z should be able to express the semantic concept of Z, which is the application of an abstract expression, e.g. "$speaker(@)$", to a context. A context is a real situation centered around a linguistic agent. The result of applying an abstract expression is reference to a real, non-abstract, object in the context. For example, applying "$speaker(@)$" yields reference to the speaker of the context.

This desideratum is easily achieved. It is not a problem to say in Z that the application of, say, "$speaker(@)$" to a context yields the speaker of the context. However, the phrase "speaker of the

context" is itself said abstractly, so this statement in Z is trivial. We get the feeling that the heart of the idea of application, the move from abstractness to reality, is not captured by Z's statement of its own semantics. The problem is that to the extent that it is missing from Z, it is missing also from the statement of the desideratum. That statement too, if it is to apply in full generality, must be stated abstractly. In short, there is no way to state generally what is missing from Z.

This is a slippery predicament, but let me mention a strategy for living with it. The idea is to consider particular non-abstract cases which reflect features of Z, though without its absolute generality. This strategy is not new. It is in play in, for example, set theory, where reflection principles establish that certain sets exhibit (reflect) properties of certain proper classes. We can then prove results about these sets in a non-abstract way, and draw conclusions about the abstract proper classes. In the case of Z, the reflections are be the theoretical, truth-defining, metalanguages of the hierarchy. Every such particular metalanguage lets us study truth and semantics in a concrete though restricted way. This is the *reflection strategy*, which I propose as a way to overcome the effability problem.[219]

(b)

It will be informative to explore the relations between the conceptual apparatus of the stratified conception (as developed here) and several similar or related philosophical theories. The language-level approach is not the only approach to truth based on stratification. Older and no less well-known are various type theoretic approaches. Such approaches are sometimes considered less objectionable than the language-level theory, since although they fragment the truth-predicate they keep the language in a single piece. A complete defense of the present stratified conception should show why a full-on language stratification is necessary, why we can't get the same result for a cheaper price. The answer, in an intuitive sketch, is that a type-theoretic approach doesn't have the means to avoid the regress problem. A metalanguage, on the language-level approach, is a vantage point from which truth is definable, and to which we can refer abstractly. The type-theoretic approach offers no such vantage point.[220]

---

[219] Kreisel (1967) is a classical place in which this strategy is used. A less obvious instance is, I propose, Quine's (1969c) notion of naturalized epistemology. I think a case can be made for construing Quine's objection to classical epistemology in terms of the effability problem (though this is surely not how he would describe it). His proposal to study knowing subjects as empirical objects would then be an instance of the reflection strategy. Empirical linguistics might be considered a reflection in Quine's sense.

[220] One example of such a "softer" hierarchy is Church (1976) (Church doesn't claim that the type-theoretic hierarchy is superior to the language-level one). Church shows how to define truth in a type-theoretic language, but his definitions assume that the expressions are already interpreted. See also Glanzberg (2015) for a discussion of more and less objectionable hierarchies.

(c)

Another issue is to get clearer about the notions of language and theory in play in the stratified conception and the relation between them. Logic deals with uninterpreted, *merely formal* languages. For these languages the notion of (non-logical) truth doesn't come up. Truth only comes up for regimented, or *formalized*, languages. These are languages the content of which is completely specified, up to the domain of quantification. In practice we are often not interested in one completely specified language, but in a partially specified language which defines a class of completely specified languages. A useful and common way to partially specify a language is, starting from a merely formal language, to specify a collection of sentence strings that we stipulate to be true. This constrains the range of interpretations that we can give the uninterpreted language, to various degrees of determinacy. This usage of "theory" and "language" is different from the one adopted in this essay. We can call the usage in this essay a *language-first* approach, and the other a *theory-first* approach. They are appropriate for different projects. Though I will not expand on it any further, I mention it in order to avoid confusion.[221]

(d)

Finally, there are many interpretative issues about Tarski that the stratified conception can contribute to. There are debates about the extent to which Tarskian truth can justly be considered a correspondence theory, and the extent to which it is a deflationary theory. The present development of the stratified conception gives complicated answers to both of these questions. The question of whether Tarski's theory explicates the philosophical notion of correspondence with reality receives a different answer according to whether we consider it from the standpoint of particular truth definitions or of the general character of truth. Particular truth definitions are formulated in theoretical metalanguages which make reference to non-linguistic objects. For example, the truth of a sentence such as "snow is white" can very reasonably be said to be grounded in the color of snow, which is a physical, not a linguistic circumstance. By contrast, thinking generally, we would say that the truth of "snow is white" in a language *L* is grounded by its relation to the sentence "snow is white" in *L*'s metalanguage. On this view truth is a relation between sentences (albeit of different languages) and therefore an inter-linguistic phenomenon. The metaphysical outlook of the stratified conception is therefore a *methodological idealism and theoretical realism*.[222]

---

[221] Tarski is sometimes accused of conflating the notions of language and theory. I think the point made in the text is a clue to the reason. See the exchange between DeVidi and Solomon (1999) and Ray (2005) regarding this issue. In a similar way I think we can explain the tension that we find both in Frege and Davidson, between compositionality on the one hand and the context principle (Frege) or linguistic holism (Davidson) on the other.

[222] This would explain why some writers, e.g. Popper (1979), take Tarski to be the undisputed champion of correspondence truth, while others deny it. See Patterson (2012, p. 140ff) for references. See Tarski (1944,

I hope to take up some of these issues in the near future.

§§18,19) for some discussion, which is however somewhat fleeting and superficial. This question cannot be addressed without proper consideration of Tarski's philosophical background, and the fact that attention to the correspondence theory and its proper formulation was a central feature of the philosophical tradition leading up to Tarski, from Brentano's occupation with Aristotle's definition of truth, through Twardowski and Kotarbiński. See Woleński and Simons (1989), Murawski and Woleński (2008).

The slogan I picked for the metaphysical outlook of the stratified conception is obviously meant to evoke Kant's slogan for his own metaphysical position, *transcendental idealism and empirical realism*. The similarity is not superficial, and an important direction of research is to expose the deep philosophical principles common to Kant's project and to Tarski's (at least as I develop the latter). Let me note that the heart of Kant's first Critique, the Analytic of Pure Reason, is presented by Kant himself as a (transcendental) theory of truth (see A58-9/B83). See also Posy (unpublished).

Coda

In this essay I have not dealt with what is probably the most prevalent objection to the stratified conception since Kripke – the claim that stratification is an ad-hoc device with little or no philosophical significance. Although this objection seems to underlie much of the resistance to Tarski's theory in the literature, I have never seen it supported seriously, and hardly ever even stated clearly. The main reason that Tarski's solution is considered ad-hoc is because it doesn't account for the semantic closure of ordinary language. I myself am persuaded by Tarski's own argument that ordinary language (in the sense of unregimented discourse) is simply not an object of study for precise theories of truth,[223] but the literature on truth does seem to be substantially motivated by this objection, so something must be said. My plan is to reconstruct an argument that captures the core of the objection, and to undermine it. The argument runs as follows:

The normative premise: Ordinary language is the philosophically important object for a theory of truth; formalized languages aren't.

The descriptive premise: Ordinary language is semantically closed.

Conclusion: A theory of truth that precludes semantic closure at the outset is philosophically inadequate. [224]

---

[223] Compare chapter 2 of this essay with the present section, especially sections 2.1.1 and 2.3.2.

[224] Here are some suggestive quotes:
Black (1948): "The philosophical relevance of [Tarski's] work will depend upon the extent to which something similar can be done for colloquial English". (p. 56)
Strawson (1949): "[I]n so far as [The Semantic or Meta-linguistic Theory of Truth] is simply a contribution to the construction of artificial languages, and is not intended to be regarded as relevant to the use of actual languages I am not concerned with it. But… the theory has been claimed… to throw light on the actual use of the word 'true'; or (which I take to be the same claim) on the philosophical problem of truth.". (p.83)
Martin (1970): "I see the Liar as raising questions concerning the concepts of sentence… truth, negation, reference, etc.; in short, as a problem in the philosophy of language – our language – not *primarily* as a problem having to do with formalized languages." (p. 91)
Kripke (1975): "None of [the technical] notions is to be found in natural language in its pristine purity, before philosophers reflect on its semantics (in particular, the semantic paradoxes)." (p. 714fn). Kripke is an exception in *not* thinking of ordinary language as semantically closed.
Priest (1983): "[T]he universality of language makes the metalanguage construction inherently unstable… [S]uch castles in the transfinite air [language levels] can be constructed… But they have no more significance than a mathematical game. Whatever they are, they are not English. In giving a semantical account of English the distinction between object and metalanguage is a logical apartheid which must go… In another jargon, we could say that the metalanguage is the alienated essence of (object language) truth. The alienation should of course be transcended". (p.122)
See also Barwise and Etchemendy (1987, p.5); Simmons (1993, p.62); Glanzberg (2004, p.290); and many others.

In the literature various terms in addition to "ordinary language" are used: "natural language", "our language", "English", and others. If the difference in terminology reflects a difference in concepts, this is usually not remarked on. Nor is it easy to find a detailed characterization of what is meant by these terms. This is already enough to cast doubt on the objection from ordinary language, but it is a point I will largely ignore.[225] My procedure will be to scrutinize the two premises and argue that they are, at the very least, ill-founded.

As I said, the premises are hardly ever given explicit arguments to support them. One commendable exception is McGee (1991), which I tackle now. Actually, McGee doesn't argue for the importance of *ordinary* language, but of semantic closure directly. He writes:

> If we adopt the [stratified conception], we shall find that within the object language we are unable even to describe human thought and action… intentional human activities, such as speaking, believing, willing and acting will be indescribable and inexplicable. Thus, if we accept the limitations imposed by Tarski's proposal for avoiding antinomies, we forfeit one of the highest aspirations of the human spirit, the aspiration to self-understanding. (p.79)

I take McGee's worry seriously. I agree that self-understanding is one of the principal goals of philosophy, and that if the stratified conception were shown to force us to forfeit this goal, that would be a mighty argument against it. But this is not the case. Quite the contrary, I think that with the stratified conception we are making an important step in the achievement of this goal. The project that McGee is hinting at, that of describing intentional human activities such as speaking, willing etc., is (or is intimately related to) the project of giving a philosophical account of the propositional attitudes. McGee implicitly assumes that a model of the reflective aspects of human subjectivity, i.e. a theory of propositional attitudes, requires a semantically closed language. But this assumption is unfounded. The distinctive feature of propositional attitude reports is that they represent the linguistic or cognitive agency of the attitude holder. The stratified conception does not preclude such a representation; it complicates it by showing that it requires stratification and an appeal to abstract discourse. If this is correct, then the stratified conception does not block the way to human self-understanding, but rather opens it up by blocking an apparent shortcut, semantic closure, that in fact leads nowhere.

Why have philosophers been inclined to hold the normative premise? I can think of three contrasts that might mistakenly be associated with the contrast between ordinary and regimented language. The first, least likely, is the one between interpreted and merely formal language. This association is easy to resist when paying attention, but it might be that the term "formalized languages", and the kinship between the concept of truth and the concepts of logic, put on us a subtle but constant pressure to associate the

[225] In §7.3 we saw how insensitivity to the ambiguity of the term "language" can lead to fallacies.

two contrasts. This is why, as a matter of terminological hygiene, I prefer to say "regimented languages".

The second contrast is the one between the natural and the artificial, suggested by the term "natural language".[226] In general, the "natural" enjoys the positive connotation of being more "real" than the artificial. But in the present case this doesn't hold. Although regimented languages are indeed "man-made" in the sense that they are a product of convention and will, this confers on them *more* philosophical relevance, not less. Using a regimented language, like, say, speaking under oath, carries with it more responsibility and commitment on the part of the user than using unregimented discourse. A theory of truth should therefore apply to them first and foremost. The negative connotation of "artificial" is something that we need to resist.

A more potent confound is the superficially similar contrast that ordinary language philosophers such as Ryle and Austin, or, in a different way, Wittgenstein, make between language in its everyday use, and a philosophical use of words which, by taking them out of context, empties them of their meaning and creates philosophical pseudo-problems. At least some thinkers might be confused by this contrast and the one between regimented and unregimented discourse.[227] Here is a famous passage from Wittgenstein (1958) that expresses this worry:

> When philosophers use a word – "knowledge", "being", "object", "I", "proposition", "name" – and try to grasp the *essence* of the thing, one must always ask onself: is the word ever actually [*tatsächlich*] used in this way in the language-game [Sprache] which is its original home [Heimat]? – What *we* do is to bring words back from their metaphysical to their everyday use. (§116)[228]

Though we can certainly find cases that fit Wittgenstein's description of philosophical practice, the use of regimented discourse is emphatically not one of them. Regimentation is not about trying to grasp "essences" of things. It is the rather humdrum procedure of accepting, in advance of discourse, certain norms of expressions that help avoid vagueness and equivocation. This procedure is used in legal matters, in science, in engineering, and in every field in which it is important to get things right. Philosophy is one of these fields.

In view of these potential pitfalls that the normative premise faces, I propose that we wait until it is stated and defended in a more detailed and explicit manner before we take it seriously.

---

[226] See Strawson's quote above.
[227] See Kripke's quote above.
[228] I gave the original German where I thought that the English translation misses something important. In particular, the German noun "Heimat", which has no exact English translation, is an extremely loaded term, opposed to terms such as "exile" and "alienation". It carries romantic and mythical connotations that suggest that Wittgenstein's use of the phrase "the everyday use of words" is anything but an everyday use.

So much for my misgivings concerning the normative premise. The situation is not much better with the descriptive premise, the claim that ordinary language is semantically closed. It too is rarely spelled out, and practically never argued for. I see two possible reasons why such a claim could be relied on so often without philosophical argument: either it is self-evident, or it is an empirical fact. In the former case, ordinary language is conceived as something the nature of which is given to our consciousness unmediatedly, perhaps in the form of "intuitions"; on the latter, ordinary language is an empirical object to be discovered by experiments and observations. Either way, I contend, the descriptive premise is ill-founded.

Let's look first at the claim that the semantic closure of natural language is an empirical fact. This might make a *philosophical* argument unnecessary, but reference to some discussion in print of this empirical discovery is surely in order. To the best of my knowledge, there is no such empirical result.[229] It is eminently plausible that most empirical linguists, if asked whether they think natural language is semantically closed, would say yes. But this doesn't make it an empirical fact. If we look at the (truth-based) semantic theories themselves, for example as reflected in standard textbooks, and assume that their object language is semantically closed, we can quite easily derive contradictions from them. The implicit working assumption of linguists is therefore that natural language is *not* semantically closed.[230]

Perhaps what is meant by the claim that semantic closure is an empirical fact is not that there is published research with that conclusion, but that it is an *obvious* empirical fact. We all, as competent speakers of natural language, have intuitions of its semantic closure. And since speaker intuitions are the primary empirical data for linguistics, it *should* be considered an empirical fact that natural language is semantically closed.

This argument misrepresents the way in which intuitions are used in empirical linguistics. Speakers are not consulted for their judgments about theoretical questions concerning language and its nature, but only about particular sentences or constructions. It is probably the case that speakers will tend to accept all sentences of the form:

(17) "snow is white" is true if and only if snow is white,

---

[229] The closest is Arne Næss's very comprehensive survey of the opinions of non-philosophers about the nature of truth. Næss concludes that there is no single common-sense conception of truth, though the T-sentences are indeed accepted, as trivial, by a significant majority. See Næss (1938a,b). Tarski refers to this research in (1944). See Ulatowski (2016) for a summary of the results and a brief discussion of the relation to Tarski.

[230] Textbooks in semantics generally take up a healthy insouciant attitude towards the semantic paradoxes. No mention is made of them in Chierchia and McGonnell-Ginet (1990), Heim and Kratzer (1998) and Jacobson (2014); they get a passing mention in Larson and Segal (1995, p.30fn).

but this in itself doesn't get us very far, for we don't know whether the sentence used on the right-hand side is the same sentence as that mentioned on the left-hand side, or just a homophonic synonym. And this is not a question we can ask our informant. The mere fact that T-sentences are intuitively unproblematic is therefore not a valid ground for the descriptive claim.

We turn to the view that says that the semantic closure of natural language is self-evident, and as such requires no support. This claim, if it is to escape the critique of the previous paragraph, must mean more than just having intuitions about the T-sentences. It is the very fact of semantic closure that should be intuited. The problem with intuitions of the sort appealed to in this claim is that it is not clear how to distinguish them from mere prejudice. Indeed, one sure way to recognize mere prejudice is to find out that it is false. Yet isn't this exactly what we did when we discovered that semantic closure is incoherent? I conclude that the view that says that it is self-evident that natural language is semantically closed is highly suspect, and cannot, on its own, ground the descriptive premise.[231]

The purpose of the foregoing was to argue that both premises of the argument against the stratified conception from the semantic closure of ordinary language are in need of clarification and grounding before the argument can be assessed. This does not refute the objection, but it gives the stratified conception some breathing space. It is ironic that the

objection from ordinary language was the first thing that Tarski sought to dissolve in *CTFL*. We have given in chapter 2 an updated version of his classical proof that ordinary language, if assumed to be universal, is inconsistent. Slightly less familiar are his comments from §6 of *CTFL*, directed expressly at philosophers who doubt the philosophical relevance of formalized languages:

> Philosophers who are not accustomed to use deductive methods in their daily work are inclined to regard all formalized languages with a certain disparagement, because they contrast these 'artificial' constructions with the one natural language – the colloquial language. For that reason the fact that the results obtained concern the formalized languages almost exclusively will greatly diminish the value of the foregoing investigations in the opinion of many readers. (p.267)

Later history shows that he should not have restricted his warnings to philosophers "not accustomed to use deductive methods in their daily work".

[231] One wonders why so many thinkers insist on holding on to something as flimsy as an intuition in the face of a refutation as decisive as a contradiction. It's as though the claim that ordinary language is semantically closed stands for a different idea altogether, a yearning after a return to a prediscursive unity of the subject. If this is so, then what the indefinability theorem shows us is that once the fruit of reflection has been eaten there is no return.

Bibliography

Abbott Barbara 1989: "Nondescriptionality and Natural Kind Terms", *Linguistics and Philosophy* 12: 269-291.

Achourioti T, Galinon H, Martínez Fernández J and K Fujimoto (eds.) (2015): *Unifying the Philosophy of Truth*. Springer.

Aczel Peter (1977): "An Introduction to Inductive Definitions", in Barwise Jon (ed.): *Handbook of Mathematical Logic*. Elsevier.

Almog, Perry and Wettstein (eds.) (1989): *Themes from Kaplan*. Oxford: Oxford University Press.

Atkin Albert (2005): "Peirce on the index and indexical reference", Transactions of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy 41:1, 161-188.

Awodey Steven and A.W. Carus (2007): "Carnap's Dream: Gödel, Wittgenstein and *Logical Syntax*", *Synthese* 159:1, 23-45.

Barwise Jon and John Etchemendy (1987): *The Liar: An Essay on Truth and Circularity*. Oxford: Oxford University Press.

Bar-Hillel Yehoshua (1954) "Indexical Expressions", *Mind* 63:251, 359-379.

Bennet Jonathan (1967): "Review of Encarnacion, Ushenko, Toms, Donnellan, etc.", *Journal of Symbolic Logic* 32:1, 108-112.

Bett, Richard (ed.) (2005): *Sextus Empiricus: Against the Logicians*. Cambridge: Cambridge University Press.

Black, Max (1948): "The Semantic Definition of Truth" *Analysis* 8:4, 49-63.

Boolos George (1995): "Quotational Ambiguity", in Leonardi and Santambrogio (eds.): *On Quine: New Essays*. Cambridge: Cambridge University Press.

Bromberger Sylvain and Morris Halle (2000): "The Ontology of Phonology (Revised)", in Burton-Roberts, Carr and Docherty (eds.): *Phonological Knowledge: Conceptual and Empirical Issues*. Oxford: Oxford University Press.

Burge Tyler (1979): "Semantical Paradox", *Journal of Philosophy*, 76:4, 169-198.

Burks Arthur (1949): "Icon, Index, and Symbol", *Philosophy and Phenomenological Research* 9:4, 673-689.

Cappelen Hermann and Josh Dever (2013): The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person. Oxford: Oxford University Press.

Cappelen Hermann and Ernie Lepore (1997): "Varieties of Quotation", *Mind* 106:423, 429-450.

Carlson Greg (1991): "Natural Kinds and Common Nouns" in Von Stechow and Wunderlich (es.) (1991).

Carnap Rudolf (1947): *Meaning and Necessity*. Chicago: University of Chicago Press.

Chierchia Gennaro and Sally McConnell-Ginet (1990): *Meaning and Grammar: An Introduction to Semantics*. Cambridge, MA: MIT Press.

Chihara Charles (2003): "Five Puzzles about Mathematics in Search of Solutions", in Downey, Decheng, Ping, Hui and Yasugi (eds.): *Proceedings of the 7th and 8th Asian Logic Conferences*. World Scientific.

Chomsky Noam (1965): *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

----- (1995): *The Minimalist Program*. Cambridge, MA: MIT Press.

Copeland B. Jack (2002): "The Genesis of Possible Worlds Semantics", *Journal of Philosophical Logic* 31, 99-137.

Corcoran J, W Frank and M Maloney (1974): "String Theory", Journal of Symbolic Logic 39:4, 625-637.

Crane Tim (1990): "The Language of Thought: No Syntax Without Semantics", *Mind & Language* 5:3, 187-212.

David Marian (1994): Correspondence and Disquotation: An Essay on the Nature of Truth. Oxford: Oxford University Press.

----- (2008): "Tarski's Convention T and the Concept of Truth", in Patterson (ed.) (2008b).

Davidson Donald (1965): "Theories of Meaning and Learnable Languages", in Bar-Hillel Yehoshua (ed.): *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science*. North Holland. Reprinted in Davidson (1984).

----- (1967): "Truth and Meaning", *Synthese* 17:1, 304-323. Reprinted in Davidson (1984).

----- (1984): Inquiries into Truth and Interpretation. Oxford: Clarendon.

Devidi David and Graham Solomon (1999): "Tarski on 'essentially richer' metalanguages", *Journal of Philosophical Logic* 28:1, 1-28.

Devitt Michael (1981): *Designation*. New York: Columbia University Press.

----- (1984): *Realism and Truth*. Princeton: Princeton University Press.

Donnellan Keith (1966): "Reference and Definite Descriptions," *Philosophical Review*, 75:3 281-304.

Drange Theodore (1969): "Paradox Regained", Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 20:4, 61-64.

Dummett Michael (1959): "Truth", *Proceedings of the Aristotelian Society* 59:1, 141-162. Reprinted in Dummett (1978).

------ (1978): *Truth and Other Enigmas*. Cambridge, MA: Harvard University Press.

------ (1991): *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.

Dyke Heather (2012): "Propositions: Truth vs. Existence", in Maclaurin (ed.): *Rationis Defensor*. Springer.

Eklund Matti (2002): "Inconsistent Languages", *Philosophy and Phenomenological Research* 64:2, 251-275.

Elbourne Paul (2005): *Situations and Individuals*. Cambridge, MA: MIT Press.

Enderton Herbert (2001): *A Mathematical Introduction to Logic*. Academic Press.

Etchemendy John (1988): "Tarski on Truth and Logical Consequence" *Journal of Symbolic Logic*, 53:1, 51-79.

Evans Gareth (1982): *The Varieties of Reference*. Oxford: Oxford University Press.

Feferman Anita and Solomon Feferman (2004): *Tarski: Life and Logic*. Cambridge: Cambridge University Press.

Field Hartry (1974): "Quine and the Correspondence Theory", *The Philosophical Review* 83:2, 200-228.

----- (1972): "Tarski's Theory of Truth", *Journal of Philosophy* 69:13, 347-375.

----- (1994): "Deflationist Views of Meaning and Content", *Mind* 103:411, 249-285.

Fine Kit (1983): “A Defence of Arbitrary Objects”, *Proceedings of the Aristotelian Society, Supplementary Volumes* 57, 55-89.

----- (1985): Reasoning with Arbitrary Objects. Blackwell.

----- (2003): "The role of variables", *Journal of Philosophy* 100:12, 605-631.

----- (2007): *Semantic Relationism*. Blackwell.

Fitting Melvin (2015): “Intensional Logic”, *The Stanford Encyclopedia of Philosophy*, Zalta Edward (ed.). URL = <https://plato.stanford.edu/archives/sum2015/entries/logic-intensional/>

Frost-Arnold, Greg (2004). Was Tarski’s theory of truth motivated by physicalism? *History and Philosophy of Logic*, 25(4):265–280.

----- (2008): “Tarski’s Nominalism”, in Patterson (ed.) (2008a).

Gaifman Haim (1992): “Pointers to Truth”, *Journal of Philosophy* 89:5, 223-261.

Garcia-Carpintero Manuel (1998): “Indexicals as Token-Reflexives”, *Mind* 107:427, 529-563.

Gauker Christopher (2006): “Against Stepping Back: A Critique of Contextualist Approaches to the Semantic Paradoxes”, *Journal of Philosophical Logic* 35, 393-422.

Geurts Bart (1997): “Good News about the Description Theory of Names”, *Journal of Semantics* 14, 319-348.

Glanzberg Michael (2001): “The Liar in Context”, *Philosophical Studies* 103, 217-251.

----- (2004a): “A Contextual-Hierarchical Approach to Truth and the Liar Paradox”, *Journal of Philosophical Logic*. 33:1, 27-88.

----- (2004b): “Truth, Reflection, and Hierarchies”, *Synthese* 142, 289-315.

----- (2014): “Explanation and Partiality in Semantic Theory”, in Burgess and Sherman (eds.). *Metasemantics: New Essays on the Foundations of Meaning*. Oxford: Oxford University Press.

----- (2015): “Complexity and Hierarchy in Truth Predicates”, in Achourioti et al. (eds.) (2015).

Geach Peter T (1972): *Logic Matters*. Oxford: Basil Blackwell.

Goodman Nelson and Willard Quine (1947): "steps toward a constructive nominalism", *Journal of Symbolic Logic* 12:4, 105-122.

Gupta Anil (1982): "Truth and Paradox", *Journal of Philosophical Logic* 11:1, 1-60.

----- (2011a): *Truth, Meaning, Experience*. Oxford: Oxford University Press.

----- (2011b): "A Critique of Deflationism", in (2011a). Originally published in *Philosophical Topics* (1993) 21:2, 57-81.

----- (2011c): "An Argument against Tarski's Convention T", in (2011a). Originally published in Schantz (2002): *What is Truth?* De Gruyter.

Gupta Anil and Nuel Belnap (1993): *The Revision Theory of Truth*. Cambridge, MA: MIT Press.

Haas-Spohn Ulrike (1995): *Versteckte Indexicalität und Subjektive Bedeutung*, Berlin: Akademie-Verlag.

Hale, Wright and Miller (eds.) (2017): A companion to the philosophy of language (2nd edition). Wiley & Sons. (1st edition 1997, Blackwell).

Hawthorne John and David Manley (2012): *The Reference Book*. Oxford: Oxford University Press.

Heim Irene and Angelika Kratzer (1998): *Semantics in Generative Grammar*. Blackwell.

Heck Richard G (1997): "Tarski, Truth and Semantics," *The Philosophical Review* 106, 533-554.

Hempel Carl (1935): "On the Logical Positivists' Concept of Truth", *Analysis* 2:4, 49-59.

Hodges Wilfrid (1986): "Truth in a Structure", *Proceedings of the Aristotelian Society*, New Series 86, 135-151.

----- (2004): "What Languages have Truth Definitions?", *Annals of Pure and Applied Logic* 126:1-3, 93-113.

----- (2008): "Tarski's Theory of Definition", in Patterson (ed.) (2008a).

Horsten Leon (2011): The Tarskian Turn: Deflationism and Axiomatic Truth. Cambridge, MA: MIT Press.

Horwich Paul (1998): *Truth* (2nd edition). Oxford: Clarendon Press.

Israel David and John Perry (1996): "Where Monsters Dwell", in Seligman and Westerståhl (eds.): *Logic, Language and Computation*, Volume 1. Stanford: Stanford University Press.

Jacobson Pauline (2014): Compositional Semantics: An Introduction to the Syntax/Semantics Interface. Oxford: Oxford University Press.

Juhl Cory (1997): "A Context Sensitive Liar", *Analysis* 57:3, 202-204.

Kaplan David (1989a): "Demonstratives", in Almog, Perry and Wettstein (eds.), pp.481-563.

----- (1989b): "Afterthoughts", Almog, Perry and Wettstein (eds.), pp. 565-614.

Kleene Stephen Cole (1952): *Introduction to Metamathematics*. New York: Van Nostrand.

Kneale William (1962): "Modality, De Dicto and De re", in Nagel, Suppes and Tarski (eds.): *Logic, Methodology and the Philosophy of Science: Proceedings of the 1960 International Congress*. Stanford: Stanford University Press.

Kneale William and Martha Kneale (1962): *The Development of Logic*. Oxford: Oxford University Press.

Kokoszyńska Maria (1936): "Über den absoluten Wahrheitsbegriff und einige andere semantische Begriffe", *Erkenntnis* 6, 143-165.

Kotarbiński Tadeusz (1929): Elementy Teorii Poznania Logiki Formalnej i Metodologii Nauk. Warszawa: Ossolineum.

Kreisel Georg (1967): "Informal Rigour and Completeness Proofs", in Lakatos (ed.): *Problems in the Philosophy of Mathematics*. North-Holland.

Kripke Saul (1975): "Outline of a Theory of Truth", *Journal of Philosophy*, 72:19, 690-716.

----- (1980): *Naming and necessity*. Harvard University Press. Originally published in Davidson and Harman (eds.) (1971): *The Semantics of Natural Language*. Springer.

Künne Wolfgang (2003): *Conceptions of Truth*. Oxford: Oxford University Press.

Larson Richard and Gabriel Segal (1995): *Knowledge of Meaning: An Introduction to Semantic Theory*. Cambridge, MA: MIT Press.

Leeds Stephen (1978): "Theories of Reference and Truth", *Erkenntnis* 13:1, 111-129.

Lewis David (1968): “Counterpart Theory and Quantified Modal Logic”, *Journal of Philosophy* 65:5, 113-126.

----- (1970): “General Semantics”, *Synthese* 22:1, 18-67.

----- (1975): “Languages and Language” in Gunderson (ed.): *Language, Mind and Knowledge*. University of Minnesota Press.

----- (1979): “attitudes de dicto and de se”, *The Philosophical Review* 88:4, 513-543.

----- (1980): “Index, Context, Content”, in Kanger and Ohman (eds.): *Philosophy and Grammar*. Reidel.

----- (1986): *On the plurality of worlds*. Oxford: Oxford University Press.

Liao Shen-Yi (2012): “What are Centered Worlds?”, *The Philosophical Quarterly*, 62:247, 294-316.

Leonard Henry and Nelson Goodman (1940): “The Calculus of Individuals and its Uses” *Journal of Symbolic Logic*, 5:2, 45-55.

Lowe E.J. (2017): “Objects and Criteria of Identity”, in Hale, Wright, Miller (eds.) (2017). pp. 990-1012.

Lynch Michael (ed.) (2001): *The Nature of Truth*. Cambridge, MA: MIT Press.

Magidor Ofra (2015): “the myth of the de se”, *Philosophical Perspectives* 29:1, 249-283.

Mancosu Paul (2008): “Tarski, Neurath and Kokoszyńska on the Semantic Conception of Truth”, in Patterson (ed.) (2008a).

Martin Richard (1964): “The Philosophical Import of Virtual Classes”, *Journal of Philosophy* 61:13, 377-387.

Martin Robert (ed.) (1970): *The Paradox of the Liar*. New Haven: Yale University Press.

McDowell John (1978): “Physicalism and Primitive Denotation: Field on Tarski”, *Erkenntnis,* 13:1, 131-152.

McGee Vann (1991): *Truth, Vagueness and Paradox*. Hackett.

Mendelson Elliot (1997): *Introduction to Mathematical Logic* (4th edition). Chapman & Hall.

Menzel, Christopher (2016): "Possible Worlds", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/possible-worlds/>.

Montague Richard (1963): “Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability”, *Acta Philosophica Fennica* 16, 153-67. Reprinted in Montague (1974), 287-302.

----- (1968): "Pragmatics", in Klibansky (ed.): *Contemporary Philosophy: A Survey*. La Nuova Italia Editrice. Reprinted in Montague (1974), pp. 95-118.

----- (1970a): "Pragmatics and Intensional Logic", *Synthèse* 22, 68-94. Reprinted in Montague (1974), pp. 119-147.

----- (1970): "English as a Formal Language", in Visentini Bruno et al. (eds.): *Linguaggi nella Società e nella Tecnica*. Edizioni di Comunità, Milan. Reprinted in Montague (1974).

----- (1974): Formal Philosophy: Selected Papers of Richard Montague (edited by Richmond Thomason). Yale University Press.

Moschovakis Yiannis (1974): *Elementary Induction on Abstract Structures*. North-Holland Publishing Company.

Mueller Ian (1981): Philosophy of Mathematics and Deductive Structure in Euclid's Elements. Cambridge, MA: MIT Press.

Muraswki Roman and Jan Woleński (2008): "Tarski and his Polish Predecessors on Truth", in Patterson (ed.) (2008a).

Murzi Julien and Carrara Massimiliano (2015): "Paradox and Logical Revision. A Short Introduction", *Topoi* 34:1, 7-14.

Odden David (2005): *Introducing Phonology*. Cambridge: Cambridge University Press.

Pap Arthur (1954): "Propositions, Sentences, and the Semantic Definition of Truth", *Theoria* 20:1-3, 23-35.

Parsons Charles (1971): "A Plea for Substitutional Quantification", *Journal of Philosophy* 68:8, 231-237.

Patterson Douglas (2002): "Theories of Truth and Convention T", *Philosophers Imprint*. 2:5, 1-16.

----- (2003): "What is a Correspondence Theory of Truth?" *Synthese* 137:3, 421-444.

----- (2005): "Deflationism and the Truth Conditional Theory of Meaning", *Philosophical Studies* 124:3, 271-294.

----- (2006): "Tarski on the Necessity Reading of Convention T", *Synthese* 151:1, 1-32.

----- (ed.) (2008a): *New Essays on Tarski and Philosophy*. Oxford: Oxford University Press.

----- (2008b): "Tarski's Conception of Meaning", in Patterson (ed.) (2008a).

----- (2008c): "Truth-definitions and Definitional Truth", *Midwest Studies in Philosophy* 32, 313-328.

----- (2009): "Inconsistency Theories of Semantic Paradox", *Philosophy and Phenomenological Research* 79:2, 387-422.

----- (2012): Alfred Tarski: Philosophy of Language and Logic. Palgrave Macmillan.

Perry John (1979) "the problem of the essential indexical", *Noûs* 3-21.

----- (2017) "the semantics and pragmatics of indexicals" in Hale, Wright and Miller (eds.) (2017), pp. 970-989.

Pietroski Paul (2005) "Meaning before Truth" in Preyer and Peter (eds.) (2005).

Popper, Karl (1955): "A Note on Tarski's Definition of Truth", *Mind* 64:255, 388-391.

----- (1979): Objective Knowledge: An Evolutionary Approach. Oxford: Clarendon Press.

Posy Carl (2015): "Realism, Reference, and Reason: Remarks on Putnam and Kant", in Auxier, Anderson and Hahn (eds.): *The Philosophy of Hilary Putnam,* in the Library of Living Philosophers series. Open Court.

----- (forthcoming): "Of Griffins and Horses: Mathematics, Metaphysics and Kant's Critical Turn", in Posy and Rechter (eds.): *Kant's Philosophy of Mathematics, Vol. 1*. Cambridge: Cambridge University Press.

----- (unpublished): "Systematicity and Semantics"

Potter Michael (2004): *Set Theory and its Philosophy: A Critical Introduction*. Oxford: Oxford University Press.

Prauss Gerold (1969): "Zum Wahrheitsproblem bei Kant", *Kant-Studien* 60:2, 166-182.

Predelli Stefano (2011): "I'm Still Not Here Now", *Erkenntnis* 74:3, 289-303.

Preyer Gerhard and Georg Peter (eds.) (2005): *Contextualism in Philosophy: Knowledge, Meaning and Truth*. Oxford: Clarendon Press.

Priest Graham (1984): "Semantic Closure", *Studia Logica* 43:1, 117-129.

----- (2006): *In Contradiction (2nd edition)*. Oxford: Oxford University Press. (1st edition 1987, Martinus Nijhoff).

Pullum Geoffrey and Barbara Scholz (2010): "Recursion and the Infinitude Claim", in Van der Hulst (ed.): *Recursion and Human Language*. Mouton de Gruyter

Putnam Hilary (1985): "A Comparison of Something with Something Else" *New Literary History* 17:1, 61-79.

Quine Willard van Orman (1936): "Definition of Substitution", *Bulletin of the American Mathematical Society* 42:8, 561-569.

----- (1945): "On Ordered Pairs", *Journal of Symbolic Logic* 10:3, 95-96.

----- (1946): "Concatenation as a Basis for Arithmetic", *Journal of Symbolic Logic* 11:4, 105-114.

----- (1951): "Ontology and Ideology", Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition. 2:1, 11-15.

----- (1960): *Word and Object*. Cambridge, MA: MIT Press.

----- (1961): *From a Logical Point of View (2nd edition)*. 1st ed. (1953). Cambridge, MA: Harvard University Press.

----- (1962): Set Theory and its Logic. Belknap

----- (1969): *Ontological Relativity and Other Essays*. New York: Columbia University Press..

----- (1977): "Intensions Revisited", *Midwest Studies in Philosophy* 2:1, 5-11.

----- (1981): *Mathematical Logic (revised edition)*. 1st ed. (1940). Cambridge, MA: Harvard University Press.

----- (1987): *Quiddities*. Cambridge, MA: Harvard University Press.

Rattan Gupreet (2004): "The Theory of Truth in the Theory of Meaning," *European Journal of Philosophy* 12:2, 214-243.

----- (2005): "Davidson, Semantic Deflationism, and Dummett's Dilemma", *Iyyun: The Jerusalem Philosophical Quarterly* 54, 39-63.

Ray Greg (2005): "On the Matter of Essential Richness", *Journal of Philosophical Logic* 34:4, 433-457.

Rayo Agustín and Gabriel Uzquiano (2006): *Absolute Generality*. Oxford: Oxford University Press.

Richard Mark (1986): "Quotation, Grammar, and Opacity", *Linguistics and Philosophy* 9:3, 383-403.

Rozeboom William (1958): "Is Epimenides Still Lying?", *Analysis* 18:5, 105-113

Rojszczak Artur (2002): "Philosophical Background and Philosophical Content of the Semantic Definition of Truth", *Erkenntnis* 56, 29-62.

----- (2006): From the Act of Judging to the Sentence: the Problem of the Truth Bearers from Bolzano to Tarski. Springer.

Rosen, Gideon, "Abstract Objects", *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2017/entries/abstract-objects/

Russell Bertrand (1911): "Knowledge by Acquaintance and Knowledge by Description", *Proceedings of the Aristotelian Society* 11, 108-128.

----- (1912): *The Problems of Philosophy*. Williams and Norgate.

----- (1972): The Philosophy of Logical Atomic. Fontana.

Ryll-Nardzewski C. (1953): "The Role of the Axiom of Induction in Elementary Arithmetic" *Fundamenta mathematicae* 39, 239-263.

Salmon Nathan (1982): *Reference and Essence*. Blackwell.

Schantz Richard (ed.) (2002): *What is Truth?* De Gruyter.

Scharp Kevin (2013): *Replacing Truth*. Oxford: Oxford University Press.

Schlenker Phillipe (2003): "A Plea for Monsters", *Linguistics and Philosophy* 26:1, 29-120.

----- (2011): "Indexicality and *De Se* Reports", in von Heusinger, Maienborn and Portner (eds.): *Semantics: An International Handbook of Natural Language Meaning. Volume 2*. De Gruyter Mouton.

Serény György (2006): "The Diagonal Lemma as the Formalized Grelling Parado*x"* in Baaz and Preining (eds.): *Gödel Centenary*, Collegium Logicum vol. 9, Kurt Gödel Society, Vienna, 2006, pp. 63-66.

Sher Gila (1999): "What is Tarski's Theory of Truth?" *Topoi* 18, 149-166.

Sher Gila (2016): "Substantivism about Truth", *Philosophy Compass* 11:12, 818-828.

Simmons Keith (1993): *Universality and the Liar*. Cambridge: Cambridge University Press.

Simons Peter (2009): "Twardowski on Truth", The Baltic International Yearbook of Cognition, Logic and Communication 4, 1-14.

Soames Scott (1999): *Understanding Truth*. Oxford: Oxford University Press.

----- (2002): Beyond Rigidity: The Unfinished Agenda of Naming and Necessity. Oxford: Oxford University Press.

Stephanou Yannis (2001): "Indexed Actuality", *Journal of Philosophical Logic* 30:4, 355-393

Stanley, Jason (2008): "Philosophy of Language in the Twentieth Century", in *the Routledge Companion to Twentieth Century Philosophy*: 382-437.

----- (2017): "Names and Rigid Designation", in Hale, Wright and Miller (eds.) (2017), pp. 920-947.

Stanley Jason and Zoltan Gendler Szabo (2000): "On Quantifier Domain Restriction", *Mind & Language*, 15:2-3, pp.219-261.

Strawson P.F (1949): "Truth", *Analysis* 9:6, 83-97.

----- (1950): "On Referring", *Mind* 59:235, 320-344.

Tarski Alfred (1933) [*CTFL*]: *Pojęcie prawdy w językach nauk dedukcyjnych*. Warszawa: Nakładem Towarzystwa Naukowego Warszawskiego. German translation in Tarski (1936a); English translation under the title "The Concept of Truth in Formalized Languages" in Tarski (1956). Page numbers are to the English edition.

----- (1936a): "Der Wahrheitsbegriff in den formalisierten Sprachen", *Studia Philosophica* 1, 261-405.

----- (1936b): "O pojęciu wynikania logicznego", *Przegląd Filozoficzny* 39, 58-68. English translation under the title "On the Concept of Logical Consequence" in Tarski (1956).

----- (1944): "The Semantic Conception of Truth: and the Foundations of Semantics", *Philosophy and Phenomenological Research*, 4:3, 341-376.

----- (1956): *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. Translated by J.H. Woodger. Oxford: Clarendon Press.

------ (1992): "Alfred Tarski: Drei Briefe an Otto Neurath", *Grazer Philosophische Studien* 43, 1-32.

Tarski Alfred and John Corcoran (1986): "What are Logical Notions?" *History and Philosophy of Logic* 7:2, 143-154.

Thomason Richmond (1975): "Necessity, Quotation, and Truth: An Indexical Theory", in Kasher (ed.): *Language in Focus: Foundations, Methods and Systems*. Springer. pp. 119-138.

Ulatowski Joseph (2016): "Ordinary Truth in Tarski and Næss", In Kuzniar & Odrowąż-Sypniewska (eds.), *Uncovering Facts and Values*. Brill.

Ushenko Andrew P (1937): "A New 'Epimenides'", *Mind* 46:184, 549-550.

Van Fraassen Bas (1966): "Singular Terms, Truth-Value Gaps, and Free Logic", *Journal of Philosophy* 63:17, 481-495.

------ (1968): "Presupposition, Implication, and Self-Reference", *Journal of Philosophy* 65:5, 136-152.

------ (1970): "Rejoinder: On a Kantian Conception of Language", in Martin (ed.) (1970).

Von Fintel, Kai and Irene Heim (2011): *Lecture Notes in Intensional Semantics*, available at http://mit.edu/fintel/fintel-heim-intensional.pdf.

Von Stechow Arnim and Dieter Wunderlich (1991): Semantik/Semantics: ein internationaler Handbuch der zeitgenössichen Forschung. Berlin.

Wang Hao (1986): Beyond Analytic Philosophy: Doing Justice to What We Know. Cambridge, MA: MIT Press.

Weyl Hermann (1946): "Mathematics and Logic. A brief survey serving as a preface to a review of '*The Philosophy of Bertrand Russell*'", *American Mathematical monthly* 53:1, 2-13.

Whitehead and Russell (1925): *Principia Mathematica, Volume 1 (2nd edition)*. Cambridge: Cambridge University Press.

Williams Michael (1999): "Meaning and Deflationary Truth" *Journal of Philosophy* 96:11, 545-564.

Wittgenstein Ludwig (1953): Remarks on the Foundations of Mathematics. Blackwell.

------ (1958): *Philosophical Investigations (2nd Edition)*. 1st Edition 1953. Blackwell.

Woleński Jan and Peter Simons (1989): "De Veritate: Austro-Polish Contributions to the Theory of Truth from Brentano to Tarski", in Szaniawski (ed.): *The Vienna Circle and the Lvov-Warsaw School*. Nijhoff.

Zimmerman Thomas Ede (1997): “The addressing Puzzle” in Künne, Newen and Anduschus (eds.): *Direct Reference, Indexicality and Propositional Attitudes*. Stanford: Stanford University Press.

הוראה מוחשית, אבל במחיר של הגבלה של תחום הדיון. הבחנה זו היא בעלת השלכות חשובות להבנה של מושגי האמת, המובן והלשון. ;

תקציר

חיבור זה מכיל ניתוח פילוסופי של מושג האמת. זהו פיתוח של והגנה על התפיסה "המדורגת" של אמת, לפיה הגדרת מושג האמת עבור שפה יכולה להתבצע רק מתוך שפה עשירה יותר. מקורה של תפיסה זו במונוגרפיה של אלפרד טרסקי (1933) בשם **מושג האמת בשפות מוצרנות**. לעבודתו של טרסקי נודעה השפעה עמוקה, גם בפילוסופיה וגם בתחומים טכניים, אך הטענה הפילוסופית המרכזית שלה לא התקבלה בדרך כלל. לחיבור הנוכחי שתי מטרות עיקריות: א. להציע הצגה מפורטת ואנליטית של תורתו של טרסקי ושל הבעיות שעומדות בפניה; ב. לחפש פתרון לבעיות אלה ולפרש מה מתחייב מפתרון כזה.

החיבור מורכב משני חלקים גדולים. החלק הראשון מכיל הצגה מפורטת ופרשנית של תפיסת האמת המדורגת של טרסקי. ההצגה בנויה מהשלבים הבאים: א. הנחות יסוד חשובות, כגון ההגבלה של השיטה לשפות מוצרנות, ניתנות במפורש, ומשמעותן הפילוסופית נבחנת. ב. התוצאה השלילית העיקרית של השיטה, בדבר אי-אפשרות של שפה סגורה סמנטית, ניתנת בפרוטרוט ועם פרשנות. ג. הקריטריון לנכונות תוכנית של הגדרות, **מוסכמה א** (Convention T), מפורשת יחד עם השיקולים שמובילים אליה. ד. מבנה העומק של הגדרות אמת נחקר, כולל התנאים הנחוצים על מנת שהגדרה כזו תתאפשר. בפרט, פירוש מדויק ניתן למושג של "עושר לשוני". ה. לבסוף, מספר בעיות מועלות, בפרט **בעיית האחדות**, שלפיה אם בתפיסה המדורגת ניתנת הגדרה נפרדת לכל שפה מוצרנת, אזי לא מילאנו את המטלה שבה פתחנו – לספק הגדרה למושג האמת.

החלק השני של החיבור מפתח פתרון לבעיות שעומדות בפני התפיסה המדורגת. היסוד המרכזי בתשובה לבעיית האחדות הוא הטענה שמוסכמה א, הקריטריון לנכונות תוכנית, הוא הגורם שמאחד את הגדרות האמת השונות לגדי מושג יחיד. הבעיה התשובה זו היא שהיא מניחה שלמוסכמה א יש תוקף אוניברסלי, אך מהתוצאה השלילית של חלק א נובע שאין שפה אוניברסלית. המטלה בחלק ב' אפוא היא לפתח אופן ביטוי שיאפשר החלה אוניברסלית של מוסכמה א מבלי לדרוש שפה אוניברסלית. המטלה מבוצעת בשלבים הבאים: א. מוסכמה א מוצרנת על מנת לבודד את הצורך בכלי החדש. ב. התורה של אמצעי ביטוי חדש, בשם "כלליות מופשטת", מפותחת מתוך התורה הפרגמטית של ביטויים מצביעים (אינדקסיקלים) בשפה טבעית. ג. הכלי מותאם לשפות ללא ביטויים מצביעים, ומוסכמה א מנוסחת בעזרתו ללא ההנחה שיש שפה אוניברסלית. ג. המטלה של ניתוח מושגים מוכללת על מנת להכיל את התשובה לבעיית האחדות.

לעתים קרובות מועלה נגד תפיסת האמת של טרסקי הטענה שהיא בעלת משמעות פילוסופית מוגבלת. בחיבור זה אני טוען שההיפך הוא הנכון, ושהבעיות שתפיסה זו נתקלת בהן הן בעלות השלכות פילוסופיות כבדות משקל. מושג הכלליות המופשטת מביא להבחנה בין שני סוגי שיח שונים באופן יסודי. סוג אחד הוא שיח **מתודולוגי** מאפשר כלליות מוחלטת אבל במחיר של מופשטות; הסוג השני הוא שיח **תיאורטי** שמאפשר

עבודה זו נעשתה בהדרכתו של פרופ' קרל פוזי

הערה על מושג האמת (עם השלכות על מושגי המובן והלשון)

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

דוד קשתן

הוגש לסנט האוניברסיטה העברית בירושלים

נובמבר 2017