David Makinson

Gödel's Master Argument: what is it, and what can it do?

Gödel's Master Argument: What is it, and what can it do?

David Makinson
Department of Philosophy Logic and Scientific Method, London School of Economics

Abstract

This text is expository. We explain Gödel's 'Master Argument' for incompleteness as distinguished from the 'official' proof of his 1931 paper, highlight its attractions and limitations, and explain how some of the limitations may be transcended by putting it in a more abstract form that makes no reference to truth.

Keywords: Gödel, Master Argument, Incompleteness.

1 Introduction

Gödel's 'Master Argument' is sketched in his brief correspondence with Zermelo in late 1931. It is discussed in an influential 1984 article of Feferman [2], and may be found in books by several authors, most accessibly [8] and its website spin-off [9] . However, the argument is not as widely known as it should be, and its strengths and shortcomings compared to the 'official' proof appear not to have received much discussion. Moreover, an interesting abstraction on the Master Argument that overcomes some of the shortcomings, can be found only deep within the pages of specialist presentations such as [10] and [3], difficult to untangle from other material. The present article may thus be useful for those with limited time and energy but still wishing to have a proper understanding of what is going on in the Master Argument.

We begin by recalling the 1931 exchange of letters between Zermelo and Gödel, and itemize the background needed to continue reading. The Master Argument is then presented in its simplest available form, followed by a discussion balancing its attractions and limitations as well as an alleged philosophical weakness. We finally give a more abstract and powerful, but still easy, version of the Master Argument in which arbitrary 'oracles' take the place of 'truth in the intended model', thus transcending some of its limitations.

The author wishes to thank Jon Burton and Peter Smith for remarks on an ancestor of this text.

2 Autumn 1931

Gödel announced his incompleteness results in an abstract of 1930 and published them with proofs in his celebrated paper of 1931. Ernst Zermelo, already famous for his work on the axiom of choice and what we now call the Zermelo-Fraenkel axiomatization of set theory, read the 1931 paper and heard Gödel speak on it at a conference that summer. But he saw it as fatally flawed and ultimately not of great significance.

Both views appear to have stemmed from his disinterest in studying axiomatic systems that are formulated in finitary languages, *a fortiori* in doing so only by finitary means. Roughly speaking, Zermelo believed that we should be studying systems that embody broad swathes of mathematics, and that we should feel free to use any of the resources of mathematics in doing so. Both the formal systems studied and the reasoning used in that study could be infinitary along lines that he hoped, in the letters, to make precise at a later date.

This perspective evidently contrasts with that of Hilbert, which was adopted by Gödel in his published paper. The formal object-language that Gödel examines is defined by finite means and the investigation, conducted in a distinct and rather informal language, uses only finitary and constructive reasoning.

To be sure, in following decades logicians began relaxing these restrictions. Some investigated languages that are in one way or another infinitary, while others used free-wheeling methods with infinite sets, transfinite ordinals and the axiom of choice even when studying finitely generated systems. But in all cases they, like Hilbert and Gödel, continued to maintain a clear distinction between the system that is under study, formulated in an 'object-language', and the means used to study it, expressed in a (usually less formal) 'metalanguage'.

In contrast, Zermelo was unable or unwilling to make the distinction between object and metalanguage, and it seems to be that which led him to believe that Gödel's proofs harboured paradox. On 21 September 1931 he wrote to Gödel, hinting at his own general perspectives and outlining explicitly a contradiction that he claimed to have discovered in the paper.

Gödel replied on 12 October. He did not comment on the differences in general perspective, but responded in detail to the specific claim of paradox, carefully showing why his proof did not generate the contradiction that Zermelo thought he had found. At the same time, in an effort to help Zermelo see what was going on, he outlined the essence of his argument in a manner quite different from that of the version published earlier in the year. He did this again in an address in Princeton in 1934, but never elaborated it in print. After his death in 1978, the three letters constituting the Zermelo/Gödel exchange were found, published and translated. The

proof there sketched came to be known as 'the Master Argument'.[1]

3 Background needed

We assume that the reader is familiar with the notation of first-order logic, and has seen the standard first-order axiomatization of the arithmetic of the natural numbers. We write *PA* (Peano Arithmetic) for the axiomatization, *LPA* for its formal language, *N* for the set of all natural numbers themselves.

On the semantic level, we presume familiarity with the notion of a model for a first-order theory, the recursive definition of satisfaction/truth in a model, and the concepts of soundness and completeness of a given theory with respect to a given model. On the syntactic level, the concepts whose definitions should already be familiar are those of a sentence (closed formula) of the language, free and bound variables, the consistency and negation-completeness of an arbitrary first-order theory and, for the particular case of *PA*, the notion of ω-consistency. With that basis, the reader will be able to verify from the definitions the easy parts of Figure 1 (all of them for *PA* and some for arbitrary first-order theories) namely the three vertical arrows and, given them, the following interrelations between the arrows:

- The diagonal full arrow follows from the top one,
- The diagonal dotted arrow follows from the diagonal full one,
- The bottom arrow follows from the diagonal dotted one,
- Conversely (and a little less obviously), the diagonal dotted arrow follows from the bottom one (the verification of this will be recalled in Section 4).

4 The Master Argument

The Master Argument has two parts: an Inexpressibility Lemma and an Expressibility Lemma; its conclusion arises from the collision of the two.

[1] Who coined the term 'Gödel's Master Argument'? The author has not been able to determine this with certainty. It is used as if familiar in [8], and already appeared tentatively in the first edition of that book (2007). In response an inquiry from the present author, Smith recalled that he had been using the phrase for some time in lectures in Cambridge, but could not remember whether he devised it himself or took it from another source. Of course, the term 'master argument' had already been used for certain other celebrated, although highly contested, demonstrations. It was applied in Greek antiquity to an argument of Diodorus Cronos about future and necessity, and since 1974 has been used to highlight one of Berkeley's arguments about existence and the mind (see the relevant Wikipedia articles).

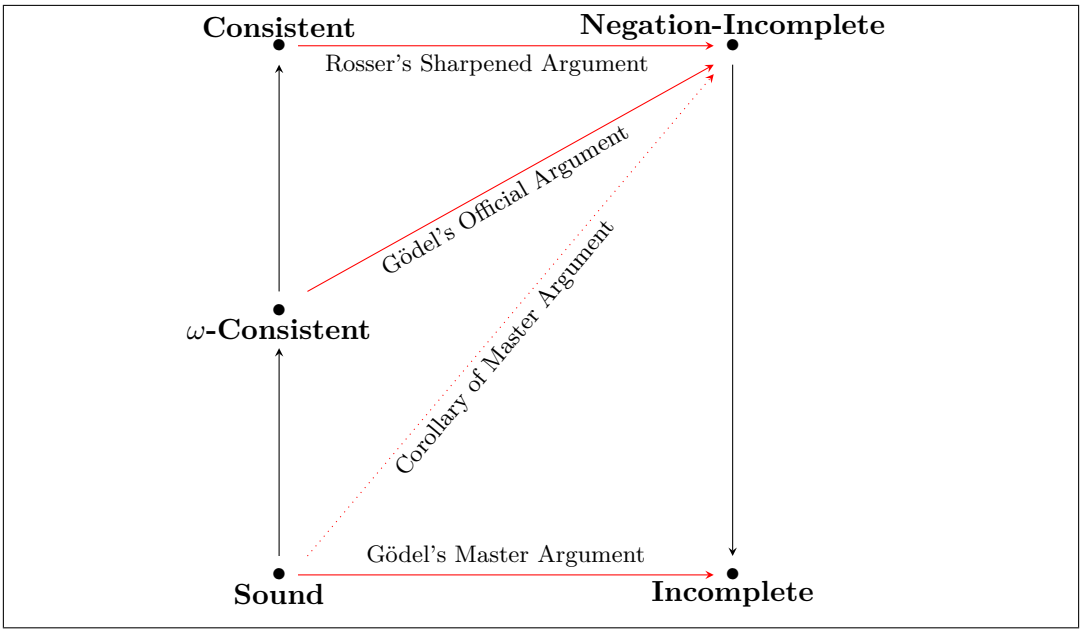

Figure 1: Gödel's first incompleteness theorem for *PA*

Definition 4.1. *A set $S \subseteq N$ is said to be* expressible *in LPA iff there is a formula $\varphi(x)$ of that language, with one free variable x, such that for all $n \in N$,*

$$n \in S \text{ iff } \varphi(\overline{n}) \text{ is true in the intended model for } PA$$

where $\overline{n}$ is the LPA numeral for n.

Fix any enumeration $\varphi_1, \varphi_2, \ldots, \varphi_i, \ldots, (i < \omega)$ of all the formulae in *LPA* whose sole free variable is x. Put D^+ to be the set of all natural numbers n such that $\varphi_n(\overline{n})$ is true in the intended model for *PA*, and let D^- be the set of all n such that $\varphi_n(\overline{n})$ is not true (i.e. false) in the same model. Clearly these two sets are complements of each other wrt. N, that is, $D^- = N \backslash D^+$ and $D^+ = N \backslash D^-$.

Lemma 4.2 (Inexpressibility Lemma[2])**.** *Neither D^- nor D^+ is expressible in the language of PA.*

[2]The formulation of the Inexpressibility Lemma 4.2 that is given here differs slightly from that sketched by Gödel in his letter to Zermelo, which has been followed in later presentations (e.g. Feferman, Smullyan, Fitting, Smith). On those accounts the lemma states the inexpressibility of truth itself (in other words, is exactly Tarski's Theorem), while on our account it states the inexpressibility of the sets D^+, D^-. Our formulation has the advantage that it simplifies the proof of the Inexpressibility Lemma 4.2, at the cost of then having to derive Tarski's Theorem from it

Proof. For D^-, suppose for reductio that it is expressible in *LPA*. Then by the definition of expressibility 4.1, there is a formula $\varphi(x)$ with x as sole free variable such that for all $n \in N, n \in D^-$ iff $\varphi(\overline{n})$ is true in the intended model. Now, $\varphi(x) = \varphi_k(x)$ for some $k \in N$. So, instantiating n to k we have: $k \in D^-$ iff $\varphi_k(\overline{k})$ is true in the intended model. But by the definition of D^-, we also have that $k \in D^-$ iff $\varphi_k(\overline{k})$ is not true in that model, giving a contradiction. Turning to D^+, it suffices to note that if D^+ is expressed by formula $\varphi(x)$ then D^- is expressed by $\neg\varphi(x)$. □

The second part of the Master Argument is a contrasting Expressibility Lemma. The sets D^+, D^- were defined using the notion of truth in the intended model of *PA*. We may also consider what happens if in the definitions we replace that notion by provability in the axiom system *PA*. Fix separate numberings of all formulae of *LPA* with just one free variable x, and of all derivations of *PA*. For brevity, write $|PA|$ for the set of all sentences that are theorems of *PA*. Put $D^+_{|PA|}$ to be the set of all natural numbers n such that $\varphi_n(\overline{n})$ is provable in *PA*, and let $D^-_{|PA|}$ be the set of all n such that $\varphi_n(\overline{n})$ is not provable in *PA*. Again these sets are complements of each other, so that one of them is expressible in the language of *PA* iff the other one is. But their behaviour is different from that of D^+, D^-. Indeed, as Gödel showed:

Lemma 4.3 (Expressibility Lemma)**.** *If PA is sound wrt. its intended model, then both $D^+_{|PA|}$ and $D^-_{|PA|}$ are expressible in the language of PA.*

Proof. (sketch) It suffices to show this for $D^+_{|PA|}$. Consider the relation that holds between a derivation δ_m and a formula $\varphi_n(x)$ with just one free variable x iff the former is a derivation of $\varphi_n(\overline{n})$. Then (as outlined by Gödel with more detailed verifications in later presentations, e.g. [9]), assuming that the enumerations are in a certain technical sense 'acceptable', this relation is primitive recursive and so is *captured* in *PA* by some formulae $\psi(y, x)$ in the following sense: for all $m, n \in N$,

1. if δ_m stands in the relation to $\varphi_n(x)$ then $\psi(\overline{m}, \overline{n}) \in |PA|$ and
2. if δ_m does not stand in the relation to $\varphi_n(x)$ then $\neg\psi(\overline{m}, \overline{n}) \in |PA|$.

Now suppose that *PA* is sound wrt. its intended model. We want to show that the formula $\exists y\psi(y, x)$ expresses $D^+_{|PA|}$. That is, we need to check that for all $n \in N$,

$$n \in D^+_{|PA|} \text{ iff } \exists y\psi(y, \overline{n}) \text{ is true in the intended model.}$$

as is done in Section 5 point iii. The more common formulation eliminates any need for the latter derivation, but at the cost of a more complex proof of inexpressibility. This is a small matter of trade-offs.

Left to right: Suppose $n \in D^+_{|PA|}$. Then by definition, $\varphi_n(\overline{n}) \in |PA|$. Hence there is a derivation δ_m of $\varphi_n(\overline{n})$, so δ_m stands in the relation to $\varphi_n(x)$ so, by (1), $\psi(\overline{m}, \overline{n}) \in |PA|$, so by first-order logic $\exists y \psi(y, \overline{n}) \in |PA|$. Thus by the supposition of soundness, $\exists y \psi(y, \overline{n})$ is true in the intended model as desired.

Right to left: Suppose $n \notin D^+_{|PA|}$. Then by definition, $\varphi_n(\overline{n}) \notin |PA|$. Hence there is no derivation δ_m of $\varphi_n(\overline{n})$, so no δ_m stands in the relation to $\varphi_n(x)$ so, by (2), $\neg\psi(\overline{m}, \overline{n}) \in |PA|$ for all $m \in N$. Thus by soundness, $\neg\psi(\overline{m}, \overline{n})$ is true in the intended model for each $m \in N$, hence $\forall y \neg\psi(y, \overline{n})$ is true in the intended model, so $\exists y \psi(y, \overline{n})$ is not true in the intended model, as required. □

The semantic incompleteness of *PA* emerges immediately from the collision between these two Lemmas.

Theorem 4.4 (Gödel's First Incompleteness Theorem (semantic version))**.** *If PA is sound wrt. its intended model then it is incomplete (wrt. the same).*

Proof. It is immediate from the two results that the set of all sentences of *PA* that are *true* in the intended model is not the same as the set of all *provable* sentences of *PA*. Recall that soundness means that the latter set is included in the former; completeness is the converse. Thus if *PA* is sound, it is not complete. □

Corollary 4.5. *If PA is sound with respect to its intended model then it is negation-incomplete.*

Proof. This is simply an application to *PA* of the fact, mentioned in the last bullet point of Section 3, that for arbitrary first-order theories and intended models, the bottom arrow of the diagram implies the diagonal dotted one. Details: Suppose *PA* is sound wrt. its intended model. Then by incompleteness there is a sentence φ of *LPA* such that φ is true in the intended model of the theory but is not derivable in the theory. Hence $\neg\varphi$ is false in the intended model, so by soundness it is not derivable in the theory. Thus neither φ nor $\neg\varphi$ is derivable in the theory, which is to say that it is negation-incomplete. □

5 Three attractions

We articulate three important attractions of the Master Argument, before looking at shortcomings in the following section.

i. A striking feature of the Argument is the way that it decomposes the proof of the incompleteness theorem into two contrasting lemmas, thus providing a

simple overall architecture. Moreover the diagonal argument for the first, 'negative' lemma is (in the present formulation) of the utmost simplicity, almost equal to that of Cantor's theorem in set theory. It is true that the second, 'positive' lemma remains long and tedious to check out in full detail, but at least the tedium is localized. Overall, one can say that the Master is the simplest and most transparent argument available for the semantic incompleteness, given soundness, of systems such as *PA*.

ii. Another feature of the proof is the prominent place that it gives to the notion of *expressibility* of subsets of N in the language *LPA* of Peano arithmetic, alongside the quite distinct notion of derivability of sentences from the axioms of *PA*. Since the formulation of the incompleteness theorem for *PA* speaks only of truth and provability, students easily assume that that is all it is about. Yet there are less expressive sub-languages of *PA* (indeed quite interesting ones) for which complete (and natural) axiomatizations are available. A well-known example is the sub-language in which zero, successor and addition remain but multiplication is absent.

   Expressive power and derivational power are thus two quite different capacities and can run in opposite directions. The former concerns the language alone and has nothing to do with provability from axioms; the latter is about which sentences, among those available in the language, are provable. The distinction is easily obscured by loose talk of the 'strength' of a system where it is left vague what kind of strength is meant. The salience that the Master Argument accords to the notion of expressibility has the merit of putting it on a par with that of provability.

iii. A third advantage of the Master Argument is that it reveals the close connection between Gödel's first incompleteness results and another celebrated theorem of mathematical logic: Tarski's 1933 theorem [11] on the indefinability of the notion of truth in *LPA* (or more expressive systems). We can state it (for *LPA*) as follows.

   Consider any enumeration $\psi_1, \psi_2, \ldots, \psi_i, \ldots, (i < \omega)$ of all sentences (formulae with no free variables) of *LPA*, and let T (the 'truth set') be the set of all $m \in N$ such that ψ_m is true in the intended model of *PA*. Then:

   Theorem 5.1 (Tarski's Theorem)**.** *T is not expressible in LPA.*

   Proof. (sketch)

We apply the Inexpressibility Lemma 4.2. By definition, $n \in D^+$ iff $\varphi_n(\overline{n})$ is true in the intended model for *PA*, which holds iff there is an m with $\psi_m = \varphi_n(\overline{n})$ and $m \in T$. Now, it is not difficult (though tedious) to show that the relation that holds between n and m iff $\psi_m = \varphi_n(\overline{n})$, is expressible in *LPA*, so if T is also expressible in *LPA* then D^+ must also be so, contrary to the Inexpressibility Lemma 4.2. Thus T is not expressible in *LPA*. □

6 Two limitations

However, as presented above the Master Argument has two important limitations.

i. It yields only the semantic version of Gödel's first incompleteness theorem (soundness implies incompleteness and thus also negation-incompleteness). It does not give us Gödel's stronger syntactic version (ω-consistency implies negation-incompleteness); nor the yet stronger syntactic version obtained by Rosser (plain consistency implies negation-incompleteness); nor again Gödel's second incompleteness theorem (to the effect that if *PA* is consistent then its consistency cannot be proven by means executable within the system itself).

   However this ceiling on content can be raised. In the next section we formulate a more powerful version of the Inexpressibility Lemma 4.2, in which 'truth in the intended model' is replaced by an arbitrarily chosen 'oracle', and show how this, when combined with a corresponding variant of the Expressibility Lemma 4.3, gives us Gödel's syntactic version of the incompleteness result (the diagonal full arrow in the diagram) by a proof just as short and transparent as before.

   Nevertheless, it must be conceded that there is no visible way of using such oracles to obtain a similar proof of the Rosser version (top arrow), nor of Gödel's second theorem (not in the diagram), so there still remains a ceiling, albeit rather higher, on what the argument can achieve.

ii. This limitation is related to the fact that The Master Argument is *not constructive*. As Gödel put it in his letter of 12 October 1931, "it furnishes no construction of the undecidable statement". That is to say, it does not exhibit a specific sentence of *LPA* that is both true in the intended model and underivable from the axioms of *PA* (under the supposition of soundness). Nor does it give us a recipe for constructing such a sentence – it merely guarantees that there must be one. In contrast, the proof in the 1931 paper is constructive in every detail.

Assessment of this feature will vary with one's philosophy of mathematics. Doctrinal constructivists take the view that non-constructive proofs are invalid: they fail to show existence and at best can be taken as heuristic encouragements for devising properly constructive proofs of their results. There are few mathematicians and philosophers who would take such a position, but any who do must see the Master Argument as incorrect, with the real proof being the considerably more complex constructive one.

On the other hand, one may be constructively inclined without being doctrinal about it. There are results that most logicians are quite happy to establish non-constructively, for example the completeness of classical first-order (or even propositional) logic using the Lindenbaum-Henkin method of maxi-consistent sets of formulae. Since the 1930s mathematicians and logicians have gradually become more comfortable working with explicitly non-constructive principles, in particular the axiom of choice.

So a less radical view of the situation is that non-constructive proofs are perfectly valid, but give less information than their constructive counterparts when the latter are available. If one sees the additional information as important for one's purposes – as will often be the case in computer science – then one might describe oneself as a 'light constructivist'. While granting that non-constructive arguments are valid and frequently shorter and more transparent than constructive ones, the light constructivist is happy to put up with additional complexity in constructive reasoning to get further information.

In the present instance, the additional information that can be provided by a constructive proof appears to be needed for going on to Rosser's improved version of the completeness theorem, and likewise for Gödel's second incompleteness theorem. Its absence from the Master Argument is thus an important limitation. However, it may be said that the non-constructive proof still deserves a place *alongside* constructive ones because of the attractions mentioned above – simplicity and transparency of architecture, salient role of the notion of expressibility, and close connection to Tarski's theorem. In the author's opinion, it is the version to teach to non-specialists seeking a good understanding within a limited time-frame.

7 A philosophical weakness?

A further shortcoming, or at least potential one, was mentioned by Gödel in his letter to Zermelo: unlike the published proof, the Master Argument is "not intuitionistically unobjectionable".

Of course, non-constructivity is already objectionable to intuitionists, but the Master has a further feature that they do not accept: its use of the notions of *truth and falsehood* in mathematics. Intuitionists baulk at the idea that mathematical propositions have an objective truth-value beyond our ability to give intuitively satisfying demonstrations or refutations of them. For this reason, they do not accept in their full generality certain principles of classical propositional (and first-order) logic, most conspicuously the law of excluded middle, double negation elimination and one half of contraposition. But the very notion of expressibility, which features essentially in both lemmas for the Master Argument, is defined in terms of truth and falsehood in an intended model, and the law of excluded middle is implicit in e.g. the last sentence of the proof of the Inexpressibility Lemma 4.2.

In the early 1930s, intuitionism was a live option as a philosophy of mathematics, and its perspectives influenced the way in which Gödel presented his official proof of the incompleteness theorems. This is clear from a famous "crossed-out passage in an unsent reply" (Feferman's memorable phrase) written on 27 May 1970 to graduate student Yossef Balas. There Gödel said: "However in consequence of the philosophical prejudices of our time ... a concept of objective mathematical truth as opposed to demonstrability was viewed with greatest suspicion and widely rejected as meaningless." As Feferman observes in his paper of 1984, it is clear that when establishing his incompleteness result Gödel did not himself share that suspicion. But he nevertheless refrained from using the notions of mathematical truth and intended model out of an abundance of caution or, to put it more plainly, from fear of adverse reception by the mathematical community of the time.

Today intuitionistic logic is more an object of study than a code to live by. Few logicians and fewer mathematicians have any qualms about using the law of excluded middle or double negation elimination. Should we still retain any suspicions about the notion of *truth in the intended model* of a first-order theory? This is a philosophical question, and it would be rash to think that only one answer is possible. But many feel that there is no intrinsic difficulty with this concept. On the one hand, we can define the domain of the intended model, and the values to be given to the primitive operations of successor, addition and multiplication, within the confines of a quite small fragment of set theory; on the other hand we can define the truth-values of complex formulae, in that model, by recursion in the manner that was articulated by Tarski and is now standard.

On this view, there are really only two shortcomings to the Master Argument: a ceiling on what it shows and its non-constructivity. The canvassed philosophical weakness of relying on the notion of 'truth in the intended model of *PA*', is not a ground for serious concern.

Nevertheless, it is interesting to see that the Master Argument may be re-run

on a purely syntactic plane without any reference to truth in the intended model. Thus, even if one has residual worries about that notion, they become irrelevant. The re-run has, moreover, a technical benefit: it can be done in such a way as to yield Gödel's syntactic version of the first incompleteness theorem (ω-consistency implies negation-incompleteness), thus raising somewhat the ceiling on content although remaining non-constructive. While a little more abstract than the basic version of the Master Argument, it is no more complex. We turn to it now.

8 The Master Argument without truth

We begin by generalizing the definition of expressibility. The definition in Section 4 took a set $S \subseteq N$ to be expressible in the language of *PA* iff there is a formula $\varphi(x)$ with one free variable x, such that for all $n \in N$, $n \in S$ iff $\varphi(\overline{n}) \in T$ where, we recall, T stands for the set of sentences that are true in the intended model for *PA*. Evidently, this definition continues to make sense if we replace T by another set of sentences of *LPA*; we can indeed generalize from T to an arbitrary set X of sentences, as follows.

Definition 8.1. *Let X be any set of sentences in the language of PA; we call it an 'oracle'. We say that that a set $S \subseteq N$ is* expressible according to (the oracle) *X iff there is a formula $\varphi(x)$ with one free variable x in LPA such that for all $n \in N$,*

$$n \in S \text{ iff } \varphi(\overline{n}) \in X.$$

In particular, S is expressible according to the oracle T iff it is expressible *tout court*. The generalized definition 8.1 allows us to formulate a syntactic version of the Master Argument, using lemmas that follow the originals but with certain small changes.

As before, fix an enumeration of all formulae in the language of *PA* with just one free variable x. Generalize the definitions of D^- and D^+ thus: for any set X of sentences in the language of *PA*, put D^-_X (resp. D^+_X) to be the set of all natural numbers n such that $\varphi_n(\overline{n}) \notin X$ (resp. $\in X$). Clearly these two sets are complements of each other wrt. N, and as particular cases we have $D^-_T = D^-$ and $D^+_T = D^+$.

The Inexpressibility Lemma modulo an oracle 8.2 for D^-_X is formulated just as for D^-, but for D^+_X we need the hypothesis that X is well-behaved wrt. negation, in the sense that for every sentence φ in the language of *PA*, exactly one of $\varphi, \neg\varphi$ is in X. To appreciate the force of that hypothesis, note that one half of it (at least one of $\varphi, \neg\varphi$ is in X) is just negation-completeness, while the other half (at least one

of $\varphi, \neg\varphi$ is not in X) is immediately implied by consistency. Indeed, if one assumes that X is closed under classical consequence, then the second half is equivalent to consistency. We have no need to make that assumption, but doing so would cause no harm to the argument.

Lemma 8.2 (Inexpressibility Lemma (modulo an oracle))**.** *Let X be any set of sentences in the language of PA. Then D^-_X is not expressible according to the oracle X. Moreover, if X is well-behaved wrt. negation, then D^+_X is not expressible according to X.*

Proof. For D^-_X, we argue exactly as before with X in place of T. Suppose for reductio that it is expressible according to X. Then by the definition of expressibility according to X, there is a formula $\varphi(x)$ with x as the sole free variable such that for all $n \in N$, $n \in D^-_X$ iff $\varphi(\overline{n}) \in X$. Now $\varphi(x) = \varphi_k(x)$ for some $k \in N$. So, instantiating n to k we have in particular $k \in D^-_X$ iff $\varphi_k(\overline{k}) \in X$. But by the definition of D^-_X we have $k \in D^-_X$ iff $\varphi_k(\overline{k}) \notin X$, giving a contradiction.

For D^+_X, suppose for *reductio* that it is expressible according to X by a formula $\varphi(x)$ and that X is well-behaved wrt. negation. Then for all $n \in N$, $n \in D^+_X$ iff $\varphi(\overline{n}) \in X$ iff $\neg\varphi(\overline{n}) \notin X$. But then $n \in D^-_X$ iff $n \notin D^+_X$ iff $\neg\varphi(\overline{n}) \in X$, so that D^-_X is expressed by the formula $\neg\varphi(x)$ according to X, contrary to what we have just shown. □

The Expressibility Lemma modulo an oracle 8.3 runs parallel to its unmodulated counterpart, with $D^+_{|PA|}$ replacing $D^+ = D^+_{|T|}$ and expressibility according to the oracle $|PA|$ replacing expressibility *tout court*. This forces two rejigs. Since truth is no longer involved, the lemma requires the condition of ω-consistency rather than soundness in the intended model; since expressibility of a set $S \subseteq N$ according to the oracle $|PA|$ does not in general follow immediately from the same for its complement $N \backslash S$, the lemma covers only $D^+_{|PA|}$ and not $D^-_{|PA|}$.

As before, we fix separate acceptable numberings of all formulae with just one free variable x and of all derivations of PA.

Lemma 8.3 (Expressibility Lemma (modulo an oracle))**.** *If PA is ω-consistent, then $D^+_{|PA|}$ is expressible according to the oracle $|PA|$.*

Proof. As in the proof of the Expressibility Lemma 4.3, consider the relation that holds between a derivation δ_m and a formula $\varphi_n(x)$ with just one free variable x iff the former is a derivation of $\varphi_n(\overline{n})$, and recall that this relation is captured in PA by some formulae $\psi(y, x)$ in the sense that for all $m, n \in N$,

1. if δ_m stands in the relation to $\varphi_n(x)$ then $\psi(\overline{m}, \overline{n}) \in |PA|$ and

2. if δ_m does not stand in the relation to $\varphi_n(x)$ then $\neg\psi(\overline{m},\overline{n}) \in |PA|$.

Now suppose that PA is ω-consistent. We want to show that the formula $\exists y\psi(y,x)$ expresses $D^+_{|PA|}$ according to the oracle $|PA|$; that is: for all $n \in N$, $n \in D^+_{|PA|}$ iff $\exists y\psi(y,\overline{n}) \in |PA|$.

Left to right: Suppose $n \in D^+_{|PA|}$. Then by definition, $\varphi_n(\overline{n}) \in |PA|$. Hence there is a derivation δ_m of $\varphi_n(\overline{n})$, so δ_m stands in the relation to $\varphi_n(x)$ so, by (1), $\psi(\overline{m},\overline{n}) \in |PA|$, so by classical logic $\exists y\psi(y,\overline{n}) \in |PA|$ as desired.

Right to left: Suppose $n \notin D^+_{|PA|}$. Then by definition, $\varphi_n(\overline{n}) \notin |PA|$. Hence there is no derivation δ_m of $\varphi_n(\overline{n})$, so no δ_m bears the relation to $\varphi_n(x)$ so, by (2), $\neg\psi(\overline{m},\overline{n}) \in |PA|$ for all $m \in N$. Thus by the ω-consistency of PA, $\exists y\psi(y,\overline{n}) \notin |PA|$ as desired. □

Those who relish fine detail may compare the verifications contained in the last two paragraphs with their counterparts in the Expressibility Lemma 4.3. Both directions have indeed become a little simpler as a result of dealing with the oracle $|PA|$ rather than the truth-set T: in the left to right direction, we could simply omit the last sentence of the previous version; in the converse direction, the last sentence brings ω-consistency into play in lieu of soundness modulo PA.

The syntactic version of Gödel's first incompleteness theorem appears immediately from the collision between the two oracular lemmas.

Theorem 8.4 (Gödel's First Incompleteness Theorem (syntactic version))**.** *If PA is ω-consistent then it is negation-incomplete.*

Proof. Suppose for *reductio* that PA is ω-consistent and negation-complete. Using ω-consistency, the Expressibility Lemma modulo an oracle 8.3 tells us that $D^+_{|PA|}$ is expressible according to $|PA|$. But ω-consistency immediately implies consistency so, combining that with negation-completeness, $|PA|$ is well-behaved wrt. negation. So the second part of the Inexpressibility Lemma modulo an oracle 8.2 tells us that $D^+_{|PA|}$ is not expressible according to $|PA|$, giving a contradiction. □

It is also possible to put Tarski's theorem on the undefinability of truth into a form that is no longer about truth in the intended model, but about an arbitrary oracle satisfying certain syntactic conditions. However, the details of both formulation and proof are a little more complex than we wish to handle here. We have gone only so far as is needed to render the Master Argument immune to the criticism that it uses the general notions of truth and falsehood of sentences of PA, and to raise the ceiling on its content to cover the syntactic version of Gödel's first incompleteness theorem. Readers who would like to see 'oracular' versions of Tarski's Theorem are directed to the texts of Smullyan and Fitting in the list of resources that follows.

References

[1] Hans-Dieter Ebbinghaus. *Ernst Zermelo: An Approach to his Life and Work*. Berlin: Springer, 2010. The correspondence with Gödel is discussed in section 4.10.

[2] Solomon Feferman. Kurt Gödel: conviction and caution. *Philosophia Naturalis*, 21:546–562, 1984. This influential paper was republished with minor additions as chapter 7 of the same author's book *In the Light of Logic*, Oxford: Oxford University Press 1998. Its theme is Gödel's outer caution about using the notion of truth in mathematics, as contrasted with his inner confidence in its meaningfulness.

[3] Melvin Fitting. *Incompleteness in the Land of Sets*. London: College Publications, 2007. An abstract form of the Master Argument is given in section 8.5. Also contains an abstract form of Tarski's Theorem.

[4] Kurt Gödel. Uber formal unentscheidbare Sätze der Principia Mathematica und Verwandter Systeme I. *Monatshefte für Mathematik und Physics*, 38:173–198, 1931. The 'official' version in its German original. A number of English translations are available, notably in Volume 1 of Gödel's collected works.

[5] Kurt Gödel. *Collected Works*, volume 1-5. Oxford: Clarendon Press, 1986-2003. The definitive, dual-language, collection. The letters are in volumes 5 and 6, ordered alphabetically by correspondent. Readers are urged to examine for themselves Zermelo's letter, Gödel's reply, and the response of Zermelo that ended the correspondence.

[6] Ivor Grattan-Guinness. In memoriam Kurt Gödel: his 1931 correspondence with Zermelo on his incompletability theorem. *Historia Mathematica*, 6:294–304, 1979. The first publication (in German) of Gödel's reply to Zermelo, dated 12 October 1931, and the latter's response of 29 October. The initial letter of Zermelo, of 21 September, was found later by John Dawson in Gödel's Nachlass and first published (in German) in his note 'Completing the Gödel-Zermelo correspondence' *Historia Mathematica* 12 (1985): 66-70.

[7] Roman Murawski. Undefinability of truth. the problem of priority: Tarski vs Gödel. *History and Philosophy of Logic*, 19:153–160, 1998. Discusses the historical relationship between Gödel's work on incompleteness and Tarski's work on the undefinability of truth.

[8] Peter Smith. *An Introduction to Gödel's Theorems*. Cambridge: Cambridge University Press, 2nd edition, 2013. The Master Argument is sketched in section 27.5, but without abstraction to a version without truth.

[9] Peter Smith. *Gödel without (too many) tears*. version of 20 February 2015. The text is based on Smith 2013, but is more selective, concise and lively. It is perhaps the most readable among accounts that go deeply into the machinery of Gödel's incompleteness theorems. The Master Argument is in a box in section 50.

[10] Raymond Smullyan. *Diagonalization and Self-Reference*. New York: Oxford University Press, 1994. Seeks maximum generality; not to be tackled lightly. Abstract versions of the Master Argument appear several times, beginning with Theorem 2.2. Also contains an abstract version of Tarski's Theorem.

[11] Alfred Tarski. *Pojęcie prawdy w językach nauk dedukcynjnych*. Nakladem Towarzysta Naukowego Warszawskiego: Warsaw. An English translation, 'The concept of truth in formalized languages', may be found in pp. 152-278 of the same author's collection *Logic, Semantics Metamathematics*, Oxford: Clarendon Press.

Received March 2015