HELDA
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Gödelian sentences and semantic arguments

Sandu, Gabriel

2020-08-06

Sandu , G 2020 , ' Gödelian sentences and semantic arguments ' , Logical Investigations , vol. 26 , no. 1 , pp. 60-77 . https://doi.org/10.21146/2074-1472-2020-26-1-60-77

http://hdl.handle.net/10138/340645
https://doi.org/10.21146/2074-1472-2020-26-1-60-77

cc_by_nc
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Логические исследования
2020. Т. 26. № 1. С. 60–77
УДК 16

Logical Investigations
2020, Vol. 26, No. 1, pp. 60–77
DOI: 10.21146/2074-1472-2020-26-1-60-77

Gabriel Sandu

Gödelian sentences and semantic arguments

Gabriel Sandu
University of Helsinki,
P.O. Box 24 (Unioninkatu 40 A), 00014 Helsinki, Finland.
E-mail: gabriel.sandu@helsinki.fi

Abstract: This paper contains some philosophical reflections on Gödelian (undecidable) sentences and the recognition of their truth using semantic arguments. These reflections are not new, similar matters have been extensively addressed in the philosophical literature. The matter is rather one of emphasis.

Keywords: Gödelian sentences, Gödel's incompleteness theorem, semantical argument, truth theory, arithmetic, proof, provability

For citation: Sandu G. "Gödelian sentences and semantic arguments", *Logicheskie Issledovaniya / Logical Investigations*, 2020, Vol. 26, No. 1, pp. 60–77. DOI: 10.21146/2074-1472-2020-26-1-60-77

To the memory of Alexandr Karpenko, such a great friend

1. Gödel incompleteness theorem

Let L be the *language of arithmetic*, consisting of

- variables, x_0,$x_1, x, y, ...$
- logical constants: $\neg, \vee, \exists x, =$
- nonlogical constants: $\mathbf{0}, \mathbf{S}, +, \times$.

(Here '$\mathbf{0}$' is an individual constant, '$\mathbf{S}$', is a one-place function symbol and '$+$', and '$\times$' are two place function symbols.)

From these items, the terms and formulas of the language of L are formed in the standard way.

As Tarski observed, the object language of a formalized science, comes together with a theory, usually given by listing its axioms and rules of inference. In our case the starting point is the theory Q (minimal arithmetic) which is the set of logical consequences of the following axioms:

© Sandu G.

1. $\forall x \forall y(\boldsymbol{S}x = \boldsymbol{S}y \rightarrow x = y)$

2. $\forall x(\boldsymbol{S}x \neq \mathbf{0})$

3. $\forall x(x \neq \mathbf{0} \rightarrow \exists y(s = \boldsymbol{S}y))$

4. $\forall x(x + \mathbf{0} = x)$

5. $\forall x \forall y(x + \boldsymbol{S}y = \boldsymbol{S}(x + y))$

6. $\forall x(x \times \mathbf{0} = \mathbf{0})$

7. $\forall x \forall y(x \times \boldsymbol{S}y = (x \times y) + x)$.

Notice that this theory is finitely axiomatizable. The language of Q is interpreted in a metalanguage in which '**0**' is assigned the the natural number zero, '**S**' is assigned the successor function, '+' is assigned the operation of addition '×' is assigned multiplication. It is known that Q is a rather strong theory which is able to represent all recursive functions (in a technical sense of the notion of 'representation', which is assumed to be known. It is also known that Q defines (in a technical sense assumed to be known) its own syntax and many semantical notions. This happens, as shown by Gödel, via the notion of gödel numbering. As a result, each term t in the language L gets associated with a gödel number $\ulcorner t \urcorner$; and each formula A receives its gödel number $\ulcorner A \urcorner$. Recalling that every natural number m has a name $\underline{m}$ in L, where $\underline{m}$ is an abbreviation for (the numeral) $\underbrace{\boldsymbol{S}\boldsymbol{S}...}\mathbf{0}$ (m times), we see that every term t and every formula A have names in the arithmetical language, $\underline{\ulcorner t \urcorner}$ and $\underline{\ulcorner A \urcorner}$, respectively. This fact, together with the ones mentioned earlier, makes possible to introduce, for any formula A in the language of arithmetic, the diagonalization of A, which is the expression

$$\exists x(x = \underline{\ulcorner A \urcorner} \wedge A).$$

When A is a formula with one free variable, then we see that asserting the diagonalization of A amounts to predicating A of its own gödel number.

From Gödel's results, it follows that for any theory T extending Q, the set of gödel numbers of theorems of T is not definable in T, from which it can be further inferred that the set of Gödel numbers of true arithmetical sentences ("true in the standard model") is not definable. This last statement is usually known as "Tarski's theorem"; it is somehow debatable in the literature whether Gödel himself was aware of this result or not, but this matter will not concern us here. The first statement is standardly proved by reductio using the

diagonalisation lemma which asserts that for any theory T which extends Q, for any formula $B(y)$ there is a sentence A such that

$$T \vdash A \leftrightarrow \neg B(\underline{\ulcorner A \urcorner}).$$

The second statement follows directly from it, by observing that the set of true arithmetical sentences is an extension of Q.

The variant of the Gödel's incompleteness theorem we are interested in is proved by first showing that for every extension T of Q there is a formula $Pr_T(x)$ in the language of arithmetic which has the form $\exists y Prov_T(x, y)$ and is such that for any sentence A in the language of arithmetic:

- $T \vdash A$ if and only if $\exists y Prov_T(\underline{\ulcorner A \urcorner}, y)$ is true (in the standard model) if and only if for some natural number m, $Prov_T(\underline{\ulcorner A \urcorner}, \underline{m})$ is true if and only if (given the representability of $Prov_T$ in Q), $Q \vdash Prov_T(\underline{\ulcorner A \urcorner}, \underline{m})$ for some m.

Here $Prov_T(x, y)$ is a primitive recursive formula, that is, a formula which contains only bounded quantifiers and is closed under the standard propositional connectives. Thus, from the above we get that if $T \vdash A$ then $Q \vdash Prov_T(\underline{\ulcorner A \urcorner}, \underline{m})$ for some m, and given that T is an extension of Q we also get $T \vdash \exists y Prov_T(\underline{\ulcorner A \urcorner}, y)$, i.e., $T \vdash Pr(\underline{\ulcorner A \urcorner})$. Now applying the Diagonalization lemma to the formula $\exists y Prov_T(\underline{\ulcorner A \urcorner}, y)$ Godel showed that there is a sentence, usually denoted by G such that

$$T \vdash G \leftrightarrow \neg \exists y Prov_T(\underline{\ulcorner G \urcorner}, y)$$

The sentence G is called a *Gödel sentence for* T. It is taken to say: "I am unprovable".

We recall that a theory T is called $\omega-$inconsistent if there is a formula $F(x)$ such that $T \vdash \exists x F(x)$ but $T \vdash \neg F(\underline{0})$, $T \vdash \neg F(\underline{1})$, $T \vdash \neg F(\underline{2})$,...(for every natural number $0, 1, 2, ...$). T is called $\omega-$consistent if it is not $\omega-$inconsistent. Now Gödel proved

Theorem 1. *(Gödel First Incompleteness Theorem). Let T be a consistent, axiomatizable extension of Q and let G be a Gödel sentence for T. Then $T \nvdash G$. If T is $\omega-$consistent, then $T \nvdash \neg G$.*

The proof is well known but we rehearse it here (we follow Boolos, Jeffrey and Burgess), because it serves as a basis for extracting, later on, a semantic argument. Suppose that $T \vdash G$. Hence, by our previous comments, $\exists y Prov_T(\underline{\ulcorner G \urcorner}, y)$ is true (in the standard model) and by a well known result, $Q \vdash \exists y Prov_T(\underline{\ulcorner G \urcorner}, y)$; given that T is an extension of Q we also have

$T \vdash \exists y Prov_T(\underline{\ulcorner G \urcorner}, y)$. From the Diagonalization lemma we also know that $T \vdash \neg\exists y Prov_T(\underline{\ulcorner G \urcorner}, y)$. Thus T is inconsistent, a contradition. Hence $T \nvdash G$. For the second claim, suppose that $T \vdash \neg G$. By the diagonalization lemma, $T \vdash \exists y Prov_T(\underline{\ulcorner G \urcorner}, y)$. But given that T is consistent and $T \vdash \neg G$, we must have $T \nvdash G$. This implies that for no natural number n, n is the code of a proof of G in T, that is, $\neg Prov_T(\underline{\ulcorner G \urcorner}, \underline{0})$, $\neg Prov_T(\underline{\ulcorner G \urcorner}, \underline{1})$, $\neg Prov_T(\underline{\ulcorner G \urcorner}, \underline{2})$..., are all true (in the standard model), where each of these formulas are primitive recursive. Hence $Q \vdash \neg Prov(\underline{\ulcorner G \urcorner}, \underline{0})$, $Q \vdash \neg Prov(\underline{\ulcorner G \urcorner}, \underline{1})$, $Q \vdash \neg Prov(\underline{\ulcorner G \urcorner}, \underline{2})$.... and since T is an extension of Q we also have $T \vdash \neg Prov(\underline{\ulcorner G \urcorner}, \underline{0})$, $T \vdash \neg Prov(\underline{\ulcorner G \urcorner}, \underline{1})$, $T \vdash \neg Prov(\underline{\ulcorner G \urcorner}, \underline{2})$.... Hence T is $\omega-$inconsistent, which contradicts our assumption. We conclude $T \nvdash \neg G$.

After reviewing these results, let us return to the question which is the main concern in this paper, namely Gödel's method to produce undecidable sentences such as G, and especially a claim often made in this connection to the effect that these sentences are true and *recognized to be true*. Here is, for instance, how Dummett describes Gödel's result:

> By Gödel's theorem there exists, for an intuitively correct formal system for elementary arithmetic, a statement $[G]$ expressible in the system but not provable in it, which not only is true but can be recognized by us to be true... [Dummett, 1963].

The puzzling question is: how do we "recognize" that G (or any statement equivalent to it) is true?

The above proof of the theorem does not give an explicit argument about how we come to recognize G as true, neither did Gödel provide one. But it is not very difficult to extract one. From the Diagonalization lemma we know that the statement G is equivalent to a universal statement, viz. $\neg\exists y Prov_T(\underline{\ulcorner G \urcorner}, y)$ (i..e $\forall y \neg Prov_T(\underline{\ulcorner G \urcorner}, y)$). From the second part of the proof we see that every numerical instance is provable (and true) in the system. Since G is the universal quantification over all these numerical instances, then G is true. Of course in this last step we rely on our grasp of the standard model (this is what the $\omega-$consistency is supposed to ensure).

In fact, this is Dummett's argument for the truth of Gödel's sentence:

> The statement $[G]$ is of the form $\forall x A(x)$,where each one of the statements $A(\underline{0})$,$A(\underline{1})$,$A(\underline{2})$, ...is true: since $A(x)$ is recursive, the notion of truth for these statements is unproblematic. Since each of the statements $A(\underline{0})$,$A(\underline{1})$,$A(\underline{2})$, ...is true in every model of the formal system, every model of the system in which G is false must be a non-standard model...whenever, for some predicate $B(x)$, we

> can recognize all of the statements $B(\underline{0})$,$B(\underline{1})$,B$A(\underline{2})$, ...as true in the standard model, then we can recognize that $\forall x A(x)$ is true in that model. This fact ...we know on the strength of our clear intuitive conception of the structure of the model [Dummett, 1963, p. 191].

As we see from this quote, we come to appreciate that the undecidable Gödel sentence G for Q is true not by working inside the system but rather by conducting a so called *semantical argument* which makes an essential use of the concept of truth itself. Dummett is not the only one to have seen the importance of semantical arguments. There is another semantical argument which uses the truth predicate, distinct from Dummett's argument, which goes back to Alfred Tarski [Tarski, 1956]. In order to present it, we need to say somehting about arithmetical induction.

The system Q of minimal arithmetic is knowingly defficient in that it fails to prove many universal statements about numbers which are usually proved by mathematical induction. Typically, if we want to prove that every number has a given property, we prove it by showing that 0 has that property, and then we show, from the assumption that an arbitrary number x has that property, that the successor Sx has that property. To accommodate induction one needs a more adequate set of axioms for number theory. To this effect we add to the 7 axioms of the system Q all sentences of the form

8. $[A(\mathbf{0}) \wedge \forall x(A(x) \rightarrow A(S(x))] \rightarrow \forall x A(x)$

(8) is usually known as the *Induction axiom scheme*. The theory which is the set of all sentences in the language of arithmetic which are logical consequences of (1)–(8) is known as Peano Arithmetic (PA). It is a simple mathematical fact that definability and representability in Q entail definability and representability in any extension of Q and thus in PA in particular. From now on we shall operate with PA. Tarski's semantical argument which proves the truth of the Gödelian statement G for PA, uses a universal statement which cannot be proved in Q but needs PA.

1.1. The representability of the syntax in arithmetic

Tarski's truth-definition for arithmetic exploits the representability of the syntax of PA in PA.

It is a mathematical fact that there are functions $f_{\neg}, f_{\vee}, f_{\exists}$ defined on the natural numbers such that the following hold:

- $f_{\neg}(\ulcorner A \urcorner) = \ulcorner \neg A \urcorner$, for every formula A in the object language;
- $f_{\vee}(\ulcorner A \urcorner, \ulcorner B \urcorner) = \ulcorner A \vee B \urcorner$, for every formulas A, B in the object language;
- $f_{\exists}(\ulcorner A \urcorner, n) = \ulcorner \exists x_n A \urcorner$, for every formula A and natural number n.

There is also a function f_{sub} (the susbstitution function) which has the property:

$$f_{sub}(\ulcorner A\urcorner, \ulcorner x_i\urcorner, \ulcorner t\urcorner) = \ulcorner A(t)\urcorner$$

for every formula A in the language of arithmetic, variable x_i and term t in the same language.

All these functions are recursive, thus representable in Q and hence in PA which means there are formulas $Neg(x,y), Dis(x,y,z), Ex(x,y,z)$ and $Sub(x,y,z,w)$ in the language of arithmetics so that for all formulas A, B, term t, and natural number n we have

a) $PA \vdash \forall y\,(Neg(\underline{\ulcorner A\urcorner}, y) \leftrightarrow y = \underline{\ulcorner \neg A\urcorner})$

b) $PA \vdash \forall y\,(Dis(\underline{\ulcorner A\urcorner}, \underline{\ulcorner B\urcorner}, y) \leftrightarrow y = \underline{\ulcorner A \vee B\urcorner})$

c) $PA \vdash \forall y\,\left(Ex(\underline{\ulcorner A\urcorner}, \underline{n}, y) \leftrightarrow y = \underline{\ulcorner \exists x_n A\urcorner}\right)$

d) $PA \vdash \forall y\,\left(Sub(\underline{\ulcorner A\urcorner}, \underline{\ulcorner x_i\urcorner}, \underline{\ulcorner t\urcorner}, y) \leftrightarrow y = \underline{\ulcorner A(t)\urcorner}\right)$

Similarly, the function $f_=$ on the natural numbers such that

$$f_=(\ulcorner t\urcorner, \ulcorner s\urcorner) = \ulcorner t = s\urcorner$$

for all terms t, s in the language of arithmetic is representable in PA by, say, the expression $Id(x,g,z)$, that is,

$$PA \vdash \forall y\,(Id(\ulcorner t\urcorner, \ulcorner s\urcorner, y) \leftrightarrow y = \ulcorner t = s\urcorner)\,.$$

If in (a) we instantiate y with $\underline{\ulcorner \neg A\urcorner}$ we get

$$PA \vdash Neg(\underline{\ulcorner A\urcorner}, \underline{\ulcorner \neg A\urcorner}) \leftrightarrow \underline{\ulcorner \neg A\urcorner} = \underline{\ulcorner \neg A\urcorner}.$$

The formula on the right side is a theorem of the predicate calculus (with identity), hence PA proves it. Thus $PA \vdash Neg(\underline{\ulcorner A\urcorner}, \underline{\ulcorner \neg A\urcorner})$. We can show that for each formula A of the object language there is exactly one formula B of the object laguage such that $PA \vdash Neg(\underline{\ulcorner A\urcorner}, \underline{\ulcorner B\urcorner})$ and B is $\neg A$. Therefore we can take Neg to be a function and write $Neg(\underline{\ulcorner A\urcorner}) = \underline{\ulcorner \neg A\urcorner}$.

In a similar way we can also take $Dis, Ex, Sub, Id, Less$ to be also functions. Thus we shall have

a*) $PA \vdash Neg(\underline{\ulcorner A\urcorner}) = \underline{\ulcorner \neg A\urcorner}$, for every formula A in the object language.

b*) $PA \vdash Dis(\underline{\ulcorner A\urcorner}, \underline{\ulcorner B\urcorner}) = \underline{\ulcorner A \vee B\urcorner}$, for every formulas A, B in the object language

c*) $PA \vdash Ex(\underline{\ulcorner A \urcorner}, \underline{n}) = \underline{\ulcorner \exists x_n A \urcorner}$, for every formula A in the object language and natural number n.

d*) $PA \vdash Sub(\underline{\ulcorner A \urcorner}, \underline{\ulcorner x_i \urcorner}, \underline{\ulcorner t \urcorner}) = \underline{\ulcorner A(t) \urcorner}$, for every formula A and term t of the object language and every natural number i.

e*) $PA \vdash Id(\ulcorner t \urcorner, \ulcorner s \urcorner) = \ulcorner t = s \urcorner$, for all terms t, s of the object language.

In a similar way it can be shown that PA defines its own syntax: being a closed term, a variable, a formula and a sentence (of the language of arithmetic). That is, there are formulas $ct(x), var(x), form(x)$ and $sen(x)$ in the object language such that the following holds:

f) $PA \vdash ct(\underline{\ulcorner t \urcorner})$, for every closed term t.

g) $PA \vdash var(\underline{\ulcorner x_i \urcorner})$, for every natural number i.

h) $PA \vdash form(\underline{\ulcorner A \urcorner})$, for every formula A.

j) $PA \vdash sen(\underline{\ulcorner A \urcorner})$, for every closed sentence A.

PA also defines some semantical properties. There is a formula $Den(x)$ in the object language (that we can take to be a function) such that

k) $PA \vdash t = s \leftrightarrow Den(\underline{\ulcorner t \urcorner}) = Den(\underline{\ulcorner s \urcorner})$, for all terms t, s in the object language.

2. Tarski's truth theory

In the case of Tarski's truth theory for arithmetic we do not need to go via the notion of satisfaction but use directly the truth-predicate Tr. The reason for this is that each natural number has a name in the object language.

The axioms of the truth-definition are given in the metalanguage containing Tr is a predicate symbol:

Ax1 $\forall x(Tr(x) \rightarrow sen(x))$

(If x is true, then x is of a sentence)

Ax2 $\forall x \forall y(ct(x) \wedge ct(y) \rightarrow (Tr(Id(x, y)) \leftrightarrow Den(x) = Den(y)))$

(The identity between two closed terms x and y is true iff their denotations are the same)

Ax3 $\forall x(Sen(x) \rightarrow (Tr(Neg(x)) \leftrightarrow \neg Tr(x)))$

(The negation of the sentence is true iff the sentence is not true)

Ax4 $\forall x \forall y(sen(x) \wedge sen(y) \rightarrow (Tr(Dis(x,y)) \leftrightarrow Tr(x) \vee Tr(y)))$

(A disjunction is true iff either sentence is true)

Ax5 $\forall x_1 \forall x_2(form(x_1) \wedge var(x_2) \rightarrow (Tr(Ex(x_1,x_2)) \leftrightarrow \exists t(Tr(Sub(x_1,x_2,t))))$

(An existential sentence is true iff there is a closed term t such that the sentence which is the result of the substitution of the free variable x_2 in x_1 by t is true.)

Let $PA(Tr)$ be the set of sentences which are the logical consequences of the 7 axioms of PA, the five axioms (Ax1)–(Ax5), and plus the Induction schema (8) which allows occurrences of the truth-predicate in the formulas $A(x)$. It can be shown that $PA(Tr)$ is materially adequate, that is,

$$PA(Tr) \vdash Tr(\underline{\ulcorner A \urcorner}) \leftrightarrow A,$$

for any sentence A in the language of arithmetic.

It is well known that the Tarskian truth theory proves the following universal statements:

- *The principle of noncontradiction (consistency).* For every sentence y of the object language it is not the case that both y and its negation are true:
$$PA(Tr) \vdash \forall y\,(Sen(y) \rightarrow \neg(Tr(y) \wedge Tr(neg(y))))\,.$$
This property follows directly from Ax3.

- *The principle of excludded middle.* Every sentence of the object language is true ot its negation is true:
$$PA(Tr) \vdash \forall y\,(Sen(y) \rightarrow Tr(y) \vee Tr(neg(y)))\,.$$
This property follows from the other direction of Ax3.

- *The principle of soundness.* All theorems are true:
$$PA(Tr) \vdash \forall x(Pr_{PA}(x) \rightarrow Tr(x)).$$

This principle fully exploits the occurrence of the truth-predicate in the Induction scheme. We omit its proof but it consists, informally, of the following steps:

1. All the axioms of PA are true.
2. The rules of inference of PA preserve truth.
3. Hence every theorem of PA is true (i.e. $PA(Tr) \vdash \forall x(Pr_{PA}(x) \rightarrow Tr(x))$.

2.1. Tarski's semantical argument

In the postscript to the English translation of his seminal article, Tarski adds some interesting parallels between his results and those of Gödel:

> Moreover Gödel has given a method for constructing sentences which- assuming the theory concerned to be consistent- cannot be decided in any direct way in this theory. All sentences constructed according to Gödel's method possess the property it can be established whether they are true or false on the basis of the metatheory of higher order having a correct definition of truth [Tarski, 1956, p. 274].

To establish the truth of such a Gödelian sentence Tarksi uses the principle of soundness listed in the previous sesction. We present Tarski's semantical argument (Tarski, 1936, Theorem 5) for the Gödelian sentence $\neg Pr_{PA}(\underline{\ulcorner \neg \mathbf{0} = \mathbf{0} \urcorner})$ (that we shall abbreviate by Con_{PA}) which is taken to express the consistency of PA. The semantical argument for G is similar. There is nothing original in my presentation, this argument has been rehearsed many times [Ketland, 1999] and [Shapiro, 1998].

Gödel's second incompleteness theorem shows that $PA \nvdash Con_{PA}$ and $PA \nvdash \neg Con_{PA}$. But Tarski shows

$$PA(Tr) \vdash Con_{PA}.$$

The argument is straightforward. From the soundness principle we get

$$(i) \quad PA(Tr) \vdash Pr_{PA}(\underline{\ulcorner \neg \mathbf{0} = \mathbf{0} \urcorner}) \rightarrow Tr(\underline{\ulcorner \neg \mathbf{0} = \mathbf{0} \urcorner}).$$

We also know that the theory of truth proves all the T-instances, i.e.,

$$(ii) \quad PA(Tr) \vdash Tr(\underline{\ulcorner \neg \mathbf{0} = \mathbf{0} \urcorner}) \leftrightarrow \neg \mathbf{0} = \mathbf{0}.$$

But PA proves $\mathbf{0} = \mathbf{0}$, and thus $PA(Tr) \vdash \mathbf{0} = \mathbf{0}$, which together with (ii) entails

$$(iii) \quad PA(Tr) \vdash \neg Tr(\underline{\ulcorner \neg \mathbf{0} = \mathbf{0} \urcorner}).$$

From (i) and (iii) we get

$$(iv) \quad PA(Tr) \vdash \neg Pr_{PA}(\underline{\ulcorner \neg \mathbf{0} = \mathbf{0} \urcorner})$$

that is, $PA(Tr) \vdash Con_{PA}$.

Tarski's *semantical argument* is usually expressed in words, in order to enhance its explanatory power:

- In a first step we establish the principle of soundness as we showed earlier:

1. All the axioms of PA are true.
2. The rules of inference of PA preserve truth.
3. Hence every theorem of PA is true,

$$PA(Tr) \vdash \forall x(Pr_{PA}(x) \rightarrow Tr(x)).$$

- A second step established that the sentence '$\neg\mathbf{0} = \mathbf{0}$' is not true:

$$PA(Tr) \vdash \neg Tr(\underline{\ulcorner \neg\mathbf{0} = \mathbf{0} \urcorner})$$

(see (iii))

- In a third step we combined the conclusion of the first and of the second step and concluded that '$\neg\mathbf{0} = \mathbf{0}$' is not a theorem:

$$PA(Tr) \vdash \neg Pr_{PA}(\underline{\ulcorner \neg\mathbf{0} = \mathbf{0} \urcorner})$$

(see (iv))

- Finally we note that $\neg Pr_{PA}(\underline{\ulcorner \neg\mathbf{0} = \mathbf{0} \urcorner})$ is the Consistency statement Con_{PA}.

The crucial role in this argument is the universal generalization which is the Principle of soundness. It confers the semantic argument the form of a nomological argument which shows the *explanatory role of the truth predicate:*

> Let us return to the Gödelian statement G (or Con_{PA}). Let us suppose a logic teacher asserts that Con_{PA} is true, and the puzzled student asks for an explanation. The student believes the teacher's word that Con_{PA} is true, but he wants to be shown why Con_{PA} is true. The student wants something like a convincing proof or an explanatory proof. The natural answer is to remark that all the axioms of PA are true and the rules of inference preserve truth. Thus every theorem of PA is true. It follows that '$\neg 0 = 0$' is not a theorem and thus PA is consistent.... It seems to me that this informal version of the derivability of Con_{PA} is as good an *explanation* as there is. The argument shows why Con_{PA} is true or why Con_{PA} is a consequence- and the move through the notion of truth provides the explanation [Shapiro, 1998, p. 505].

3. Feferman's program

Tennant [Tennant, 2002] argues against Ketland [Ketland, 1999] and Shapiro [Shapiro, 1998] that Tarski's theory of truth is not the only way we can come to recognize the truth of the Gödel sentence. In particular, Tennant claims, the generalization "All theorems are true" is not the only way to express the soundness of an arithmetical system S. There is, instead, another way to express it, viz., using reflection principles of the form

(pa) If $\overline{\varphi}$ is a primitive recursive sentence and $\overline{\varphi}$ is provable in S, then φ.

As we see, this reflection principle does not use the truth-predicate. Tennnat follows here Feferman [Feferman, 1962], who emphasizes that "Reflection principles are axioms schemata ...which express, insofar as is possible without use of the formal notion of truth, that whatever is derivable in S is true".

Let us take stock. We have discussed two semantic arguments invoked in how we come to recognize that Gödelain sentences are true.

One such argument, due to Tarski, and explicitly described in Shapiro's quote in the last section, uses the generalization "All theorems are true" and can be run in an extension $PA(Tr)$ of PA which, in addition to the truth axioms, allows occurrences of the truth predicate in the induction scheme.

The other semantic argument, described earlier in the second quote from Dummett also uses the truth-predicate. However, Tennant [Tennant, 2002] rephrases it, so that the reference to "the structure of the model" is deleted and the truth-predicate lifted out as required by Feferman's reflection principles. Here is Tennant's formulation of his own semantic argument:

> G is a universally quantified sentence (as it happens, one of Goldbach type, that is, a universal quantification of a primitive-recursive predicate). Every numerical instance of that predicate is provable in the system S. (This claim requires a subargument exploiting Gödel numbering and the representability in S of recursive properties.) Proof in S guarantees *truth*. Hence every numerical instance of G is *true*. So, since G is simply the universal quantification over those numerical instances, it too must be *true* [Tennant, 2002, p. 556].

Tennant shows that this argument can be faithfully represented in a "sufficiently strong" arithmetical system S enriched with reflection principles (with no occurrence of the truth-predicate) in Feferman's style.

I will now describe shortly the main lines of Tennant's argument. Before doing that let me mention what it means for a formal system of arithmetic S to be "sufficiently strong": S represents recursive properties and proves the Diagonalization lemma (i.e., there is a proof in S leading from G to $\neg\exists y Prov_T(\underline{\ulcorner G \urcorner}, y)$;

and there is a proof in the other direction too), and S also proves the equivalence between the Gödelian sentence G and the consistency sentence Con_S. It is known that there are several systems which satisfy this requirement, e.g. PA.

Tennant proposes an extension of S with Feferman's *principle of uniform primitive recursive reflection* (which is more general than the principle (pa) mentioned above):

(UR) Add to S all sentences of the form

$$\forall n(Pr_S(\underline{\ulcorner \psi(\underline{n}) \urcorner})) \rightarrow \forall m \psi(m)$$

where ψ is a primitive recursive formula and $\underline{n}$ is, as before the numeral corresponding to the natural number $n$and $Pr_S(\underline{\ulcorner \psi \urcorner})$ is, like before, an abbreviation for $\exists y Prov_S(\underline{\psi}, y)$

He then shows that in this extension a faithful formalization of the semantical argument described above can be run. The proof goes like this [Tennant, 2002, p. 577]. (We let S^* denote the system S plus (UR)).

Suppose m codes a proof of G in S. Hence by representability (a natural number being the code of a proof in S of a formula is a primitive recursive relation), $S \vdash Prov_S(\underline{\ulcorner G \urcorner}, \underline{m})$, where $Prov_S$ is a primitive recursive formula. But S proves also, from the assumption G, the sentence $\forall y \neg Prov_S(\underline{\ulcorner G \urcorner}, y)$ (i.e. the diagonalization lemma), which by universal instantiation implies $\neg Prov_S(\underline{\ulcorner G \urcorner}, \underline{m})$. Given our assumption that S is consistent, we have a contradiction, from which we conclude that m does not code a proof of G in S. Again by representability we get $S \vdash \neg Prov_S(\underline{\ulcorner G \urcorner}, \underline{m})$. But n has been chosen arbitrarily, hence for every n, there is some proof of $\neg Prov_S(\underline{\ulcorner G \urcorner}, \underline{n})$ in S, from which with the help of (UR) we derive (in S^*) that $\forall y \neg Prov_S(\underline{\ulcorner G \urcorner}, y)$. Finally, by the Diagonalization Lemma, we get G (in S^*).

The penultinate steps requires perhaps some additional clarification. If I understood correctly, "for every n, there is some proof of $\neg Prov_S(\underline{\ulcorner G \urcorner}, \underline{n})$ in S" is just the sentence $\forall n Pr_S(\underline{\ulcorner \psi(\underline{n}) \urcorner})$ in the antecedent of (UR), where $\psi(\underline{n})$ is the primitive recursive sentence $\neg Prov_T(\underline{\ulcorner G \urcorner}, \underline{n})$.

We are then told:

> The foregoing proof justifies the assertion of G. The stronger system S^* contains methods for reflecting on the justification resources of the weaker system S. These methods can be seen at work, in the application, in the proof just give, of various rules of inference that are available in S^* but not in S [Tennant, 2002, p. 577].

The thing which I find somehow problematic in the proof are the penultimate steps:

...But n has been chosen arbitrarily, hence for every n, there is some proof of $\neg Prov_S(\underline{\ulcorner G \urcorner}, \underline{n})$ in S, from which, with the help of (UR), we derive (in S^*) that $\forall y \neg Prov_S(\underline{\ulcorner G \urcorner}, y)$.

I take them to correspond to the informal steps of Tennant's own semantic argument listed earlier in this section. It seems to me that we can justify these steps only on the basis of our intuitive understanding of the standard model, as Dummett pointed out. The principle of uniform recursive reflection (UR) just expresses this understanding in a formal way. We may have eliminated the truth-predicate as required by a minimlist conception of truth, but the justification of (UR) is still grounded in such understanding. This matter is orthogonal to the goal of this essay, so I will not dwell on it.

One can still perhaps argue that Tarski's truth-definition is more general, because it can also account for the intuition that all $S-$theorems are true (sound), and not just the primitive recursive ones. Tennant's response to this objection is that we could add as well to S^* the schema (soundness principle)

$$Prov_S(\underline{\ulcorner \varphi \urcorner}) \rightarrow \varphi,$$

where φ is any sentence in the language of arithmetic. It is known from Löb's theorem that this principle cannot be derivable in S without making S inconsistent. But in the present case we add the soundness principle not to S directly but to S extended with the principle of uniform primitive recursive reflection, and this avoids the inconsistency.

To sum up, I agree with Tennant that the difference between the two semantic arguments is that between saying (Tarski) and showing (Feferman). That is, Tarski's truth theory can state the principle of soundness in one single universal statement "All theorems are true". In this case the "recognition" of the truth of the Gödelian sentence takes the form of a nomological explanation which uses that universal statement [Ketland, 1999; Shapiro, 1998]. On the other side, the Feferman-Tennant framework (S^* extended with the soundness axiom scheme) uses an axiom scheme which can be seen as a list of the infinitely many instances of the universal statement $\forall x(Pr_S(x) \rightarrow Tr(x)$:

$$\begin{array}{c} Pr_S(\underline{\ulcorner \varphi_1 \urcorner}) \rightarrow Tr(\underline{\ulcorner \varphi_1 \urcorner}) \\ Pr_S(\underline{\ulcorner \varphi_2 \urcorner}) \rightarrow Tr(\underline{\ulcorner \varphi_2 \urcorner}) \\ \vdots \end{array}$$

in which the truth-predicate has been eliminated in virtue of the equivalences

$$\begin{array}{c} Tr(\underline{\ulcorner\varphi_1\urcorner}) \leftrightarrow \varphi_1 \\ Tr(\underline{\ulcorner\varphi_2\urcorner}) \leftrightarrow \varphi_2 \\ \vdots \end{array}$$

In this case the recognition of the truth of G does not take the form of a nomological argument (because there is no collection of all these instances into one universal statement). It consists in the apprehension of the proof of G in the extension of e.g. PA with the soundness principle. Truth does not "transcend" proof, truth is just proof (in the extended system).

4. The justification of the extensions

A question which arises quite naturally at this stage is about the justification of different extensions which settle the Gödelian statements, and about the nature of these statements themselves. Is a given extension more justified than another? This question revives an older discussion which goes back to Gödel concerning intrinsic versus extrinsic extensions of a theory which has been the inspiring source for the Feferman program.

Gödel's reflections took place in the context of set theory (*What is Cantor's continuum problem*? [Gödel, 1947]) but they also apply *mutatis mutandis* to arithmetic. Gödel introduced a distinction between an *intrinsic* and *extrinsic* extension of an axiom system. An intrinsic extension, unlike an extrinsic one, is justified on the basis of one grasping the concepts of the base theory. Gödel gave as an example the *Axiom of Determinacy* in set theory that he regarded as an extrinsic axiom because it is not justified by our understanding of sets, in contrast to *Mahlo's axioms* for big cardinals. In addition, Gödel also mentioned intrinsic extensions with undecidable statements (Gödelian sentences) that one recognizes as true in virtue of their meaning, that is, by reflecting on their undecidability.

Gödel's remarks suggest the idea to treat the truth axioms of Tarski's theory of truth as examples of intrinsic extensions of the base theories, whose justification is grounded in our grasping of the concepts of the base theory, that is, natural numbers and operations on natural numbers. In fact this suggestion, which was not made by Gödel, has been explicitly advocated later on by Koellner in his reflections on Gödel's distinctions:

> Let us consider first our conception of natural numbers which is underlying PA. This conception of natural numbers not only justifies the principle of mathematical induction for the language of PA, but for any other extension of the language of PA which has

> a sense. For instance if we extend the language of PA by adding the tarskian truth-predicate and we extend the axioms of PA by adding the tarskan axioms for truth, then, on the basis of our conception of natural numbers, we are justified in accepting the instances of the induction scheme in which the truth-predicate occurs. In the resulting system one can prove Con_{PA}....By contrast, the Axiom of determinacy AD is not justified by our understanding of natural numbers [Koellner, 2006].

Similar ideas have been expressed by Feferman. Starting with the 60's and inspired by Gödel, he addressed the question of the extensions of *schematic formal systems* (formal systems which contain axiom shemes, like ZFC and PA) with new axioms. He started looking for the possibility to generate systematically extensions of such systems whose acceptance was already implicit in the base theory. One of the mechanisms Feferman proposed is *reflection principles*. We saw an illustration of this mechanism when presenting Tennant's ideas. Little by little Feferman also came to consider extensions which contains explicitly a truth-predicate and developed the notion of *reflexive closure of a schematic theory* [Feferman, 1991], which allows for the Induction scheme to range over the truth-predicate. In this case the extended system can prove statements of the form $\forall x(Pr_{PA}(x) \rightarrow Tr(x))$. This has been, as we saw, Tarski's way.

I think there is an important difference between Gödel's notion of intrinsic extension where the new axioms display or unfold the content of the notions of the base theory, and the two extensions of PA introduced in this paper. It seems to me that neither Tarski's extension of PA with his theory of truth, nor Tennant's extension of a sufficiently strong arithmetical system S (e.g. PA) with reflection principles $Prov_S(\underline{\ulcorner\varphi\urcorner}) \rightarrow \varphi$, "unfold" the content of the notion of natural number. None of this extensions is, in my opinion, grounded in our knowledge and understanding of natural numbers but rather "reflect" on the properties of certain methods of proof that have been adopted. That is, although these methods of proof operate on arithmetical and logical resources, they also possess certain properties confered to them by certain philosophical positions which are constitutive of their definitions. The extension axioms or schemata are about these properties (e.g. soundness, truth, consistency) and not about the content of the notion of natural number. Gödelian arithmetical statements as well as their analogues in set theory contain explicit references to these methods of proof, as a consequence of which they inherit an additional content which is not purely arithmetical, or set-theoretical, for that matter. One can find a partial recognition of this point in [Horsten, 2011]:

> Gödelian proofs of G_{ZFC} and Con_{ZFC} are certainly *partly* mathematical in nature. The proof cited above, for example, involves an instance of the principle of mathematical induction, which is a mathematical principle if there ever was one. It is just that such Gödelian proofs are not *purely* mathematical proofs. For they essentially contain the notion of *truth*, which is itself not a mathematical but a philosophical notion. This is not to deny that mathematics can be applied to produce interesting theories of truth. It is just that mathematical theories of truth do, on this view, belong not to pure mathematics but at least to *applied* mathematics, or to the more mathematical part of philosophy [Horsten, 2011].

Horsten refers here to the philosophical notion of truth, and to Gödelian proofs using a truth-predicate, but my main point in this paper is slightly different. It concerns the notions of proof and provability. It is a metamathematical notion which reflects a certain finitistic, philosophical standpoint. By making explicit reference to such notions, Gödelian sentences acquire also a higher-order, not purely numerical content, which depends on the properties of these notions and cannot be reduced to the concept of natural number. One possible way to be more explicit about the higher-order content of Gödelian sentences is through some remarks made by Isaacson [Isaacson, 1991; Isaacson, 1996]. He contrasts arithmetical sentences provable in PA with the Gödelian sentences: the former have a pure arithmetical content, and the system PA which proves them arises out of our undertanding of natural numbers. On the other side, the meaning of Gödelian statements involve our reflections on our understanding of natural numbers.

The ideas discussed in this paper have been debated many times in the post Gödelian era. The contribution of the paper is simply one of emphasis. Myhill, for instance expresses similar ideas in an often quoted passage:

> Indeed it seems to me that the use of the word 'proof' in ordinary non-philosophical mathematical discussion is rather clearly neither a syntactical nor a semantical term. It is as self-contradictory to use methods of proof without admitting their correctness, as it is to make statements without admitting their truth. (I am not using 'self-contradictory' in the sense of formal logic, but roughly as a synonym for 'irrational'.) Therefore if a person who has been using certain methods for proving arithmetical theorems succeeds in making these methods explicit, he is ipso facto committed to the perfectly definite proposition that the use of those methods cannot lead to a false arithmetical statement, for example the statement

> that 0 is equal to 1. By Gödel's technique of arithmetization, which translates every statement of formal deducibility into a statement of arithmetic, any such person is compelled to admit a new arithmetical statement, namely the arithmetized version of the statement that his methods cannot lead to a proof of the statement that 0 is equal to 1. By Gödel's theorem, he could not have established this statement by his previous methods. Hence, as soon as a person makes explicit the tools which he has been using in the construction of arithmetical proofs, he is ipso facto in a position to obtain new arithmetical proofs which he could not have obtained by using those tools alone. The whole process is closely related to what the British philosophical logician W.E. Johnson called 'intuitive induction'; we find ourselves making certain inferences and we thereupon realize that the pattern of those inferences is such as to confer validity on arguments in which they occur. This realization is a demonstrative and rational step quite apart from any question of formalization, though of course the *results* of an intuitive induction can be formalized after the induction has taken place [Myhill, 1960, p. 461].

It is difficult to disagree with these remarks. Myhill, like other commentators I discussed (Horsten) is concerned with the distinction between different kind of proofs. My concern in this paper was, however, with the other side of the coin: the meaning of the Gödelian sentences which are settled by these proofs. The minor point I tried to make was that, by making reference to notions like proof (provability), these sentences have a content which transcend the arithmetical content of purely numerical statements. This is the internal, conceptual reason for which, in some cases (not all; there are Gödelian statements like "I am provable" which are provable), their proof has to mobilize higher-order (meta-theoretical) resources, be they in the form of a truth-theory, a la Tarski, or reflection principles, a la Feferman. I think that Gödel was aware of this fact when he made a distinction between intrinsic extensions with Gödelian sentences and intrinsic extensions with other kind of axioms which unfold the content of the basic notions like natural numbers.

References

Dummett, 1963 – Dummett, M. "The Philosophical Significance of Gödel's Theorem", *Ratio*, Vol. 5, pp. 140–155. Reprinted in: Dummett, M. *Truth and Other Enigmas*, London, Duckworth, 1979, pp. 186–201.

Feferman, 1962 – Feferman, S. "Transfinite recursive progressions of axiomatic theories", *The Journal of Symbolic Logic*, 1962, Vol. 27, pp. 259–316.

Feferman, 1991 – Feferman, S. "Reflections on incompleteness", *Journal of Symbolic Logic*, 1991, Vol. 56, pp. 1–49.

Feferman, 2005 – Feferman, S. "Predicativity", in: *The Oxford handbook of philosophy of mathematics and logic*, ed. by S. Shapiro, Oxford: Oxford University Press, 2005, pp. 590–624.

Gödel, 1947 – Gödel, K. "What is Cantor's continuum problem?", *The American Mathematical Monthly*, 1947, Vol. 54, pp. 515–525.

Horsten, 2011 – Horsten, L. *The Tarskian Turn. Deflationism and Axiomatic Truth*, MIT Press, 2011.

Hyttinen, Sandu, 2004 – Hyttinen, T., Sandu, G. "Deflationism and Arithmetical Truth", *Dialectica*, 2004, Vol. 58, pp. 413–426.

Isaacson, 1991 – Isaacson, D. "Some considerations on arithmetical truth and the omega-rule", in: *Proof, logic, and formalization*, ed. by M. Detlefsen, Routledge, 1991, pp. 49–138.

Isaacson, 1996 – Isaacson, D. "Arithmetical truth and hidden higher-order concepts", in: *Logic Colloquium '85*, Amsterdam: North-Holland, 1987, pp. 147–169. Reprinted in: *The philosophy of mathematics*, ed. by W.D. Hart, Oxford University Press, 1996, pp. 203–224.

Ketland, 1999 – Ketland, J. "Tarski's Paradise and Deflationist Truth", *Mind*, 1999, Vol. 108, pp. 69–94.

Koellner, 2006 – Koellner, P. "On the question of absolute undecidability", *Philosophia Mathematica*, 2006, Vol. 14, pp. 153–188. Revised and reprinted in: *Kurt Gödel: Essays for his Centennial*, S. Feferman, C. Parsons, S.G. Simpson (eds.), Lecture Notes in Logic, Vol. 33. Association of Symbolic Logic, 2009.

Myhill, 1960 – Myhill, J. "Some Remarks om the notion of proof", *Journal of Philosophy*, 1960, Vol. 57, pp. 461–471.

Shapiro, 1998 – Shapiro, S. "Truth and Proof: Through Thick and Thin", *Journal of Philosophy*, 1998, Vol. 95, pp. 493–521.

Tarski, 1956 – Tarski, A. "The Concept of Truth in Formalized Languages", in: *Logic, Semantics, Metamathematics*, ed. by A. Tarski, 2d edition, Oxford University Press, 1956, pp. 152–278.

Tennant, 2002 – Tennant, N. "Deflationism and the Gödel Phenomena", *Mind*, 2002, Vol. 111, pp. 551–582.