## Interpretations

The first incompleteness theorem applies not only to the language of arithmetic but to other languages into which one can translate the language of arithmetic. The notion of "translation" we'll be using comes from Tarski, Mostowski, and Robinson's book *Undecidable Theories*.[1] Let $\mathfrak{L}$ be a language that includes the first-order predicate calculus and might or might not include other logical apparatus as well. We translate the language of arithmetic into $\mathfrak{L}$ by picking a formula $N(x_0)$ of $\mathfrak{L}$ to represent "$x_0$ is a natural number," picking a formula "$Z(x_0)$" to represent "$x_0 = 0$," and picking formulas "$S(x_0,x_1)$," "$A(x_0,x_1,x_2)$," "$M(x_0,x_1,x_2)$," "$E(x_0,x_1,x_2)$," and "$L(x_0,x_1)$," to represent "$x_1$ is the successor of $x_0$," "$x_2 = (x_0 + x_1)$," "$x_2 = (x_0 \cdot x_1)$," "$x_2 = (x_0 \text{ E } x_1)$," and "$x_0 < x_1$," respectively.

As an example, let us take $\mathfrak{L}$ to be the language of set theory, which is the language whose only non-logical symbol is the binary predicate "$\in$" ("is an element of"). Our technique for formulating arithmetical statements within the language of arithmetic is due to John von Neumann.[2] We use the empty set, $\varnothing$, to represent 0, we use $\{\varnothing\}$ to represent 1, we use $\{\varnothing, \{\varnothing\}\}$ to represent 2, and so on, representing a number $n$ as the set of the sets that we use to represent

---

[1] Amsterdam: North-Holland, 1953. The notion we are developing here is what they call *relative interpretation.*

[2] An earlier proposal for reducing number theory to set theory, put forward by Ernst Zermelo, was to identify 0 with the empty set and to identify a successor with the unit set of its immediate predecessor, so that we associate 1 with $\{\varnothing]$, 2 with $\{\{\varnothing\}\}$, 3 with $\{\{\{\varnothing\}\}\}$, and so on. If we're only interested in arithmetic, one suggestion is as good as the other. The reason von Neumann's technique has generally been preferred is that it extends seamlessly into the transfinite. You can read the papers by Zermelo and von Neumann in Jean van Heijenoort's *From Frege to Gödel* (Cambridge, Mass.: Harvard University Press, 1965). The fact that there are multiple equally effective reductions of number theory to set theory has been thought to have profound and disturbing implications for the philosophy of mathematics. See Paul Benacerraf, "What Numbers Could Not Be," in Benacerraf and Hilary Putnams, eds. *Philosophy of Mathematics*, 2nd ed. (Cambridge: Cambridge University Press, 1984).

the numbers less than $n$. One way to think of it is that we reduce number theory to set theory by taking numerals to refer to certain sets. In terms of the reduction, we can say that a number is equal to the set of its predecessors.

"$Z(x)$" will just be "$\sim (\exists y)y \in x$"; this identifies the zero element with the empty set. The successor function is represented by the operation that takes $x$ to $x \cup \{x\}$, so we set "$S(x,y)$" equal to "$(\forall z)(z \in y \leftrightarrow (z \in x \vee z = x))$." "$N(x)$" will say that $x$ has four properties:

$x$ is transitive: $(\forall y)(\forall z)((y \in x \wedge z \in y) \rightarrow z \in x)$.

$x$ is connected: $(\forall y)(\forall z)((y \in x \wedge z \in x) \rightarrow (y \in z \vee y = z \vee z \in y))$.

$x$ is well-founded: $(\forall y)((\exists z)(z \in y \wedge z \in x) \rightarrow (\exists z)((z \in x \wedge z \in y) \wedge \sim (\exists w)(w \in x \wedge w \in y \wedge w \in z)))$.

$x$ contains no limits: $(\forall y)((y \in x \wedge (\exists z)z \in y) \rightarrow (\exists z)(z \in x \wedge (\forall w)(w \in y \leftrightarrow (w \in z \vee w = z))))$.

The third conditions is what gives us the principle of mathematical induction. The most common formalization of the axioms of set theory has an axiom that says every set is well-founded. In the presence of such an axiom, the third clause is superfluous. The fourth clause tells us that every number is either 0 or a successor.

In order express addition, multiplication, and exponentiation set-theoretically, we first need a set-theoretic analogue of Pair, a function that encodes an ordered pair of sets as a single set.[3] For that purpose, we define, for any sets $a$ and $b$,

$$<a.b.> =_{\text{Def}} \{\{a\}, \{a.b\}\}.$$

---

[3] The specific function we use was devised by Kazimierz Kuratowski, although the basic idea was due to Norbert Weiner. Both their papers are in the van Heijenoort volume.

It is easy to prove that, for any *a, b, c,* and *d*, we have $\langle a.B.\rangle = \langle c,d\rangle$ if and only if $a = c$ and $b = d$. Once we have the pairing function, we can get our formulas "$A(x,y,z)$," "$M(x,y,z)$," and "$E(x,y,z)$" by applying our usual method for converting recursive definitions into explicit definitions. "$L(x,y)$" is just "$x \in y$."

Once we have our translation scheme we translate arithmetical formulas into $\mathfrak{L}$ by the following procedure: Given a sentence $\phi$, we eliminate the nesting of function signs. rewriting $\phi$ as a logically equivalent formula in which the atomic formulas take one of the following seven forms:

$$x_i = x_j$$

$$x_i < x_j$$

$$0 = x_i$$

$$x_j = \mathrm{s}x_i$$

$$x_k = (x_i + x_j)$$

$$x_k = (x_i \cdot x_j)$$

$$x_k = (x_i \mathrm{E}\, x_j)$$

We leave the first of these alone, and we replace the others by "$L(x_i, x_j)$," "$Z(x_i)$," "$S(x_i,x_j)$," "$A(x_i,x_j,x_k)$," "$M(x_i,x_j,x_k)$," and "$E(x_i,x_j,x_k)$," respectively, changing bound variables as needed to avoid collisions. Finally, we replace "$(\forall x_i)$" and "$(\exists x_i)$" by "$(\forall x_i)(N(x_i) \rightarrow \ldots)$" and "$(\exists x_i)(N(x_i) \wedge \ldots)$."

To take an example, let's translate (Q4), "$(\forall x)(\forall y)(x + \mathrm{s}y) = \mathrm{s}(x + y)$." We first find an equivalent sentence in which the atomic formulas all have the prescribed form. There are a number of ways to do this, but they're all logically equivalent. This is one:

$$(\forall x_0)(\forall x_1)(\forall x_2)(\forall x_3)(\forall x_4)(\forall x_5)(((x_2 = \mathrm{s}x_1 \wedge x_3 = (x_0 + x_2)) \wedge$$

$$(x_4 = (x_0 + x_1) \wedge x_5 = \mathrm{s}x_4)) \rightarrow x_3 = x_5).$$

Making the substitutions, we get:

$$(\forall x_0)(N(x_0) \rightarrow (\forall x_1)(N(x_1) \rightarrow (\forall x_2)(N(x_2) \rightarrow (\forall x_3)(N(x_3) \rightarrow (\forall x_4)(N(x_4) \rightarrow (\forall x_5)(N(x_5)$$

$$\rightarrow(((S(x_1,x_2) \wedge A(x_0,x_2, \mathrm{x}_3)) \wedge (A(x_0,x_1,x_4) \wedge S(x_4,x_5))) \rightarrow x_3 = x_5))))))).$$

Given an arithmetical theory $\Gamma$ and a theory $\Delta$ expressed in $\mathfrak{L}$, we say that $\Delta$ *interprets* $\Gamma$ if $\Delta$ entails each of the following:

The translation of each of the axioms of $\Gamma$.

$$(\exists x)(\forall y)((N(y) \wedge Z(y)) \leftrightarrow y = x).$$

$$(\forall x)(N(x) \rightarrow (\exists y)(\forall z)((N(z) \wedge S(x,z)) \leftrightarrow z = y))$$

$$(\forall x)(\forall y)((N(x) \wedge N(y)) \rightarrow (\exists z)(\forall w)((N(w) \wedge A(x,y,w)) \leftrightarrow w = z))$$

$$(\forall x)(\forall y)((N(x) \wedge N(y)) \rightarrow (\exists z)(\forall w)((N(w) \wedge M(x,y,w)) \leftrightarrow w = z))$$

$$(\forall x)(\forall y)((N(x) \wedge N(y)) \rightarrow (\exists z)(\forall w)((N(w) \wedge E(x,y,w)) \leftrightarrow w = z))$$

These sentences express the constraints about what's related to what that are built into the function-sign notation.

The following, very weak theory of sets is able to interpret Q:

$$(\forall x)(\forall y)((\forall z)(z \in x \leftrightarrow z \in y) \rightarrow x = y)$$

$$(\exists x)\sim(\exists y)y \in x$$

$$(\forall x)(\forall y)(\exists z)(\forall w)(w \in z \leftrightarrow (w \in x \vee w = y))$$

Let $\Delta$ be a recursively axiomatized theory into which Q can be interpreted. In talking about a theory in $\mathfrak{L}$ being "recursively axiomatized," I am presuming that a system of code numbers has been assigned to the expressions of $\mathfrak{L}$ is a reasonable way. What will be required of

this coding, for our purposes here, is that the function that takes an arithmetical sentence to its translation be recursive. Another way to express the thesis that $\Delta$ is recursively axiomatized is to say that the set of consequences of $\Delta$ is effectively enumerable; this way of putting things depends on the Church-Turing thesis. Another way of saying it is that $\Delta$ can be axiomatized by a single axiom schema. Here we use a theorem of Robert Vaught[4] that a theory in a language built from a finite vocabulary into which we can interpret Q is recursively axiomatizable if and only if it is axiomatizable by a single axiom schema.

Given $\Delta$ a recursively axiomatized theory into which we can interpret Q, there can't be any recursive set *D* that includes the theorems of $\Delta$ and excludes all the sentences refutable in $\Delta$, since if there were such a set, the set of arithmetical sentences whose translations are elements of *D* would be a recursive set of arithmetical sentences that included the theorems of Q and excluded the sentences refutable in Q. It follows that, if $\Delta$ is consistent, it is incomplete. Virtually every known example of an undecidable theory has been obtained this way.

Our notion of interpretation requires that a formula of the language of arithmetic be translated as a formula of $\mathfrak{L}$. It does not require that an arithmetical term be translated as a term of $\mathfrak{L}$, since $\mathfrak{L}$ might be a language like the language of set theory, which has no terms other than variables. What we can do, however, is to translate arithmetical terms into definite descriptions in $\mathfrak{L}$. For example the numeral "[2]" is translated "$(\iota x)(\exists y)(\exists z)((N(x) \wedge N(y) \wedge N(z)) \wedge (Z(z) \wedge S(z,y) \wedge S(y,x))$." We can then use Russell's technique[5] to eliminate the definite descriptions

---

[4] "Axiomatizability by a Schema," *Journal of Symbolic Logic* 32 (1967): 473-479.

[5] "On Denoting," *Mind* n.s. 14 (1905): 479-493. We talked about Russell's account in Logic I.

from formulas of $\mathfrak{L}$. For now, let me ignore this complication and pretend that the formulas of $\mathfrak{L}$ are represented by code numbers that are denoted by numerals in $\mathfrak{L}$, just the way we have it in the language of arithmetic.

Let $\psi(x)$ be a formula of $\mathfrak{L}$. The self-reference lemma, applied within the language of arithmetic, gives us an arithmetical sentence $\phi$ such that

$$Q \vdash (\phi \leftrightarrow (\exists y)(y \text{ is the translation into } \mathfrak{L} \text{ of } [\ulcorner\phi\urcorner] \wedge \psi(y))).$$

Here I am taking advantage of the fact that the function that takes the code number of an arithmetical formula to the code number of its translation can be functionally represented by a $\Sigma$ formula. Let $\theta$ be the translation into $\mathfrak{L}$ of $\phi$. Then

$$Q \vdash (\forall y)(y \text{ is the translation into } \mathfrak{L} \text{ of } [\ulcorner\phi\urcorner] \leftrightarrow y = [\ulcorner\theta\urcorner]).$$

Because $\Delta$ interprets Q,

$$\Delta \vdash (\theta \leftrightarrow \psi(\ulcorner\theta\urcorner)).$$

Thus the self-reference lemma applies not only to the language of arithmetic but to languages into which we can translate the language of arithmetic.

I have been taking it for granted that the "=" sign of the language of arithmetic is translated as the "=" sign of $\mathfrak{L}$. This isn't obligatory. We can pick a formula $I(x,y)$ of $\mathfrak{L}$ to translate "$x = y$," as long as we make sure that our interpreting theory $\Delta$ proves statements that correspond to the facts about identity that are used in proving the theorems of $\Gamma$. Specifically, $\Delta$ should prove that "$I$" designates an equivalence relation:

$$(\forall x)(N(x) \rightarrow I(x,x))$$

$$(\forall x)(\forall y)((N(x) \wedge N(y)) \rightarrow (I(x,y) \rightarrow I(y,x)))$$

$$(\forall x)(\forall y)(\forall z)((N(x) \wedge N(y) \wedge N(z)) \rightarrow ((I(x,y) \wedge I(y,z)) \rightarrow I(x,z)))$$

and it should prove that "*I*" is a congruence:

$$(\forall x)(\forall u)((N(x) \wedge N(u)) \rightarrow ((Z(x) \wedge Z(u)) \rightarrow I(x,u)))$$

$$(\forall x)(\forall y)(\forall u)(\forall v)((N(x) \wedge N(y) \wedge N(u) \wedge N(v)) \rightarrow (((S(x,y) \wedge S(u,v)) \wedge I(x,u)) \rightarrow I(y,v)))$$

$$(\forall x)(\forall y)(\forall z)(\forall u)(\forall v)(\forall w)((N(x) \wedge N(y) \wedge N(z) \wedge N(u) \wedge N(v) \wedge N(w)) \rightarrow (((A(x,y,z) \wedge A(u,v,w)) \wedge (I(x,u) \wedge I(y,v))) \rightarrow I(z,w)))$$

$$(\forall x)(\forall y)(\forall z)(\forall u)(\forall v)(\forall w)((N(x) \wedge N(y) \wedge N(z) \wedge N(u) \wedge N(v) \wedge N(w)) \rightarrow (((M(x,y,z) \wedge M(u,v,w)) \wedge (I(x,u) \wedge I(y,v))) \rightarrow I(z,w)))$$

$$(\forall x)(\forall y)(\forall z)(\forall u)(\forall v)(\forall w)((N(x) \wedge N(y) \wedge N(z) \wedge N(u) \wedge N(v) \wedge N(w)) \rightarrow (((E(x,y,z) \wedge E(u,v,w)) \wedge (I(x,u) \wedge I(y,v))) \rightarrow I(z,w)))$$

$$(\forall x)(\forall y)(\forall u)(\forall v)((N(x) \wedge N(y) \wedge N(u) \wedge N(v)) \rightarrow (((I(x,u) \wedge I(y,v)) \wedge L(x,y)) \rightarrow L(u,v)))$$