Tarski's Undefinability Theorem

Brendan Cordy

Tarski's undefinability theorem states, roughly speaking, that there is no way to express arithmetical truth in first-order logic. The goal of the following is to give a precise presentation of the theorem which should be accessible to anyone with some experience working with first-order logic.

1 The Language of Arithmetic

To begin, we'll fix a language in which we can formally express properties of the natural numbers. Our first order language $\mathcal{L}_A$ has one constant, one unary function symbol, three binary function symbols, and two binary relation symbols, all of which are given below. An expression in $\mathcal{L}_A$ will contain only characters taken from this alphabet of seventeen symbols.

$$\underbrace{0}_{constant} \qquad \underbrace{S \;\; + \;\; \times \;\; E}_{function\ symbols} \qquad \underbrace{= \;\; \leq}_{relation\ symbols} \qquad \underbrace{v \;\; ' \;\; (\;\;) \;\; \neg \;\; \wedge \;\; \vee \;\; \rightarrow \;\; \forall \;\; \exists}_{logical\ symbols}$$

The symbol 0 is the only constant, and the symbol S denotes a unary function, which we'll write using postfix notation. This seems a bit odd but it will clean things up later. You can probably guess the arities of the the other non-logical symbols, which we'll write using infix notation, as usual. Note that the symbol v is a logical variable, as are the expressions v', v'', v''', and so on. We'll often make abbreviations as below when writing expressions of $\mathcal{L}_A$.

$$\overline{3} \;:=\; 0SSS \qquad v_3 \;:=\; v''' \qquad a^b \;:=\; aEb$$

The set $\mathbb{N}$ can be made into an $\mathcal{L}_A$-structure, which I'll call $\mathcal{N}$, in a fairly obvious way. The constant 0 is interpreted as zero, S is interpreted as the successor function, $+$, $\times$, and E are interpreted as addition, multiplication, and exponentiation respectively, and the interpretations of the two relation symbols are also exactly what you'd expect them to be.

A sentence of $\mathcal{L}_A$ will be called *true* when it's true in the structure $\mathcal{N}$. Let's see some examples of true sentences and false sentences of $\mathcal{L}_A$.

$$\forall v_1(v_1 S = v_1 + \overline{1}) \qquad \forall v_1(\exists v_2(v_1 = \overline{2} \times v_2) \rightarrow \exists v_3(v_1 = \overline{2}^{\,v_3}))$$

The first sentence is true in $\mathcal{N}$, by definition of the successor function. The second sentence is false. It says that every even number is a power of 2. It's important to recognize that both of these sentences use only the seventeen symbols above. The abbreviations are just for brevity.

2 Arithmetic Sets and Relations

If $\varphi(v_1)$ is a formula in which v_1 is the only free variable, then we say that $\varphi(v_1)$ *expresses* the set of all $n \in \mathbb{N}$ such that $\varphi(\overline{n})$ is true. Equivalently, $\varphi(v_1)$ expresses the subset A of the natural numbers if $\varphi(\overline{n})$ is true $\iff n \in A$.

Similarly, if $\varphi(v_1, ..., v_k)$ is a formula in which $v_1, ..., v_k$ are the only free variables, then $\varphi(v_1, ..., v_k)$ expresses the k-ary relation $R(x_1, ..., x_k)$ if $\varphi(\overline{n}_1, ..., \overline{n}_k)$ is true $\iff R(n_1, ..., n_k)$.

Often it's not that hard to come up with a formula expressing a given set or relation. The two formulae $\varphi(v_1)$ and $\psi(v_1, v_2)$ below express the set of prime numbers and the less than relation respectively.

$$\varphi(v_1) := \forall v_2 \forall v_3((v_2 \times v_3 = v_1) \to ((v_2 = \overline{1} \wedge v_3 = v_1) \vee (v_2 = v_1 \wedge v_3 = \overline{1})))$$

$$\psi(v_1, v_2) := (v_1 \leq v_2) \wedge \neg(v_1 = v_2)$$

When a set of natural numbers or a relation between natural numbers can be expressed by a formula of $\mathcal{L}_A$, we say that it is *arithmetic* (with emphasis on the third syllable). Note that a function is a special case of a relation, so the same definition works for functions. To spell it out again, a function $f : \mathbb{N} \to \mathbb{N}$ is arithmetic if there is a formula $\varphi(v_1, v_2)$ of $\mathcal{L}_A$ such that $\varphi(\overline{n}, \overline{m}) \iff f(n) = m$.

3 Concatenation in Base Seventeen

It's going to be particularly important that a certain function is arithmetic: the function that takes two natural numbers and concatenates them. We're going to do this in base seventeen for reasons that will soon become apparent. Following the usual convention for hexidecimal numbers, our base seventeen digits will be 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, G.

The base seventeen concatenation function will be denoted $* : \mathbb{N}^2 \to \mathbb{N}$, and written using infix notation. To give a concrete example of the way it works, one could write: $79B5 * CE5F2 = 79B5CE5F2$. The function performs a very simple string manipulation, but it's a bit tricky to describe in arithmetic terms. Intuitively, we need to find out how many digits the second number has, put that many zeros on the end of the first number, and then add the two.

Lemma 3.1 The length function $\ell : \mathbb{N} \to \mathbb{N}$, where $\ell(n)$ gives the number of base seventeen digits of n, is an arithmetic function.

Pf The trick here is to notice that the number of digits in the base b representation of n is given by the smallest number, k, such that $b^k > n$ (try it in base ten).

$$len(v_1, v_2) := (v_1 < \overline{10}^{\,v_2}) \wedge (\forall v_3(v_1 < \overline{10}^{\,v_3} \to v_2 \leq v_3)).$$

Note that 10 here is the number seventeen written in base seventeen, so as a term of $\mathcal{L}_A$, it's a 0 followed by seventeen Ss. Also note that we can use the strict less than relation since we showed that it was arithmetic above, so all instances can be replaced by the formula ψ given earlier.

There is actually one small problem here. The above formula succeeds in capturing the idea that the number of digits in the base seventeen representation of n is the smallest power that seventeen

needs to be raised to to exceed n, but this actually fails for $n = 0$. So to be completely correct, we need to rewrite the formula as below.

$$len(v_1, v_2) := (v_1 = \overline{0} \wedge v_2 = \overline{1}) \vee ((\overline{0} < v_1) \wedge (v_1 < \overline{10}^{\,v_2}) \wedge (\forall v_3(v_1 < \overline{10}^{\,v_3} \rightarrow v_2 \leq v_3)))$$

Thus, we have a formula of $\mathcal{L}_A$ which expresses the function ℓ, and hence ℓ is arithmetic. ■

Theorem 3.1 The base seventeen concatenation function $* : \mathbb{N}^2 \rightarrow \mathbb{N}$ is an arithmetic function.

Pf Now that we have established the lemma, writing down a formula that expresses the concatenation function is actually quite simple.

$$cat(v_1, v_2, v_3) := \exists v_4(len(v_2, v_4) \wedge v_1 \times \overline{10}^{\,v_4} + v_2 = v_3)$$

There it is, the concatenation function $* : \mathbb{N}^2 \rightarrow \mathbb{N}$ is expressed by this formula, and is thus an arithmetic function. ■

Note that the concatenation function is *almost* associative, because zero can cause a problem, namely that $((5C * 0) * 7) = 5C07$, but $(5C * (0 * 7)) = 5C7$. To remedy this, let's make the convention that any non-parenthesized sequence of concatenations associates to the left. I'll now use the concatenation function $*$ in formulae of $\mathcal{L}_A$, with the understanding that every time an atomic formula $x * y = z$ appears, it can be replaced by the actual $\mathcal{L}_A$-formula $cat(x, y, z)$. This business can get a little bit trickier too, for example, take the sentence φ below.

$$\varphi := \exists v_1 \exists v_2 \ (v_1 * \overline{5F7B} * v_2 = \overline{455F7B05})$$

To turn this into an actual $\mathcal{L}_A$-formula, one needs to do the fairly standard trick of composition using the existential quantifier to obtain the formula below, which is an $\mathcal{L}_A$-formula (once the *cat*s and then the *len*s are replaced by their defining formulae).

$$\varphi := \exists v_1 \exists v_2 \exists v_3 \ cat(v_1, \overline{5F7B}, v_3) \wedge cat(v_3, v_2, \overline{455F7B05})$$

This is fairly mechanical, but note that the way this formula is written, the concatenation does in fact associate to the left. For a quick exercise, rewrite it so concatenation associates right instead. For a very tedious exercise, expand the *cat*s, then the *len*s, and so on, so that φ is written using only the seventeen symbols of $\mathcal{L}_A$ and the abbreviations on page one.

4 Gödel Numbers

If you haven't guessed it by now, the reason base seventeen is significant is because $\mathcal{L}_A$ has precisely seventeen symbols, so we can put the symbols of $\mathcal{L}_A$ in correspondence with base seventeen digits.

0	S	$+$	$\times$	E	$=$	$\leq$	v	$'$	$($	$)$	$\neg$	$\wedge$	$\vee$	$\rightarrow$	$\forall$	$\exists$
1	0	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G

Converting an $\mathcal{L}_A$-expression into a number written in base seventeen by replacing each symbol with the corresponding base seventeen digit is called *encoding*. The result of encoding a formula φ is denoted $\ulcorner\varphi\urcorner$. Taking a number written in base seventeen and replacing its digits with the corresponding symbols of $\mathcal{L}_A$ is known as *decoding*. The result of decoding the number n is denoted φ_n. Before moving on, let's see a few examples of encoding and decoding.

$$\text{If } \ \varphi := \forall v_1(v_1 S = v_1 + 0S), \ \text{ then } \ \ulcorner\varphi\urcorner = F789780578210A$$

$$\text{If } \ n = G78897821000005100478 8A, \ \text{ then } \ \varphi_n := \exists v_2(v_1 + \overline{5} = \overline{2}^{\,v_2})$$

Note that the formula φ_n above has a free variable. In general, when decoding a number we can make no assumptions on the sequence of symbols that results. If one decodes an arbitrary number, the result will almost certainly not be a well formed formula. This is not a problem though, the encoding scheme still assigns a unique natural number $\ulcorner\psi\urcorner$ to every $\mathcal{L}_A$-formula ψ, and the encoding and decoding functions are mutually inverse, i.e. $\varphi_{\ulcorner\psi\urcorner} = \psi$, and $\ulcorner\varphi_n\urcorner = n$.

Why swap the digits 0 and 1 in our correspondence? Let's recall that the successor function is written in postfix notation in this document, so that $\overline{5} = 0SSSSS$, and the corresponding base seventeen digits for $0SSSSS$ are 100000, which is seventeen to the fifth power, written in base seventeen. In general, this coding scheme encodes the term $\overline{n}$ as the base seventeen number 10^n, which is particularly convenient.

5 Diagonalization

Still with me? The last and most crucial ingredient we need for Tarski's theorem is the *diagonal function*. Informally, the diagonal function is going to take a natural number n, decode it to produce an expression in the language $\mathcal{L}_A$, which we'll suppose is well-formed and has one free variable, substitute $\overline{n}$ in for that free variable, and then take the Gödel number of the resulting $\mathcal{L}_A$-sentence.

$$n \ \overset{decode}{\longrightarrow} \ \varphi_n \ \overset{substitute}{\longrightarrow} \ \varphi_n(\overline{n}) \ \overset{encode}{\longrightarrow} \ \ulcorner\varphi_n(\overline{n})\urcorner$$

But wait, decoding an arbitrary number produces an arbitrary sequence of $\mathcal{L}_A$ symbols, so why should φ_n be a well-formed formula with one free variable? Of course it isn't in general, but this won't actually pose a problem, as we'll see below. I assumed that φ_n has one free variable to give an intuitive picture of how the diagonal function works.

Definition 5.1 The diagonal function $d : \mathbb{N} \to \mathbb{N}$ is defined by $d(n) = \ulcorner \forall v_1(v_1 = \overline{n} \to \varphi_n) \urcorner$.

Assuming that φ_n is a well-formed formula with one free variable, v_1, the sentences $\varphi_n(\overline{n})$ and $\forall v_1(v_1 = \overline{n} \to \varphi_n)$ are equivalent. Proving their equivalence is an easy exercise in formal derivation, and I want to emphasize this, because we'll use it later on. However, $\forall v_1(v_1 = \overline{n} \to \varphi_n)$ is an $\mathcal{L}_A$-expression even if φ_n is not a well-formed formula with one free variable. This means that the diagonal function defined above is a well-defined total function.

If we actually do the encoding on the right hand side of the definition, we can write $d(n) = F789785 * \ulcorner\overline{n}\urcorner * E * \ulcorner\varphi_n\urcorner * A$. By definition of the encoding in the last section, $\ulcorner\overline{n}\urcorner = 10^n$ and $\ulcorner\varphi_n\urcorner = n$, so we obtain $d(n) = F789785 * 10^n * E * n * A$.

Theorem 5.2 The diagonal function $d : \mathbb{N} \to \mathbb{N}$ is arithmetic.

Pf This is essentially a corollary of Theorem 3.1. The diagonal function is expressed by the formula $\delta(v_1, v_2)$ written below.

$$\delta(v_1, v_2) := \exists v_3(v_3 = \overline{10}^{\,v_1} \wedge\ v_2 = \overline{F789785} * v_3 * \overline{E} * v_1 * \overline{A}) \quad \blacksquare$$

Definition 5.3 Given $S \subseteq \mathbb{N}$, define $S^* \subseteq \mathbb{N}$ by $n \in S^* \iff d(n) \in S$.

S^* is the set of natural numbers that end up in the set S after being diagonalized, in other words, S^* is the preimage of S under the diagonal function.

Lemma 5.4 If S is arithmetic, then so is S^*.

Pf Since S is arithmetic, there is a formula $\psi(v_1)$ expressing it, and as we saw earlier, the diagonal function is expressed by the formula $\delta(v_1, v_2)$. By definition, $n \in S^*$ if $d(n) = m$ and $m \in S$. It should now be easy to see that S^* is expressed by the formula $\sigma(v_1) := \exists v_2(\delta(v_1, v_2) \wedge \psi(v_2))$. $\blacksquare$

Theorem 5.5 (Tarski) The set of Gödel numbers of true sentences, T, is not arithmetic.

Pf Suppose that T is arithmetic, then there is a formula $\varphi(v_1)$ with $\varphi(\overline{n}) \iff n \in T$. Therefore, $\neg\varphi(\overline{n}) \iff n \in \widetilde{T}$, where $\widetilde{T}$ denotes the complement of T, and by Lemma 5.4, there is a formula $\psi(v_1)$ such that $\psi(\overline{n}) \iff n \in \widetilde{T}^*$. Recall that $n \in \widetilde{T}^*$ means that $d(n) \in \widetilde{T}$, so $\psi(\overline{n})$ is true when $d(n)$ is *not* the Gödel number of a true sentence.

Alright, now here comes the clever bit. Let's encode the formula $\psi(v_1)$, and call its Gödel number g, that is, let $g = \ulcorner\psi(v_1)\urcorner$, and consider the sentence $\psi(\overline{g})$. Let's try to figure out if this sentence is true, using the definition of $\psi(v_1)$.

$$\begin{aligned}
\psi(\overline{g}) &\iff g \in \widetilde{T}^* \\
&\iff d(g) \in \widetilde{T} \\
&\iff \ulcorner\forall v_1(v_1 = \overline{g} \to \varphi_g)\urcorner \in \widetilde{T} \\
&\iff \ulcorner\forall v_1(v_1 = \overline{g} \to \psi(v_1))\urcorner \in \widetilde{T} \\
&\iff \ulcorner\psi(\overline{g})\urcorner \in \widetilde{T} \\
&\iff \ulcorner\psi(\overline{g})\urcorner \notin T
\end{aligned}$$

Thus, $\psi(\overline{g})$ is a true sentence if and only if its own Gödel number is *not* the Gödel number of a true sentence, which is a clear contradiction. Therefore, our initial assumption that T is arithmetic must be false. $\blacksquare$

6 Fixed Points & Liars

Where did the clever idea to consider the sentence $\psi(\overline{g})$ come from? Well, $\psi(\overline{n})$ asserts '$d(n)$ is not the Gödel number of a true sentence', so $\psi(\overline{n})$ is a true sentence exactly when $\varphi_{d(n)}$ is not. It would clearly be absurd then if $\psi(\overline{n})$ and $\varphi_{d(n)}$ had the same truth value. But the proof above is by contradiction, so that's just what we need. We're looking for a natural number x that satisfies the 'equation' $\psi(\overline{x}) \Leftrightarrow \varphi_{d(x)}$.

If we, by abuse of notation, write $d(x)$ as $\ulcorner \varphi_x(\overline{x}) \urcorner$ (recall from the discussion following Definition 5.1 that this is actually not such a terrible offense), then the right hand side of the equation above becomes $\varphi_{\ulcorner \varphi_x(\overline{x}) \urcorner}$, which is just $\varphi_x(\overline{x})$ by definition of the encoding and decoding operations (recall $\varphi_{\ulcorner \varphi_x(\overline{x}) \urcorner}$ should be read as 'the expression whose Gödel number is the Gödel number of the expression $\varphi_x(\overline{x})$'). So in spirit, we're trying to solve $\psi(\overline{x}) \Leftrightarrow \varphi_x(\overline{x})$. At this point, it's not hard to see that $g = \ulcorner \psi(\overline{x}) \urcorner$ is the number we're looking for.

Note that $\psi(\overline{g})$ is a self-referential sentence that asserts its own falsehood. The proof of Tarski's theorem works by using the diagonal function to create a formal version of the well-known liar paradox. There's a great informal introduction to the diagonal function and how it can be used to create self-referential statements in [4].

The proof of Gödel's first incompleteness theorem uses a very similar argument. However, much of that proof is devoted to establishing that the set of *provable* statements in any recursively axiomatized first-order theory *is arithmetic*, which is a long and involved process. From that fact alone and Tarski's theorem, a weakened version of the first incompleteness theorem immediately follows. Namely, that there is no recursively axiomatized first-order theory of arithmetic having the property that the statements provable in that theory are exactly the statements true in the structure $\mathcal{N}$.

Acknowledgements

This work is based on Raymond Smullyan's excellent exposition in the second chapter of [3]. He attributes the Gödel numbering used here to Quine [1], who did it in characteristically minimalist fashion, using an alphabet of only nine symbols. If you're interested in reading more about Gödel's theorems, I would recommend [6] as an all around great introductory logic book featuring the usual modern approach, and [2] if you have some background in logic, and want to see all the gory details. The original presentation of Tarski's theorem given by Tarski himself can be found in [5].

References

[1] W. V. Quine. *Mathematical Logic.* Norton, 1940.

[2] J. Shoenfield. *Mathematical Logic.* CRC Press, 2001.

[3] R. Smullyan. *Gödel's Incompleteness Theorems.* Oxford University Press, 1992.

[4] R. Smullyan. *Satan, Cantor, and Infinity: Mind-Boggling Puzzles.* Knopf, 1992.

[5] A. Tarski. The concept of truth in formalized languages. In J. Corcoran, editor, *Logic, Semantics and Metamathematics.* Hackett, 1983. (English language translation of Tarski's 1936 *Der Wahrheitsbegriff in den Formalisierten Sprachen*).

[6] D. Van Dalen. *Logic and Structure, 4th Ed.* Springer, 2008.