The diagonal lemma as the formalized Grelling paradox

György Serény (sereny@math.bme.hu)

Introduction

Gödel's diagonal lemma (expressing formally the ability of first-order arithmetic to 'talk about itself') plays a key role in the proof of three main limitative theorems of logic:

Diagonal lemma →
- Gödel's first incompleteness theorem
- Tarski's theorem on the undefinability of truth
- Church's theorem on the undecidability of provability

Still, the proof of the lemma as it is presented in textbooks and handouts on logic is not self-evident to say the least:

> [The] proof [is] quite simple but rather tricky and difficult to conceptualize. (*Handbook of Proof Theory*)

> [This] result is a cornerstone of modern logic. [...] You would hope that such a deep theorem would have an insightful proof. No such luck. *I don't know anyone who thinks he has a fully satisfying understanding of why the Self-referential Lemma works. It has a rabbit-out-of-a-hat quality for everyone.* (*Handout for the course 24.242 Logic II, MIT, Spring 2002*)

However, the proof of the lemma can be made completely transparent by recognizing that it is simply a straightforward translation of the Grelling paradox into first-order arithmetic.

Notation Our formal language is that of first-order arithmetic; Fm is the set of formulas with at most one free variable; g is any one of the standard Gödel numberings; N denotes the set of Gödel numbers of formulas in Fm; Q stands for Robinson arithmetic. The closed terms corresponding to natural numbers are denoted by the numbers themselves.

Diagonal Lemma

For any formula $\varphi \in Fm$, there is a sentence λ such that $\quad Q \vdash \lambda \longleftrightarrow \varphi(g(\lambda))\,.$

Proof idea

An adjective is called *heterological* if the property denoted by the adjective does not hold for the adjective itself; e.g. 'long', 'German', 'monosyllabic' are heterological.

Grelling's paradox: *'heterological' is heterological*. This sentence is true just in case it is false, therefore, in effect, SAYS OF ITSELF THAT IT IS FALSE.[1]

The Lemma is about (the existence of) a first-order sentence that, informally speaking, SAYS OF ITSELF THAT IT HAS A GIVEN PROPERTY.

Outline of the proof:

FIRST STEP: An ordinary language modification of the paradox to get a sentence that

(a) is not about an adjective but about a sentence
(b) instead of asserting its own falsehood, says of itself that it has an arbitrary (but fixed) property.

SECOND STEP: Translating the result of the first step (i.e. the sentence saying of itself that it has a given property) into first-order arithmetic.

[1] What is truly important is that, contrary to the Liar, this paradoxical sentence achieves self-reference without using an indexical.

Proof.

FIRST STEP

'heterological' is heterological →

Replace 'heterological' by 'x *is heterological*' and let the self-application of an open sentence be the substitution of its name[2] for the variable in it.

'x *is heterological*' is heterological →

x is heterological *just in case* the sentence obtained by substituting the name of x for the variable in it is false. Use this definition (twice) and replace 'is false' by 'has property p'.

the sentence obtained by substituting the name of '*the sentence obtained by substituting the name of* x *for the variable in it has property* p' for the variable in it has property p[3]

[2] The name of a sentence is the sentence itself between quotation marks.

[3] This sentence, indeed, says of itself that it has property p. Actually, let $s(x)$ be the open sentence '*the sentence obtained by substituting the name of* x *for the variable in it has property* p'. Then, by definition, for *any* open sentence o with a single variable, $s(\text{'}o\text{'})$ says that $o(\text{'}o\text{'})$ has property p. In the particular case when o is just s, we obtain: $s(\text{'}s\text{'})$ says that $s(\text{'}s\text{'})$ has property p.

SECOND STEP

Let $\varphi \in Fm$ be arbitrary.

Informal version		**Formal version** (the variable x runs over formulas in Fm)
x has property p	⟷	$\varphi(g(x))$
the sentence obtained by substituting the name of x for the variable in it	⟷	$x(g(x))$
the sentence obtained by substituting the name of x for the variable in it has property p	⟷	$\varphi\big(g[x(g(x))]\big)$
(s) the sentence obtained by substituting the name of x for the variable in it has property p	⟷	$\varphi\big(g[x(g(x))]\big)$ should be represented *within* Fm: $\eta(g(x))$
Let x be (the name of) s		Let x be η
the sentence obtained by substituting the name of '*the sentence obtained by substituting the name of* x *for the variable in it has property* p' for the variable in it has property p	⟷	$\boldsymbol{\lambda = \eta(g(\eta))}$
		$\boldsymbol{Q \vdash \lambda \longleftrightarrow \varphi(g(\lambda))}$

where $\eta \in Fm$ is such that, for *every* $\psi \in Fm$,
$Q \vdash \eta(g(\psi)) \longleftrightarrow \varphi\big(g[\psi(g(\psi))]\big)$ [4]

QED

[4] Such an η exists. Indeed, $Q \vdash \eta(g(\psi)) \longleftrightarrow \varphi\big(g[\psi(g(\psi))]\big)$ for every $\psi \in Fm$ iff $Q \vdash \eta(n) \longleftrightarrow \varphi(g[g^{-1}(n)(n)])$ for every $n \in N$ (g^{-1} is the inverse of g). Let $f(n) = g[g^{-1}(n)(n)]$ if $n \in N$ and $f(n) = 0$ otherwise. Then f is recursive and hence representable in Q, and (by elementary first-order logic), up to provable equivalence in Q, the result of substituting a representable function into a formula can also be expressed by a formula. Therefore, there is an $\eta \in Fm$ such that, for every $n \in N$, $Q \vdash \eta(n) \longleftrightarrow \varphi(f(n))$, or equivalently, for every $\psi \in Fm$, $Q \vdash \eta(g(\psi)) \longleftrightarrow \varphi\big(g[\psi(g(\psi))]\big)$.