NOTAS SOBRE EL "WAHRHEITSBEGRIFF", II*

MARIO GÓMEZ TORRENTE

6. El enunciado del "Teorema I"

Recordemos que en la primera parte de este artículo llamamos LTGC a la versión de la teoría simple de los tipos finitos de Tarski (1933). Consideremos ahora LTGC más un vocabulario y axiomas apropiados para hablar y probar cosas sobre la sintaxis de LTGC, de acuerdo con los requisitos de Tarski para la construcción de metalenguajes apropiados para definir nociones semánticas (que revisamos en la sección 3). Este lenguaje, al que podemos llamar LTGC+, es un metalenguaje apropiado para LTGC; LTGC+ contiene también traducciones de los signos de LTGC, pues incluye a LTGC. (En esencia, el aparato lógico-matemático de LTGC+ ha sido el aparato lógico-matemático en todos los metalenguajes usados por Tarski, pues, como vimos, es tan potente matemáticamente como la teoría de tipos informal que de hecho ha usado).

La cuestión que se plantea Tarski es, entonces, si es posible definir en LTGC+ un predicado que satisfaga la convención T para oraciones de LTGC. Y la respuesta, como vimos que nos adelantaba, es negativa.

Observemos que la imposibilidad de definir un predicado semejante no quiere decir que el conjunto de las oraciones

* Partes de este trabajo fueron presentadas en conferencias en la Facultad de Filosofía y Letras de la Universidad de Valencia (1997), la Facultad de Filosofía y Letras de la Universidad de Buenos Aires (2000) y el Institut d'Histoire et Philosophie des Sciences et des Techniques, Centre National de la Recherche Scientifique, Universidad de París I (2001), y en ponencias en el Congreso del Centenario de Alfred Tarski en Varsovia (2001) y el Simposio Internacional sobre el Círculo de Viena y el Empirismo Lógico en Viena (2001). Agradezco a los auditorios de estas pláticas sus útiles comentarios.

La primera parte de este trabajo se publicó en el Nº 1 del volumen XXI.

verdaderas de LTGC no sea definible en LTGC+. Puede que haya un predicado constructible en LTGC+ cuya *extensión* (dada la interpretación deseada de LTGC+) sea el conjunto de las oraciones verdaderas de LTGC, aunque no sea posible *demostrar* todos los bicondicionales pertinentes en el cálculo de LTGC+. Ahora bien, el *enunciado* que Tarski hace de su teorema incluye un resultado adicional al de la inexistencia de un predicado que satisfaga la convención T, y este resultado excluye de hecho la existencia de un predicado de LTGC+ cuya extensión sea el conjunto de las oraciones verdaderas de LTGC. Explicaremos por qué en un segundo.

Primero recordemos que la convención T para un predicado *OV* de oraciones de LTGC dirá (en parte) que

> (T) son teoremas del sistema deductivo de LTGC+ todas las oraciones de la forma *OV*(*x*) syss *p*,

donde en lugar de '*x*' ponemos un nombre descriptivo-estructural o en LTGC+ de una oración de LTGC, y en lugar de '*p*' ponemos esa oración (que es también una oración de LTGC+).

La cuestión planteada por Tarski la resuelve el siguiente teorema de dos partes:

> Teorema I. (α) Si *OV* es un predicado de expresiones definido en LTGC+, será posible derivar en el cálculo de LTGC+ la negación de una de las oraciones de la forma "*OV*(*x*) syss *p*", mencionadas en la convención (T);
> (β) si el cálculo de LTGC+ es consistente, entonces no habrá ningún predicado definido en LTGC+ que satisfaga la convención (T) (cf. Tarski (1933), p. 247).

La parte (β) es una consecuencia trivial de la parte (α), pues un sistema consistente no puede contener a una oración y a su negación.

Observemos ahora que de la parte (α) se sigue que el conjunto de oraciones *verdaderas* de LTGC *no* es definible en LTGC+, por un argumento informal similar al que usamos en

la sección 3 para mostrar que un predicado que satisficiera la convención T era coextensional con el predicado intuitivo de verdad. Ese argumento podría proceder como sigue. Supongamos que hubiera un predicado *OV* constructible en LTGC+ cuya extensión fuera precisamente el conjunto de oraciones verdaderas de LTGC. Entonces (α) nos dice que hay una oración *p* tal que "¬(*OV*('*p*') si y sólo si *p*)" es demostrable en el cálculo deductivo de LTGC+. Como el cálculo deductivo de LTGC+ sólo permite obtener oraciones verdaderas, sabemos que hay dos alternativas: *OV*('*p*') y no es el caso que *p*, o *p* y no es el caso que *OV*('*p*'). Pero claramente en ninguno de esos casos puede *OV* ser coextensional con el predicado intuitivo de verdad[1].

Tarski sin duda se refiere a este argumento o uno similar en Tarski (1933), p. 254, cuando dice que "[el Teorema I] muestra que es imposible definir en la metateoría una clase de oraciones del lenguaje estudiado que contenga exclusivamente oraciones materialmente verdaderas y que sea al mismo tiempo completo [en el sentido de que para toda oración, se deduzca de él ésta o su negación]". Una clase de oraciones "materialmente verdaderas" y completa es por ejemplo la clase de las oraciones verdaderas de LTGC. Pero su indefinibilidad en LTGC+ no se sigue de la parte (β) del Teorema I, sino sólo de la parte (α). Nótese que 'materialmente verdaderas' es aquí el predicado intuitivo de verdad para LTGC, pues el Teorema I muestra precisamente que no hay un predicado definido de verdad para LTGC.

Ahora bien, uno podría formalizar un argumento informal como el mencionado si tuviera a su disposición un predicado definido de verdad para LTGC, definido usando una

[1] El hecho de que el teorema de Tarski muestra que el metalenguaje no puede contener un predicado coextensional con el de verdad queda oscurecido en la mayoría de las presentaciones más recientes de teoremas similares al de Tarski, en las que el resultado que se enuncia como demostrado (por el razonamiento de Tarski) es el más débil de que ningún predicado del metalenguaje satisface la convención T, es decir, la parte (β) del teorema original (cf. por ejemplo, Lindström (1997), p. 17, Robbin (1969), p. 119, Smullyan (1992), p. 104). Volveremos sobre este punto.

metateoría diferente a LTGC+. Como veremos, un predicado tal lo podemos construir con la extensión de sus métodos que hace Tarski en el *Nachwort* a (1933) aparecido en (1935). Esta extensión, como también veremos, afectará a la potencia del aparato matemático aceptado en toda metateoría, que pasará a ser un aparato más potente.

Pues bien, que Tarski se refiere al argumento informal que mencionamos o uno similar lo confirma de manera muy exacta su forma de proceder en textos posteriores a la fecha de publicación del *Nachwort*. En Tarski (1939), por ejemplo, el teorema numerado como 6.2 da una formulación rigurosa de la afirmación de Tarski citada dos párrafos atrás:

> Sea *OV* un predicado que satisface la convención T para LTGC (un predicado definido en una metateoría más potente que LTGC+). Si *E*(x) es un predicado de LTGC+ del que podemos probar que $E \subseteq OV$ entonces hay una oración *p* tal que ni *p* ni $\neg p$ pertenecen a la extensión de *E*.

La prueba de Tarski es esencialmente la del argumento intuitivo anterior, aunque proporciona la información adicional de que *p* es verdadera. La prueba es la siguiente. La parte (α) de su Teorema I nos dice que hay una oración *p* tal que "$\neg(E(\text{'}p\text{'})$ si y sólo si $p)$" es demostrable en LTGC+. Hay entonces dos alternativas: (1) $E(\text{'}p\text{'})$ y $\neg p$, o (2) *p* y $\neg E(\text{'}p\text{'})$. Tarski muestra primero que *p* es *OV*: si no lo fuera, entonces $\neg p$ sería *OV* y, por la convención T, $\neg p$; se daría la alternativa (1), y tendríamos que $E(\text{'}p\text{'})$ y, por la hipótesis del teorema, $OV(\text{'}p\text{'})$. Por otro lado, ni *p* ni $\neg p$ pertenecen al conjunto definido por *E*(x). Si *p* perteneciera al conjunto, entonces se daría la alternativa (1) y $\neg p$; luego $\neg p$ es *OV* (por la convención T) y *p* no es *OV*; y al mismo tiempo, por la hipótesis del teorema, *p* es *OV*, que es absurdo. Si $\neg p$ perteneciera al conjunto definido por *E*(x), entonces tendríamos que $\neg p$ es *OV*, y por tanto que *p* no es *OV*, en contra de lo que sabemos. Tarski subraya que el núcleo de la prueba es la parte (α) de su Teorema I, a

la que refiere explícitamente al lector, y no menciona la parte (β) (cf. Tarski (1939), n. 7, p. 107)[2].

Así, en (1939) Tarski convierte el argumento informal en uno riguroso (para lo cual usa un predicado definido de verdad en una teoría más potente que LTGC+), y esto confirma que tenía en mente el argumento informal en 1933. Pero subrayemos que *la parte (α)* del Teorema I muestra la *indefinibilidad* del conjunto de oraciones verdaderas de LTGC en LTGC+ (aunque la "convicción subjetiva" de que hace esto sólo se pueda alcanzar por medio del argumento informal al que hemos aludido), *incluso sin necesidad de suponer que tenemos a nuestra disposición un predicado definido* (o primitivo) de verdad para LTGC. Esta era la situación de Tarski en 1933. Y desde luego en el enunciado de su Teorema I (tanto de la parte (α) como de la (β)) no se usa ninguna noción de verdad. Esto es un logro muy significativo, que será bueno recordar al hablar de la relación entre el resultado acerca de la verdad de Gödel y el resultado de Tarski.

7. La prueba del "Teorema I"

La prueba que da Tarski de la parte (α) está muy comprimida, pero nos da sobradas indicaciones de cómo habría de reconstruirse. Lo que sigue no es una reconstrucción completa (reconstruir los detalles técnicos sería una tarea compleja), sino una reconstrucción informal como la de Tarski pero que espero sea más accesible al lector moderno que no está tan familiarizado como el de 1933 con lenguajes como LTGC[3].

[2] Como veremos luego, en la sección 9, este argumento es una generalización de uno que ya aparece en el *Nachwort* de (1935), donde $E(x)$ es substituido por un predicado particular de demostrabilidad en el sistema formal. La conclusión es en este caso que p es una oración verdadera e indecidible, y es por tanto una versión particularmente sencilla de demostrar del primer teorema de incompleción de Gödel.

[3] Quizá también pueda ayudar el siguiente esbozo para un sistema formal más familiar. Véase a LTGC como si fuera un sistema formal L para la aritmética de primer orden. Supongamos que lo extendemos con una

Lo primero que hay que observar es que, como recuerda Tarski, la aritmética de los números naturales puede desarrollarse en la teoría simple de los tipos finitos (y por tanto en LTGC+ (y en LTGC)) usando esencialmente las definiciones logicistas. Recordemos que los números naturales se definen en la construcción logicista como ciertas clases de clases de individuos. Las relaciones aritméticas se definen como ciertas relaciones entre tales clases y las operaciones aritméticas como operaciones de tales clases en clases similares.

Para cada número natural k, en LTGC existe una fórmula con una única variable 'n' libre, que simbolizaremos '$\iota_k(n)$', que intuitivamente dice que (el valor de) n es la clase de clases de individuos identificada en la construcción logicista con el número k ('n', naturalmente, será una variable sobre clases de clases de individuos).

Tarski observa, como Gödel había observado, que las expresiones de LTGC pueden ponerse en correspondencia 1-1 con los números naturales, y por tanto con las clases correspondientes de la construcción logicista. Es más, en LTGC+

---

teoría para hablar de la sintaxis de L (éste no sería un metalenguaje tarskiano, ya que no incluiría un "sistema suficientemente desarrollado de lógica matemática", pero los propósitos ahora son meramente ilustrativos). Llamemos a la extensión L+. Entonces L+ puede ser "interpretado" en L usando técnicas gödelianas, con lo cual Tarski quiere decir, aproximadamente, que uno puede traducir L+ a L aritmetizándolo de tal manera que las traducciones de teoremas de L+ serán teoremas de L. Ahora supongamos que $\phi(n)$ es la n-ésima expresión de L (y que ϕ es definible en L+), que $d(x)$ es una función diagonal para expresiones de L (definible en L+) basada en la substitución de variables por numerales, y sea $\mathbf{n}$ el numeral de n. Entonces para todo predicado A(x) de expresiones de L definido en L+ uno puede demostrar en L+ la oración general de la forma $\forall n[\neg A(d(\phi(n)))\leftrightarrow\psi(n)]$, donde $\psi(n)$ es una fórmula con una variable libre construida enteramente en L (por medio de la aritmetización del predicado de L+ '$\neg A(d(\phi(n)))$'). Entonces uno puede demostrar también la oración $\neg A(d(\phi(\mathbf{k})))\leftrightarrow\psi(\mathbf{k})$, donde k es el número de Gödel de $\psi(n)$. $d(\phi(\mathbf{k}))$ es la oración $\psi(\mathbf{k})$, que es pues un "punto fijo" del predicado $\neg A(x)$: "dice de sí misma" que tiene la propiedad expresada por $\neg A(x)$. La oración $\neg A(d(\phi(\mathbf{k})))\leftrightarrow\psi(\mathbf{k})$ es ya equivalente a la oración $\neg[A(d(\phi(\mathbf{k})))\leftrightarrow\psi(\mathbf{k})]$, y la parte ($\alpha$) del Teorema I queda demostrada.

podemos definir una correspondencia tal: hay una fórmula $\phi_n=e$, con 'n' y 'e' como únicas variables libres, que define una función 1-1 entre números naturales (logicistas) y expresiones de LTGC. (A las antiimágenes en una correspondencia tal se las llama hoy 'números de Gödel' de las expresiones correspondientes).

"De esta forma", dice Tarski, "el metalenguaje recibe una interpretación en la aritmética de los números naturales e indirectamente en el lenguaje de la teoría general de clases" (Tarski (1933), pp. 249-250). Con esto quiere decir que la parte sintáctica de LTGC+ (el "+", por así decir) recibe una interpretación en la aritmética (el resto de LTGC+ conserva su interpretación normal, que es la misma de LTGC): a cada nombre de una expresión en LTGC+ se lo puede interpretar por medio del número natural que es su antiimagen en ϕ y a cada símbolo de relación entre expresiones de LTGC se lo puede interpretar por medio de la relación inducida entre los correspondientes números naturales por la función denotada por ϕ. Estos números naturales y relaciones entre ellos son definibles en LTGC, y además, y fundamentalmente, muchos teoremas básicos acerca de nociones sintácticas se convierten "indirectamente" en teoremas de LTGC.

La prueba es entonces como sigue. Supongamos que *OV* es un predicado de oraciones (de LTGC) definido en LTGC+. Por la última observación del párrafo anterior, a *OV* le corresponde entonces un predicado de los números naturales (logicistas) definido enteramente en LTGC.

La siguiente es una expresión (o una abreviatura de una expresión) de LTGC+ (usando el artilugio de Quine para abreviar locuciones descriptivo-estructurales):

$$\neg OV(\ulcorner \exists n(\iota_n(n) \wedge \phi_n) \urcorner).$$

Esta expresión, como dice Tarski, es una fórmula abierta del metalenguaje con una única variable libre, 'n', que recorre números naturales (construidos de la forma logicista). "Dice" que '*OV*' no es verdadero de "el resultado de concatenar el

cuantificador existencial con la variable 'n' con el paréntesis izquierdo con la fórmula que dice que el valor de 'n' es n con el signo de conjunción con la n-ésima fórmula en la enumeración ϕ con el paréntesis derecho". Es importante notar que esta expresión tanto *menciona* como *usa* la variable 'n'. Intuitivamente, las apariciones en subíndice son las apariciones donde 'n' está siendo usada, y es una auténtica variable de LTGC+. Las otras apariciones, si tienen lugar dentro de semicomillas, simbolizan *menciones* de la variable 'n' de LTGC; abrevian *nombres* de esta variable en LTGC+. A veces a la expresión $\ulcorner\exists n(\iota_m(n)\wedge\phi_m)\urcorner$ para un número particular m (o a una expresión que sirva para los mismos propósitos en el lenguaje que estemos considerando) se la llama la 'diagonalización de ϕ_m'[4], y a la función que asigna a m el número de Gödel de su diagonalización, la 'función diagonal' (de ϕ). Intuitivamente, la diagonalización de una fórmula que tenga a 'n' como única variable libre dice que esa fórmula es satisfecha por su propio número de Gödel, o que "se dice con verdad de sí misma".

La fórmula abierta $\neg OV(\ulcorner\exists n(\iota_n(n)\wedge\phi_n)\urcorner)$ de LTGC+ tiene, por las observaciones del penúltimo y antepenúltimo párrafos, una fórmula $\psi(n)$ (con 'n' como única variable libre, usada, no mencionada —nótese que no aparece dentro de semicomillas—) equivalente a ella para todos los valores de 'n' pero construida enteramente en LTGC. (Esto se demostraría en dos pasos: mostrando primero que la función diagonal es definible en LTGC, y luego que la función que asigna al número de una expresión x el número de la expresión $\neg OV(x)$ es una función de-

[4] El efecto de la diagonalización se puede conseguir con diversos tipos de fórmulas, dependiendo del lenguaje que estemos considerando. Un conocido tipo de diagonalización para fórmulas del lenguaje de la aritmética de primer orden ha sido llamado 'tarskificación' por R. Smullyan (véase Smullyan (1994), p. 82). Tiene interés histórico notar que la tarskificación (introducida en Tarski, Mostowski y Robinson (1953)) está inspirada en el tipo de diagonalización para fórmulas de LTGC que acabamos de ver, al usar las propiedades de la cuantificación existencial en lugar de la posibilidad de definir aritméticamente la relación de substitución de variables por numerales, posibilidad usada en las pruebas de Gödel.

finible en LTGC). Entonces el siguiente bicondicional generalizado es un teorema del cálculo deductivo de LTGC+:

$$(1)\ \forall n\ [\neg OV(\ulcorner \exists n(\iota_n(n) \wedge \phi_n)\urcorner) \leftrightarrow \psi(n)].$$

Por otro lado, sabemos que la fórmula abierta $\psi(n)$ es una fórmula de la enumeración ϕ. Digamos que $\psi(n)=\phi_k$. Entonces del hecho de que (1) sea un teorema del cálculo de LTGC+ se sigue que (2) también lo es:

$$(2)\ \neg OV(\ulcorner \exists n(\iota_k(n) \wedge \phi_k)\urcorner) \leftrightarrow \psi(k)^5.$$

Intuitivamente, $\ulcorner \exists n(\iota_k(n) \wedge \phi_k)\urcorner$ dice que $\psi(n)$ es verdadera de k (su propio número de Gödel), o sea que $\psi(k)^6$.

(2) es ya equivalente a la negación de una de las oraciones de la forma $OV(x)$ syss p, mencionadas en la convención (T). Para verlo basta observar que (2) es equivalente a la negación de

$$(3)\ OV(\ulcorner \exists n(\iota_k(n) \wedge \phi_k)\urcorner) \leftrightarrow \psi(k).$$

[5] (2) no es simplemente el resultado de poner un nombre de k en lugar de las apariciones libres de 'n' en la fórmula cuantificada en (1), pues LTGC+ no tiene nombres para números particulares. En LTGC+ podemos decir que un número m tiene una cierta propiedad P diciendo que hay un conjunto con tales y cuales propiedades (las que identifican a la construcción logicista de m; todas ellas son definibles en LTGC+), y tal que también tiene P. Este recurso se debe en esencia a Russell.

[6] Se atribuye a menudo a Carnap (1934) el darse cuenta por primera vez de que el razonamiento de Gödel se aplica a cualquier predicado (y no sólo al de indemostrabilidad, al que lo aplica Gödel en (1931)). (Esta atribución la hace el propio Gödel en la impresión de 1965 de Gödel (1934)). Pero como acaba de verse, Tarski (1933) lo aplica explícitamente en su prueba a todas las negaciones de predicados de LTGC+. El resultado implícito en Tarski es casi tan general como el atribuido a Carnap. Quizá el que esto no se mencione se deba a que la parte (β) del Teorema I puede obtenerse aplicando el razonamiento diagonal al caso particular de un hipotético predicado de verdad, sin pasar por una generalización a todos los predicados. Pero no ocurre así con la parte (α), que en mi opinión puede considerarse sin mayores reparos como una versión del "lema diagonal" o "teorema del punto fijo" atribuido a menudo a Carnap.

Que (3) es equivalente a un "bicondicional T" se ve observando que $\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner$ es una oración de LTGC que es verdadera intuitivamente si y sólo si hay un valor de 'n' que es (igual a la construcción logicista de) k y que satisface la fórmula abierta ϕ_k (es decir, $\psi(n)$). Y esto es así si y sólo si k satisface ϕ_k si y sólo si k satisface $\psi(n)$ si y sólo si $\psi(k)$.

Con cambios obvios, la prueba de Tarski muestra igualmente que el Teorema I vale también si se toma como metalenguaje de LTGC a LTGC mismo y no a LTGC+.

8. Los resultados de Gödel

La prueba que Tarski da de su Teorema I emplea de forma crucial métodos creados por Gödel en su clásico trabajo (1931). En una nota histórica, Tarski nos dice:

> "Debemos el método aquí empleado a Gödel, que lo ha usado para otros propósitos en su trabajo recientemente publicado (Gödel (1931) (...)). Este trabajo sobremanera importante e interesante no está directamente conectado con el tema de nuestro trabajo —trata de problemas estrictamente metodológicos: la consistencia y la compleción de sistemas deductivos—; sin embargo podremos usar los métodos y en parte también los resultados de las investigaciones de Gödel para nuestros propósitos." (Tarski (1933), p. 247).

Obsérvese la afirmación de Tarski de que el método de Gödel había sido usado por éste "para otros propósitos" y de que su trabajo "no está directamente conectado con el tema de nuestro trabajo —trata de problemas estrictamente metodológicos: la consistencia y la compleción [sintáctica] de sistemas deductivos—". Con esto Tarski quiere quizá decir, en parte, que en el trabajo de Gödel no se demuestran resultados acerca de la definibilidad o indefinibilidad de la noción de verdad.

Específicamente, Gödel demostró en su trabajo, entre otras cosas, que para todos los sistemas deductivos S de una cierta clase (sistemas que, como se suele decir, contienen su-

ficiente aritmética), se puede construir una oración p tal que si S es consistente entonces p no es demostrable en S y si S es ω-consistente entonces $\neg p$ tampoco es demostrable en S. Este resultado crucial, y los demás del artículo, no hacen referencia a ninguna noción semántica, ni definida ni primitiva.

Sin embargo, una cosa fue el resultado formal publicado y otra el razonamiento informal que llevó a Gödel a ese resultado. Ese razonamiento informal sí contenía referencia a la noción intuitiva de verdad. El proceso psicológico de Gödel ha sido reconstruido en cierto detalle por Feferman usando datos que Gödel suministró a Wang. Gödel, después de concluir su trabajo sobre la compleción de la lógica de primer orden, abordó el problema hilbertiano de dar una prueba de la consistencia del análisis. Su idea, nos dice Wang, "fue demostrar la consistencia de la teoría de números por medio de la teoría de números finitista, y demostrar la consistencia del análisis por medio de la teoría de números" (Wang (1987), p. 42). El resto de los datos proporcionados por Gödel a Wang los reconstruye Feferman de forma que nos parece iluminadora:

> "Si uno quisiera intentar dar una prueba de consistencia relativa del análisis (en su forma de teoría de números de segundo orden) en la teoría de números de primer orden por medio de un modelo o interpretación formal, la idea obvia sería interpretar las variables de conjuntos como variando sobre los conjuntos definibles aritméticamente. De forma equivalente, podríamos numerar las fórmulas de la aritmética con una variable libre x, digamos $A_n(x)$ (n= 0, 1, 2, ...) e interpretar las variables de conjuntos como variando sobre ω, con $m \in n$ interpretado como: $A_n(m)$ es verdadera. Pero para este modelo (es decir, para interpretar $x \in y$ como una fórmula de la aritmética con dos variables libres) uno necesita la *noción general de verdad* de oraciones de la teoría de números. Esto sería problemático en vista de las paradojas clásicas. El paso clave de Gödel fue darse cuenta de que podía darse un sentido definido a la locución 'este enunciado' en la formulación de la paradoja del mentiroso, por medio de una construcción de substitución diagonal llevada a cabo en la aritmética." (Feferman (1988), pp. 105-106).

Usando su ingeniosa construcción diagonal en la formulación de un razonamiento análogo al de la paradoja del mentiroso, Gödel ofreció un argumento para mostrar que la noción *intuitiva* de verdad (o, más precisamente, el conjunto de los números de Gödel de las oraciones aritméticas *intuitivamente* verdaderas) no era definible en la aritmética. El razonamiento, que Gödel ya había formulado en el verano de 1930 (cf. Wang (1987), p. 85), está implícito en una carta de 1931 a Zermelo (cf. Grattan-Guinness (1979), pp. 300-301), y debió de ser aproximadamente el siguiente.

Supongamos que tenemos una enumeración ϕ_n de las fórmulas aritméticas con una variable libre. Supongamos que el conjunto de los números de Gödel de las oraciones intuitivamente *verdaderas* fuera definible en la aritmética, digamos que por una fórmula *Ver*. Entonces tendríamos una fórmula aritmética con una variable ('x') libre $\neg Ver(Subst(x, \ulcorner \phi_x(x) \urcorner))$, satisfecha por un número n cuando el resultado de substituir 'x' por el numeral para n en $\phi_n(x)$ no es una oración verdadera. Que de hecho esta sería una fórmula aritmética naturalmente requeriría mostrar que ϕ_n, la operación de concatenación, la operación de substitución de variables libres por numerales, etc., son definibles en la aritmética. Dado el supuesto acerca de la coextensionalidad de *Ver* con 'verdadera', el resultado de substituir 'x' por el numeral para m en $\neg Ver(Subst(x, \ulcorner \phi_x(x) \urcorner))$ (abreviado, $\neg Ver(\ulcorner \phi_m(m) \urcorner)$) es una oración verdadera si y sólo si el resultado de substituir 'x' por el numeral para m en la m-ésima fórmula (abreviado, $\phi_m(m)$) no es una oración verdadera. Ahora bien, $\neg Ver(Subst(x, \ulcorner \phi_x(x) \urcorner))$ es una de las fórmulas en la enumeración ϕ_n, digamos que $\phi_k(x)$. Entonces para todo número natural m, $\phi_k(m)$ es verdadera si y sólo si $\neg Ver(\ulcorner \phi_m(m) \urcorner)$ es verdadera. Y en particular, $\phi_k(k)$ es verdadera si y sólo si $\neg Ver(\ulcorner \phi_k(k) \urcorner)$ es verdadera ($\phi_k(k)$ "dice de sí misma que es falsa"). Pero por lo dicho antes, $\neg Ver(\ulcorner \phi_k(k) \urcorner)$ es verdadera si y sólo si $\phi_k(k)$ no es verdadera. Así, $\phi_k(k)$ es verdadera si y sólo si $\phi_k(k)$ no es verdadera, y tenemos una contradicción. Concluimos que el conjunto de los números de Gödel de las oraciones intuitivamente verdaderas no es definible en la aritmética.

La esencia de este mismo razonamiento, usando una construcción diagonal como la de Gödel (pero no la substitución de variables por numerales), es, como vimos, la empleada por Tarski en la prueba de su Teorema I. Pero nótese que Tarski no menciona en ningún momento la noción de verdad, a diferencia de lo que ocurre en la reconstrucción del argumento de Gödel. El argumento reconstruido de Gödel es un argumento acerca de la noción intuitiva de verdad. El argumento de Tarski es un argumento para establecer una aserción puramente sintáctica y muy general acerca de todos los predicados posibles de LTGC+ y acerca de la noción de demostrabilidad en LTGC+[7]. El teorema de Tarski no es un teorema acerca de la noción intuitiva de verdad. Y tampoco es un teorema acerca de una noción definida de verdad –el teorema establece precisamente (para el Tarski de 1933) que no hay tal cosa en el caso de LTGC—.

9. El *Nachwort*

En el *Nachwort* a (1933), publicado en (1935), Tarski abandona la tesis, implícita en su trabajo de 1933, de que el aparato matemático de la metateoría haya de ser formalizable en la teoría de tipos finitos[8], y acepta el uso de una teoría más potente, específicamente una teoría de tipos transfinitos

[7] Recuérdese que Gödel atribuía a Carnap (1934) la primera enunciación general del "lema diagonal". Esto proporciona confirmación adicional a la idea de que el argumento implícito en su carta a Zermelo era, como el reconstruido por nosotros, un argumento particular aplicable a un predicado aritmético bajo el supuesto de que es coextensional con el de verdad. Así, no queda claro en ese argumento que la construcción de un punto fijo para un predicado aritmético sea independiente de otros supuestos acerca de sus propiedades. Como señalamos, esta independencia queda clara ya en la prueba de Tarski.

[8] No podemos ocuparnos aquí de las razones filosóficas profundas que llevan a Tarski a dejar de postular la suficiencia para la semántica del aparato matemático de la teoría de tipos finitos (ni de las razones que le habían llevado previamente a aceptarla).

donde los objetos de tipo transfinito son clases de objetos de los tipos inferiores (Tarski no da muchos detalles sobre esta teoría). Si se acepta una metateoría lo bastante potente, la relación de satisfacción para LTGC, vista como una relación entre fórmulas y secuencias formadas por una secuencia de individuos, una secuencia de clases de individuos, etc., será una relación cuya existencia es afirmada por la teoría (y a la que se puede asignar un "tipo" u "orden" transfinito apropiado). Será posible definirla en una teoría de tipos transfinitos (o en la teoría de conjuntos) cuantificando sobre secuencias de secuencias apropiadas (una secuencia de individuos, una secuencia de clases de individuos, etc.) y sobre relaciones entre secuencias (de secuencias) y fórmulas:

> "siempre es posible construir el metalenguaje de tal manera que contenga variables de orden superior al de las variables del lenguaje estudiado. El metalenguaje se convierte entonces en un lenguaje de orden superior y por tanto en uno que es esencialmente más rico en formas gramaticales que el lenguaje que estamos investigando. (...) ahora estamos en posición de definir el concepto de verdad para cualquier lenguaje de orden finito o infinito, supuesto que tomemos como la base de nuestras investigaciones un metalenguaje que sea de al menos un orden más que el lenguaje estudiado." (Tarski (1935), pp. 271-272).

Así, en una metateoría cuyo aparato matemático sea uno lo bastante potente se podrá definir un predicado de verdad para LTGC (y otro para LTGC+) siguiendo el método de Tarski. Sin embargo, Tarski señala que las consideraciones surgidas en torno al Teorema I

> "no pierden nada de su importancia y pueden extenderse a lenguajes de cualquier orden [incluso transfinito]. Es imposible dar una definición adecuada de verdad para un lenguaje en el que se pueda construir la aritmética de los números naturales, si el orden del metalenguaje en el que se lleva a cabo la investigación no excede el orden del lenguaje investigado." (Tarski (1935), p. 272).

Tarski da una formulación algo más precisa de esta afirmación en la "Tesis B" del siguiente par de tesis:

> "A. Para todo lenguaje formalizado es posible construir una definición formalmente correcta y materialmente adecuada de oración verdadera en el metalenguaje con la ayuda únicamente de expresiones lógicas generales, de expresiones del lenguaje mismo, y de términos de la morfología del lenguaje —pero dada la condición de que el metalenguaje posea un orden superior al del lenguaje que es objeto de la investigación—.
> B. Si el orden del metalenguaje es a lo sumo igual al del lenguaje mismo, una definición tal no puede construirse." (Tarski (1935), p. 273).

Afirmaciones de este tipo han provocado en intérpretes recientes algunas confusiones que es importante tratar de aclarar. Aparecen ejemplificadas en dos trabajos de David DeVidi y Graham Solomon ((1995) y, sobre todo, (1999)). DeVidi y Solomon (1999) parten de la siguiente aseveración del texto divulgativo Tarski (1944): "...la condición de 'riqueza esencial' del metalenguaje resulta ser, no sólo necesaria, sino también suficiente para la construcción de una definición satisfactoria de la verdad" (Tarski (1944), § 10). A lo largo de esta sección llamaremos a esto 'la aseveración de Tarski'.

Partiendo de la cita de la aseveración de Tarski, DeVidi y Solomon (1999) inician una exploración que consiste en considerar varias nociones relativamente precisas de "mayor riqueza esencial" que Tarski pudo haber tenido en mente, y ver si hacen verdadera la aseveración de Tarski. En particular, consideran fundamentalmente dos posibilidades: (i) un metalenguaje ML es esencialmente más rico que un lenguaje objeto LO si ML es más fuerte que LO, en el sentido (básicamente) de demostrar todo lo que demuestra LO y algo más (consideran en particular la propiedad de que ML pruebe la consistencia de LO), y (ii) ML es esencialmente más rico que LO si ML contiene más formas de expresión que LO. Y encuentran que en ninguno de estos sentidos es verdadera la dirección de necesidad de la aseveración de Tarski (comentare-

mos en notas al pie de página sus razones para concluir esto, aunque no es esencial hacerlo para nuestro propósito principal). Concluyen que la aseveración de Tarski "es falsa, a menos que 'esencialmente más rico' no signifique más que 'suficiente para contener una definición de la verdad para el lenguaje objeto'" (DeVidi y Solomon (1999), p. 1). Según ellos, ni siquiera una disyunción de las propiedades que aparecen en (i) y (ii) proporciona más que una condición necesaria puramente trivial; consideremos "la propiedad disyuntiva de *o bien* contener expresiones bien formadas adicionales *o* ser más fuerte. Pero si esto es lo que viene a ser la 'riqueza esencial', entonces la dirección de necesidad de la aseveración de Tarski es meramente una re-enunciación de una consecuencia trivial del teorema de Tarski". (DeVidi y Solomon (1999), p. 21). Además,

> "la dirección de suficiencia de su aseveración es obviamente falsa para esta formulación de la propiedad disyuntiva requerida. Conseguir una propiedad disyuntiva que sea también suficiente requerirá una especificación mucho más detallada de los disyuntos (para hacer cada disyunto suficiente para *algunos* casos), y a la luz de esta necesidad de precisión adicional necesitaremos muchos más disyuntos, puesto que habrán de cubrir todas las varias maneras en que podemos obtener un ML suficiente para contener una definición de verdad para un LO." (DeVidi y Solomon (1999), p. 21).

Antes de continuar examinando otras afirmaciones de DeVidi y Solomon, hagamos algunas observaciones. Como señalan estos autores, las propiedades (i) y (ii) (y alguna otra sólo inesencialmente diferente que consideran DeVidi y Solomon) son *obviamente no suficientes* para la existencia de una definición de verdad para LO en ML. Pero DeVidi y Solomon parecen no darse cuenta de que esto convierte sus esfuerzos por mostrar que tampoco son necesarias en una tarea irrelevante para la cuestión de qué podía entender Tarski por "esencialmente más rico", al menos si concedemos que Tarski era capaz de ver aquella obviedad. Pues no hay que perder de

vista que Tarski afirma no sólo que la "mayor riqueza esencial" es una propiedad necesaria (en lo que pudo haberse equivocado si la implicación no era obvia), sino también que es una propiedad suficiente de la existencia de una definición de verdad —y esta implicación es obviamente falsa para las propiedades consideradas por DeVidi y Solomon—.

DeVidi y Solomon no señalan que inmediatamente después de su aseveración, Tarski explica (en la § 11 de Tarski (1944)) por qué la "mayor riqueza esencial" es una propiedad suficiente para la existencia de una definición de verdad en ML para LO. Su explicación es una exposición vulgarizada de su método de Tarski (1933) para definir la verdad por medio de la satisfacción, que explicamos nosotros antes. Tarski señala que "con este tosco esquema no está claro dónde y cómo entra en la discusión el supuesto de la 'riqueza esencial' del metalenguaje; esto resulta claro sólo cuando la construcción se lleva a cabo de una manera formal y detallada" (Tarski (1944), § 11). E inmediatamente, en una nota destinada al experto y no al lego:

> "Para definir recursivamente la noción de satisfacción, tenemos que aplicar una cierta forma de definición recursiva que no se admite en el lenguaje-objeto. Por tanto la 'riqueza esencial' del metalenguaje puede consistir simplemente en admitir este tipo de definición. Por otro lado, se conoce un método general que hace posible eliminar todas las definiciones recursivas y reemplazarlas por definiciones normales, explícitas. Si intentamos aplicar este método a la definición de satisfacción, vemos que o bien hemos de introducir en el metalenguaje variables de un tipo lógico superior al de las que aparecen en el lenguaje-objeto; o bien suponer axiomáticamente en el metalenguaje la existencia de clases que son más incluyentes [comprehensive] que todas aquellas cuya existencia puede establecerse en el lenguaje-objeto." (Tarski (1944), n. 16).

Nótese que las propiedades que menciona Tarski en esta nota son propiedades suficientes para la aplicación de *su método* para definir verdad en términos de satisfacción, al menos si el metalenguaje es siempre muy potente, como requiere

Tarski[9]. Son por tanto más específicas que (i) o (ii), y no es inmediatamente obvio que alguna de ellas o una disyunción de ellas no sea una propiedad necesaria para la existencia de una definición de verdad[10]. Aclarado qué tipo de propiedades tiene en mente Tarski, podemos preguntarnos si esto ilumina la cuestión de por qué afirma que la "mayor riqueza esencial" es una propiedad necesaria para la existencia de una definición de verdad.

DeVidi y Solomon examinan el argumento que da Tarski a favor de la dirección de necesidad de su aseveración. Este es el texto que citan:

> "Si la condición de 'riqueza esencial' no se satisface, puede usualmente mostrarse que es posible una interpretación del metalenguaje en el lenguaje objeto; es decir, con todo término dado del metalenguaje puede ponerse en correlación un término bien determinado del lenguaje objeto de tal manera que las oraciones afirmables [demostrables] de un lenguaje resultan estar correlacionadas con oraciones afirmables del otro. Como resultado de esta interpretación, la hipótesis de que se ha formulado una definición satisfactoria de la verdad en el metalenguaje implica la posibilidad de reconstruir en ese lenguaje la antinomia del mentiroso; y esto a su vez nos obliga a rechazar la hipótesis en cuestión. (...) Así vemos que la condición de 'riqueza esencial' es necesaria para la posibilidad de una definición satisfactoria de la verdad en el metalenguaje." (Tarski (1944), § 10; el texto omitido es omitido por DeVidi y Solomon).

[9] Recuérdese que Tarski requiere que los axiomas lógico-matemáticos del metalenguaje "basten para un sistema suficientemente amplio de lógica matemática" (Tarski (1933), p. 173). En Tarski (1944) dice que la parte lógica "se supone que comprende toda la teoría de las clases y las relaciones (i.e., la teoría matemática de los conjuntos)" (Tarski (1944), n. 12).

[10] Como quedará claro en notas posteriores, las razones de DeVidi y Solomon para pensar que (i) y (ii) no son propiedades necesarias de la existencia de una definición de verdad para LO en ML no muestran que una disyunción de las propiedades consideradas por Tarski (1944), n. 16, no sea necesaria para la existencia de tal definición. Pero Tarski no afirmó ni siquiera esta necesidad.

Nótese que en este texto Tarski está ofreciendo una vulgarización, sin pretensiones de detalle, de la prueba, ya de por sí condensada, que ofrece en Tarski (1933) para su Teorema I. ¿Establece el argumento de Tarski la dirección de necesidad de su aseveración, cuando entiende uno "mayor riqueza esencial" como él lo hace en la § 11 de Tarski (1944)? Para responder esto es importante mencionar el contexto en que Tarski presenta su argumento. Leamos el pasaje *inmediatamente precedente*:

> "La solución [del problema de si un predicado de verdad es definible] resulta a veces positiva, a veces negativa. Esto depende de algunas relaciones formales entre el lenguaje-objeto y su metalenguaje; o, más específicamente, del hecho de si el metalenguaje en su parte lógica es *'esencialmente más rico'* que el lenguaje objeto o no. No es fácil dar una definición precisa y general de esta noción de 'riqueza esencial'. Si nos restringimos a lenguajes basados en la teoría lógica de los tipos, la condición para que el metalenguaje sea 'esencialmente más rico' que el lenguaje objeto es que contenga variables de un tipo lógico superior al de las del lenguaje objeto." (Tarski (1944), § 10).

Es pertinente notar al menos tres aspectos de este pasaje. En primer lugar, nótese que Tarski dice que la "mayor riqueza esencial" de un metalenguaje depende exclusivamente de su "parte lógica". La razón de que diga esto es que sólo está considerando (como queda claro en la § 9 de Tarski (1944)) definiciones del predicado de verdad que no contengan otros términos o apelen a otros axiomas que los de los tres grupos de un metalenguaje tarskiano: términos y axiomas del lenguaje objeto, términos y axiomas para tratar la sintaxis del lenguaje objeto, y términos y axiomas lógico-matemáticos. Esto es parte de lo que Tarski, en su aseveración que DeVidi y Solomon consideran falsa, llama una definición *satisfactoria* de verdad (una definición de verdad es satisfactoria cuando es formalmente correcta y cumple con la convención T). Todo metalenguaje *tarskiano* contiene al lenguaje objeto y los re-

cursos para tratar la sintaxis de éste, que puede ya existir en el lenguaje objeto mismo. Por tanto, en vista del teorema de Tarski, la diferencia esencial entre el lenguaje objeto y un metalenguaje que defina verdad satisfactoriamente para aquél ha de radicar *forzosamente* en la "parte lógica" (lógico-matemática).

En segundo lugar, nótese que Tarski alude *explícitamente* a la dificultad de una definición precisa y general de la noción de "mayor riqueza esencial", y no da a entender que la posea (obsérvese que pone la expresión siempre entre comillas dobles, y que ésta es una explicación de que lo haga). Tarski no llega hasta el punto de negar que pueda haber una definición precisa y general de la noción, como hacen DeVidi y Solomon, si bien afirma cautamente que tal definición no sería fácil. Debe notarse que resulta ciertamente inexplicable que estos autores, dados sus propósitos y su conclusión final, no citen este pasaje de Tarski.

En tercer lugar, nótese que Tarski dice que si uno se *limita* a (meta)lenguajes basados en la teoría de los tipos, entonces la condición de ser "esencialmente más rico" *es* algo muy preciso, a saber, que el metalenguaje "contenga variables de un tipo lógico superior al de las del lenguaje objeto". Esta diferencia en la parte lógica de estos metalenguajes es también, como vimos, la señalada en la Tesis B de Tarski (1935). Recuérdese también que en la nota 16 de Tarski (1944) dice que "la 'riqueza esencial' del metalenguaje puede *consistir* simplemente en admitir [un] tipo de definición [recursiva]" (mi cursiva). Estos usos de 'ser' y 'consistir', junto con la afirmación anterior de que una definición general de "mayor riqueza esencial" no sería fácil, sugieren que Tarski no usa en su texto 'riqueza esencial' como nombre de una propiedad general que se da en todos los casos en que un metalenguaje tarskiano define verdad para un lenguaje objeto, sino como un término *ambiguo* que nombra propiedades diferentes en diferentes contextos de investigación. Estas propiedades, como es claro por lo que dijimos más arriba, podrán ser las propiedades usadas en el método de Tarski, mencionadas en la nota 16

de Tarski (1944), pero *nada* que dice Tarski excluye que puedan también ser otras.

El argumento de Tarski para la dirección de necesidad de su aseveración ciertamente no muestra que alguna de las propiedades de la nota 16 de Tarski (1944) o su disyunción sea una propiedad necesaria de la existencia de una definición de verdad. Ni Tarski tiene intención de mostrarlo, como queda claro en el pasaje inmediatamente anterior al argumento. Es claro, independientemente de los argumentos específicos de DeVidi y Solomon, que sería difícil aspirar a dar una prueba de ese tipo que se aplicara a todos los casos, conocidos y no conocidos, en que un metalenguaje (incluso uno tarskiano) pueda contener una definición de verdad para un lenguaje objeto. Tarski no podía prever todas las formas de ampliar la "parte lógica" del lenguaje objeto que son suficientes para dar una definición de verdad en el metalenguaje. El argumento tarskiano tiene otro propósito. Dada la indicación inmediatamente anterior de que "si nos restringimos a lenguajes basados en la teoría lógica de los tipos, la condición para que el metalenguaje sea 'esencialmente más rico' que el lenguaje objeto es que contenga variables de un tipo lógico superior al de las del lenguaje objeto", y la dificultad que ve Tarski en encontrar una noción general y precisa de "riqueza esencial", lo natural es suponer que lo que hace en su argumento, así como en el resto de las §§ 10 y 11 de Tarski (1944) es razonar acerca de *esa* noción de "riqueza esencial", la que está en juego al restringir nuestra consideración a (meta)lenguajes basados en la teoría de tipos. O, a lo sumo, razonar tomando 'riqueza esencial' como ambiguo y usando consideraciones generales que se aplican a varias circunstancias *conocidas*.

Estas conjeturas quedan confirmadas al ver que, en otros textos en que hace afirmaciones similares, Tarski nunca busca tratar todos los casos con un único enunciado o argumento. Citaremos algunos de estos textos, y al mismo tiempo veremos con algo más de detalle que en el artículo divulgativo de 1944 cuáles son los argumentos tarskianos. Consideremos en primer lugar el texto en que aparece la Tesis B de

Tarski (1935). Primero es necesario reparar en que las afirmaciones de Tarski se apoyan en ciertos supuestos acerca de la estructura de los metalenguajes más potentes que LTGC y LTGC+ para los que Tarski formula sus tesis, y en particular acerca de la naturaleza de las "expresiones lógicas generales" mencionadas explícitamente en la Tesis A (e implícitamente en la Tesis B). El supuesto principal que es pertinente aquí consiste en que Tarski restringe su atención a metalenguajes cuyo grupo lógico-matemático de términos es una extensión del tipo de términos que aparecen en el lenguaje considerado para la teoría de los tipos finitos. En particular, las variables de los nuevos metalenguajes han de pertenecer a tipos que extienden la jerarquía de los tipos finitos (o a lo sumo han de tener como recorrido una clase formada por objetos que aparecen en varios tipos de esta jerarquía; a estas últimas variables Tarski las llama 'indefinidas'). Este supuesto está claramente implícito en la descripción general de los lenguajes y metalenguajes de que se va a ocupar en el texto principal del *Nachwort*, que Tarski hace en (1935), pp. 268-272. Y se manifiesta, por ejemplo, en el hecho de que Tarski contrasta estos lenguajes de los que se ocupa en el texto principal (y a los que por tanto se aplican las tesis A y B) con metalenguajes basados en el lenguaje de la teoría de conjuntos de primer orden, al que se refiere en p. 271, n. 1.

La única manera de obtener un metalenguaje matemáticamente más potente que un lenguaje objeto dado, bajo el supuesto mencionado acerca de los términos que aparecen en la parte lógico-matemática del metalenguaje, es introducir en este último variables para tipos superiores a los de las variables del lenguaje objeto, con la consiguiente ampliación del sistema deductivo para dar principios, en particular de comprensión, que gobiernen el uso de las nuevas variables. Si el metalenguaje no es "esencialmente más rico" que el lenguaje objeto en este sentido, entonces, puesto que el metalenguaje es muy rico, el lenguaje objeto *también* lo será, y será posible interpretar en él la parte del metalenguaje necesaria para desarrollar el argumento de la prueba del Teorema I (o el esbo-

zo de Tarski (1944)). Esta es la razón, relativamente simple, por la que Tarski afirma su Tesis B[11].

En varios textos de la época Tarski distingue el caso que se da al adoptar metalenguajes basados en la teoría de los tipos del caso en que se adopta como aparato lógico-matemático del metalenguaje la teoría de conjuntos. En uno de estos, no por coincidencia, señala que afirmaciones análogas a su Tesis B valen, con los cambios apropiados, para metalenguajes basados en la teoría de conjuntos en los que se puede usar el método tarskiano para definir verdad para el lenguaje objeto. El texto es el siguiente:

> "Supongamos ahora que el sistema **L** ha sido ampliado para formar un sistema $\mathbf{L}_1$ en el cual la definición de la clase *OV* (para oraciones del sistema original **L**) puede formalizarse. Si

[11] Para refutar que la propiedad (ii) mencionada en el texto sea una condición necesaria de la existencia de una definición de verdad para el lenguaje objeto en un metalenguaje, DeVidi y Solomon observan que es posible "formular una definición de verdad para Z en ZF [la teoría de conjuntos de Zermelo-Fraenkel] (...). En este caso, nótese, no necesitamos cuantificar sobre variables de ningún tipo superior" (DeVidi y Solomon (1995), p. 133; véase también DeVidi y Solomon (1999), p. 20). Con lo de que es posible formular una definición de verdad para Z en ZF se refieren a que es posible definir en ZF un predicado '*OV**' de (números de Gödel de) oraciones de Z (y ZF) del que se puede probar en ZF que verifica una convención T. En concreto, es posible demostrar en ZF todos los bicondicionales de la forma de '*OV**(*x*) si y sólo si *p*', con '*x*' substituido (mediante el truco de Russell) por un nombre de (un número de Gödel de) una oración de Z (y ZF) y '*p*' substituido por la relativización de esa oración a un cierto conjunto fijado cuya existencia se sigue de ZF, y donde valen todos los axiomas de Z. Esto indica que, si se toma como interpretación del lenguaje de Z a ese conjunto, entonces el conjunto de (números de Gödel de) oraciones verdaderas es definible en ZF. La preocupación de DeVidi y Solomon es que el lenguaje de ZF es el mismo que el de Z, y por tanto contiene las mismas formas de expresión y es del mismo orden que el de Z. Podría pensarse, pues, que este caso proporciona un contraejemplo a la Tesis B de Tarski, pero la realidad es que la Tesis B está claramente formulada exclusivamente para metalenguajes tarskianos cuya "parte lógica" amplía la del lenguaje objeto por contener variables de orden superior (y axiomas obvios, en particular de comprensión, para manejarlas), no por contener nuevos axiomas acerca de las variables y primitivos existentes.

> el sistema **L** está basado en la teoría de los tipos esta ampliación consistirá, en primer lugar, en un enriquecimiento de sus formas lingüísticas mediante la introducción de variables de un nivel superior (...). Los postulados y las reglas de inferencia han de adaptarse a los recursos ampliados de las formas lingüísticas, pero su contenido, aproximadamente hablando, no cambia. Si el sistema **L** es construido siguiendo el ejemplo del sistema de Zermelo, ni las formas lingüísticas ni las reglas de inferencia cambian, sino que el cambio consiste enteramente en el reforzamiento del conjunto de postulados." (Tarski (1939), p. 110).

En casos en que el metalenguaje se puede obtener *exclusivamente* ampliando la parte lógica del lenguaje objeto con axiomas conjuntistas que bastan para dar una definición tarskiana de verdad para el lenguaje objeto, un análogo de la Tesis B, que podemos llamar B', sería el siguiente (aunque Tarski no lo enuncia explícitamente): si el metalenguaje es a lo sumo igual de fuerte que el lenguaje objeto en su parte lógica, entonces no contiene una definición de verdad satisfactoria para el lenguaje objeto. El argumento es como antes. Como ya vimos, es un principio general aplicable a todos los metalenguajes tarskianos que su diferencia esencial con respecto al lenguaje objeto ha de radicar en la parte lógica. Si no hay tal diferencia, entonces, puesto que el metalenguaje conjuntista es muy rico matemáticamente, el lenguaje objeto también lo será, y será posible interpretar en él la parte del metalenguaje necesaria para desarrollar el argumento de la prueba del teorema de Tarski, con lo que quedará establecida la Tesis B'[12].

[12] Para refutar que la propiedad (i) mencionada en el texto sea una condición necesaria de la existencia de una definición de verdad para el lenguaje objeto en un metalenguaje, DeVidi y Solomon traen a colación una serie de resultados de Wang (1952), uno de los cuales es esencialmente el siguiente (véase el útil resumen de estos resultados en Montague (1957)). Consideremos la sintaxis de primer orden L con '$\in$' como único primitivo no lógico. Podemos extender una teoría T enunciada en L a ciertas teorías enunciadas en un lenguaje en que se añaden variables de segundo orden a L, formando extensiones tanto predicativas como impredicativas de T. La extensión predicativa más débil de T consistirá meramente en añadir unos

Concluimos aquí nuestra discusión de los problemas suscitados por la Tesis B de Tarski (1935). En resumen, podemos decir que Tarski no pretendió dar una formulación general de la Tesis B por medio de una supuesta noción general de "riqueza esencial". Usó este término en su artículo divulgativo Tarski (1944) para abreviar de forma ambigua ciertas propiedades conocidas de las que él podía mostrar que eran necesarias y suficientes para la existencia de una definición de verdad para un lenguaje objeto en un metalenguaje tarskiano para ese lenguaje objeto, dados ciertos supuestos acerca de la parte lógica del metalenguaje[13].

---

pocos axiomas que enuncian propiedades básicas de las nuevas variables y, en particular, su esencia es que contendrá axiomas de comprensión para clases cuyas condiciones para la definición de clases no podrán contener variables de segundo orden ligadas. La extensión impredicativa más débil de *T* consistirá en añadir unos pocos axiomas similares que enuncian propiedades básicas de las nuevas variables y, en particular, su esencia es que contendrá axiomas de comprensión para clases cuyas condiciones para la definición de clases podrán consistir en fórmulas cualesquiera. El resultado fundamental de Wang es que si *T* es una teoría enunciada en *L* que incluye la lógica de primer orden, el axioma de conjunto vacío y de par desordenado, entonces las extensiones predicativas de *T* satisfarán una convención T para *L*, y por tanto contendrán definiciones de verdad para *L*. Una consecuencia del resultado de Wang es, por ejemplo, el resultado conocido antes de Wang de que la teoría de conjuntos de Gödel-Bernays (GB) contiene una definición de verdad (que se construye por métodos tarskianos) para el lenguaje *L* de Z (la teoría de Zermelo), de la cual es una extensión predicativa. Sin embargo, como señalan DeVidi y Solomon, GB no es más fuerte que Z en ningún sentido interesante, y en particular es una extensión conservadora de Z. De hecho, el resultado de Wang implica que teorías claramente más débiles que Z contienen definiciones de verdad para *L*. Según DeVidi y Solomon, esto muestra que el ser más débil un lenguaje que otro no implica que el primero no pueda demostrar todos los bicondicionales T del segundo. Podría pensarse que hay aquí un contraejemplo a la Tesis B' de Tarski, pero en realidad la Tesis B' está formulada sólo para metalenguajes tarskianos cuya "parte lógica" extiende la del lenguaje objeto por contener nuevos axiomas acerca de las variables y primitivos existentes, no por contener variables de orden superior (y axiomas de comprensión para manejarlas).

[13] Una última nota sobre este asunto. DeVidi y Solomon (1999), p. 22, observan que del resultado de Wang (1952) mencionado en la nota an-

El resultado quizá más importante del *Nachwort* es sólo una consecuencia implícita pero inmediata de lo que en él se dice. Es el resultado de que con ayuda del predicado '*OV*' definido en un lenguaje de tipos transfinitos, por tanto de orden superior al de LTGC, es posible formalizar los argumentos informales que vimos en las secciones 3 y 6. Podemos así demostrar de forma matemáticamente satisfactoria que en LTGC+ no es posible definir un predicado coextensional con el predicado definido de verdad '*OV*' para LTGC. Los argumentos informales pasan a ser tan legítimos matemáticamente como lo es el argumento por el que se demuestra el Teorema I. Los argumentos legitimados son similares a los que se usan hoy en día para demostrar versiones habituales, semánticas, del teorema de la indefinibilidad de la verdad.

También en el *Nachwort*, Tarski ofrece una indicación de cómo formalizar de una manera relativamente simple la esencia del razonamiento de Gödel (1931) para la demostración de la existencia de oraciones indecidibles, usando un predicado definido de verdad. En este razonamiento se construye una oración indecidible siguiendo la idea de Gödel de obtener

---

terior se sigue que ciertas teorías considerablemente más débiles que Z (por ejemplo, extensiones predicativas de la teoría enunciada con el símbolo '$\in$' como único primitivo no lógico que incluye meramente la lógica de primer orden, el axioma de conjunto vacío y el de par desordenado) contienen una definición de verdad para Z. Al mismo tiempo, es fácil ver que Z contiene una definición de verdad para ellas. Así tenemos que hay una teoría *T* tal que *T* define verdad para Z y Z para *T*. DeVidi y Solomon infieren de esto que la idea de Tarski de que su teorema da pie a una "jerarquía de (meta)lenguajes" es confusa, pues *T* y Z no están relacionadas entre sí como lo estarían en una jerarquía. La realidad es que el que Tarski hable ocasionalmente de jerarquías de lenguajes (por ejemplo en Tarski (1944), § 9) se debe al hecho ya subrayado de que requiere que sus metalenguajes contengan al lenguaje objeto y a su teoría deductiva. Dado este supuesto, la única manera de que un metalenguaje tarskiano contenga una definición de verdad para un lenguaje objeto suficientemente potente ha de consistir en extender la "parte lógica" del lenguaje objeto. *T* no extiende la teoría deductiva de Z (aunque tenga variables de orden superior a las de Z); Z no extiende el conjunto de primitivos de *T* (aunque tenga axiomas más fuertes que los de T). Por tanto caen fuera de las consideraciones de Tarski.

una oración que "diga de sí misma" que no es demostrable. Pero la prueba no requiere todo el trabajo dentro del sistema formal que es necesario para demostrar la versión original del teorema de Gödel, en la que no se menciona la noción de verdad. La razón es que la prueba de Tarski usa el hecho, enunciable por medio del predicado de verdad, de que el sistema formal sólo demuestra oraciones verdaderas. El razonamiento de Tarski es en esencia el mismo que vimos que dará para un predicado cualquiera $E(x)$ en Tarski (1939), con $E(x)$ substituido por el predicado de demostrabilidad.

Las indicaciones de Tarski son, simplificando, las siguientes. Consideremos la demostración de la parte (α) del Teorema I. Sabemos que en LTGC+ hay un predicado $Dem(x)$ que define la clase de las oraciones demostrables en el cálculo de LTGC. Siguiendo los mismos pasos que en aquella prueba, pero poniendo '*Dem*' en lugar de '*OV*', podemos demostrar en LTGC+ un bicondicional

$$(1)\ \neg Dem(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner) \leftrightarrow \psi(k).$$

Intuitivamente, $\psi(k)$ dice de sí misma (es decir, de $\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner$) que no es demostrable. Entonces Tarski muestra, en una metateoría (más fuerte que LTGC+) donde definimos un predicado de verdad *OV* para oraciones de LTGC, que $\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner$ es indecidible pero *OV* se aplica a $\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner$: de la metateoría es consecuencia (puesto que *OV* satisface la convención T) el bicondicional

$$(2)\ OV(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner) \leftrightarrow \psi(k).$$

Por tanto, de (1) y (2) se sigue que

$$(3)\ \neg Dem(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner) \leftrightarrow OV(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner).$$

Por otro lado, de la definición de *OV* se siguen las siguientes consecuencias (siempre con base en el aparato matemático de la metateoría):

(4) $\neg OV(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$ o $\neg OV(\ulcorner \neg \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$.
(5) si $Dem(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$, entonces $OV(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$.
(6) si $Dem(\ulcorner \neg \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$, entonces $OV(\ulcorner \neg \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$.

De (3) y (5) se siguen inmediatamente:

(7) $OV(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$.
(8) $\neg Dem(\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$.

De (4) y (7) junto con (6) se sigue que

(9) $\neg Dem(\ulcorner \neg \exists n(\iota_k(n) \wedge \phi_k) \urcorner)$.

(7) establece que $\ulcorner \exists n(\iota_k(n) \wedge \phi_k) \urcorner$ es *OV*. (8) y (9) establecen que es indecidible en el cálculo de LTGC.

10. La relación entre los resultados de Gödel y los de Tarski

Gödel no dio a la imprenta el razonamiento informal que vimos en la sección 8 hasta la aparición en 1965 de una versión de sus conferencias de Princeton de 1934, donde ya aparecía un argumento similar (cf. Gödel (1934), § 7). De todos modos, las notas mimeografiadas de las conferencias tuvieron una amplia difusión, con la autorización de Gödel, ya desde 1934, y creo que pueden considerarse una publicación en el sentido pleno de la palabra. Pero un argumento análogo al de las conferencias está implícito también, como vimos, en una carta de 1931 a Zermelo. Esto sugiere preguntas interesantes que Feferman ha hecho e intentado responder con gran penetración. En particular, es interesante preguntarse por qué no dio Gödel en su artículo clásico de 1931 el argumento informal que vimos en la sección 8, del que se sigue de forma relativamente sencilla la existencia de oraciones indecidibles, pues es relativamente fácil mostrar que el conjunto de números de Gödel de las oraciones *demostrables* —en un

sistema donde se pueda desarrollar suficiente aritmética— es definible aritméticamente, y así, dado que todas las oraciones demostrables serán verdaderas, no puede coincidir extensionalmente con el conjunto de números de Gödel de oraciones verdaderas. O por qué no dio el argumento que acabamos de ver que formalizó Tarski en el *Nachwort*, por el que se construye una oración indecidible, y que Gödel podría haber enunciado y demostrado usando la noción intuitiva de verdad, de manera mucho más simple que en su trabajo de 1931.

La respuesta de Feferman me parece convincente, y consiste en atribuir a Gödel una *cautela* extrema. Gödel era muy consciente de que la noción intuitiva de verdad (aritmética) que aparecía en su resultado era una noción hacia la que los matemáticos y aun los filósofos de los círculos que le eran conocidos (los positivistas lógicos) tenían una extremada desconfianza (la misma actitud, como vimos en la sección 2, fue documentada por el propio Tarski en sus trabajos sobre semántica; cf. también Tarski (1933), p. 252). Dando resultados acerca de esa noción Gödel corría el riesgo de ser objeto de desdén intelectual por parte de los miembros de los círculos que más interesados podían estar en sus resultados (y los círculos cuya opinión contaba más para Gödel). Wang y Feferman han citado una carta escrita (pero no enviada) por Gödel a un estudiante en 1970, en la que Gödel decía, refiriéndose a la época en que publicó su (1931): "el concepto de verdad matemática objetiva en cuanto opuesto al de demostrabilidad era contemplado con la mayor sospecha y considerado en amplios círculos como carente de sentido" (Wang (1987), p. 85). Feferman concluye que Gödel no osó publicar el resultado informal en 1931 y se esforzó por obtener de él resultados enunciables haciendo referencia únicamente a nociones aceptadas, como demostrabilidad y consistencia[14]. Sólo unos pocos años más tarde, con la percepción de que el clima intelectual no era qui-

[14] Sin duda se puede decir que esto fue una buena cosa, pues llevó a Gödel a la introducción de nociones y resultados de gran importancia para los sistemas formales de la aritmética.

zá enteramente adverso, Gödel consentiría en la difusión de un texto (el de las conferencias de Princeton) donde se hacían consideraciones acerca de la noción intuitiva de verdad.

Una explicación similar a la de Feferman ha sido suscrita por otros autores, como J. Wolenski (1993) y R. Murawski (1998), que la han usado, sin embargo, para sostener posiciones algo distintas a la de Feferman en lo relativo a la cuestión de la prioridad (de Gödel o Tarski) en la demostración del teorema de la indefinibilidad de la verdad. Feferman (véase su (1988), p. 108), junto con Wang (cf. Wang (1987), p. 90) y otros autores, afirma que el teorema no debería atribuirse a Tarski, sino a Gödel, en virtud de que aparece ya anunciado por Gödel en la carta a Zermelo de 1931 que ya hemos mencionado. Wolenski y Murawski sostienen posturas algo distintas.

Murawski acepta, con Feferman, que Gödel se anticipó a Tarski en el descubrimiento de la indefinibilidad de la verdad en los lenguajes que contienen suficiente aritmética. Pero llama la atención, al igual que Feferman, sobre la aparente preocupación de Gödel por que su resultado intuitivo sobre la noción de verdad no fuera aceptado y quizá fuera denigrado. Y pone énfasis en que Tarski no tenía esas preocupaciones, lo cual le permitió adelantarse en la publicación del resultado:

> "Gödel no mencionó la indefinibilidad de la verdad en sus escritos, incluso evitó los términos 'verdad' y 'verdadero', porque temía que un trabajo que diera por sentado ese concepto sería rechazado por el *establishment* fundacional dominado por las ideas de Hilbert. Tarski carecía de tales limitaciones. De hecho, en la Escuela de Lvov-Varsovia no se daban por supuestas ningunas precondiciones iniciales restrictivas antes de que las investigaciones apropiadas pudieran comenzar." (Murawski (1998), p. 159).

Wolenski va más lejos que Murawski, y su propuesta es más favorable a Tarski. Según Wolenski,

> "el concepto de verdad [para la aritmética elemental] y el concepto de ser aritmético (aritméticamente definible) [en ella]

> son ingredientes cruciales del teorema de la indefinibilidad. Desde luego, ambos deberían estar bien definidos para servir de base conceptual para el teorema de la indefinibilidad." (Wolenski (1993), p. 136).

Pero, también según Wolenski, no está claro que Gödel haya formulado siquiera el mismo teorema que Tarski, pues a diferencia de éste, Gödel tal vez no disponía de una noción "bien definida", matemáticamente aceptable de verdad. Quizá Gödel pensaba que la noción de verdad era "inefable", que no podía aparecer en razonamientos matemáticos correctos:

> "mi tesis principal no tiene que ver con la cuestión de si Gödel ha definido la verdad o no, sino con la posibilidad de que su opinión fuese que es matemáticamente 'inefable', para usar la palabra de Hintikka. Si esta tesis es plausible, entonces [Gödel] ni enunció ni demostró el teorema de la indefinibilidad de la verdad, al menos en 1931." (Wolenski (1993), pp. 137-138).

Creo que las observaciones de Wolenski y Murawski son importantes, pero me parecen en último término erróneas. Considérese primero la afirmación de Wolenski de que la posibilidad de que Gödel creyera que la noción de verdad no es matemáticamente aceptable implicaría que no enunció ni probó el teorema de la indefinibilidad. Esto me parece falso, pues la *opinión* de un autor acerca de los conceptos que usa en sus investigaciones no influye en la cuestión de si sus investigaciones contienen esos conceptos; si los contienen y son correctas, entonces el autor es autor de investigaciones sobre esos conceptos, independientemente del valor que él mismo atribuya a sus investigaciones[15].

La afirmación de Wolenski de que el concepto de verdad y el concepto de ser aritmético (aritméticamente definible)

[15] Además, si hemos de creer a Gödel, no tenía éste en 1931 una actitud peyorativa acerca de la noción de verdad matemática, sino todo lo contrario, como se comprueba por sus afirmaciones en varias cartas citadas por Wang.

son ingredientes cruciales del teorema de la indefinibilidad es claramente falsa si aceptamos que el Teorema I de Tarski (1933) es una versión del teorema de la indefinibilidad de la verdad. Pero también es problemática su afirmación de que un concepto de verdad debería estar bien definido para servir de base conceptual para el teorema de la indefinibilidad, incluso si tenemos en mente una versión del teorema en que aparezca *un* concepto de verdad (a diferencia de la de Tarski). El problema es que, irónicamente, el concepto de "estar bien definido" es muy vago y no bien definido. No está muy claro qué haría falta para que un concepto de verdad estuviera bien definido en la formulación de un teorema. Ciertamente, una opción que no está abierta para quien quiera apoyar una atribución de prioridad a Tarski es la suposición de que un concepto de verdad está bien definido cuando está definido dentro de la matemática aceptada. Pues incluso bajo ese supuesto parecería absurdo atribuir a Tarski la versión del teorema de la existencia de oraciones indecidibles que vimos que formalizaba en el *Nachwort*, o el teorema de compleción de la lógica de primer orden (que es acerca de la noción de verdad en una estructura), meramente porque él haya sido el primero en enunciarlos sin tacha matemática. Sin embargo la situación con la versión semántica del teorema de la indefinibilidad de la verdad es análoga. (Igualmente parecería absurdo atribuir a Turing o Church el "teorema" de que la función de adición es computable).

Las observaciones de Murawski son también inexactas, pues Tarski no debe a su supuesta carencia de "precondiciones iniciales restrictivas" el haber podido enunciar y demostrar su Teorema I. El teorema, como vimos, no hace uso de ninguna noción semántica, ni primitiva o intuitiva, ni definida. Más aún, e irónicamente, es claro que cierto tipo de "precondiciones iniciales restrictivas" *influyeron* en el hecho de que formulara su teorema como lo formuló. En particular, influyó en ello su rechazo pre-teórico de (1933) a considerar metalenguajes más potentes que el de la teoría general de las clases o, de manera equivalente, la teoría simple de los tipos finitos.

También una afirmación de Feferman es inexacta, por lo tanto. Después de explicar que Gödel había obtenido ya en 1931 su resultado acerca de la noción intuitiva de verdad, afirma: "Por otro lado, no enunció esto como resultado (sólo lo hizo después Tarski, e independientemente, en 1933), y se tomó un gran trabajo en eliminar el concepto de verdad de los resultados principales de 1931" (Feferman (1988), p. 106). La afirmación de Feferman ha de basarse en la idea de que el (único) teorema demostrado por Tarski es la versión del teorema intuitivo de Gödel con la noción de verdad intuitiva substituida por la definida, pues el disponer de esa noción definida es lo que permite a Tarski "no tomarse trabajo" en eliminar el concepto de verdad de sus resultados. Pero esta idea es inexacta, como hemos visto, pues Tarski ni usó ni podía usar ningún concepto de verdad, ni primitivo ni definido, en el enunciado de su teorema. El teorema demostrado por Tarski habría sido perfectamente aceptable para el cauto Gödel, *incluso* si éste hubiera sido cauto también acerca del uso de una noción *definida* de verdad.

Todas estas consideraciones sugieren una evaluación más compleja de los logros respectivos de Gödel y Tarski que las que nos son conocidas. Todos los intérpretes mencionados han pasado por alto el hecho, subrayado aquí, de que el teorema demostrado por Tarski en 1933 no contiene referencia a ningún predicado de verdad. Y también han pasado por alto el hecho de que, a pesar de todo, su primera parte sugiere inmediatamente una versión intuitiva de la tesis de que un predicado coextensional con el intuitivo de verdad para LTGC es indefinible en LTGC y LTGC+[16]. En este sentido, el resultado de Tarski está a la par de los resultados publicados por Gödel en (1931): son formulaciones matemáticamente satisfactorias, en un sentido muy estricto de 'matemáticamente satisfacto-

[16] También parece haberse pasado por alto que la primera parte del Teorema I de Tarski es casi tan completamente general como el "lema diagonal" o "teorema del punto fijo", lo que señalamos anteriormente.

rias', no semánticas en ningún sentido, de hechos obtenidos usando intuiciones acerca de conceptos no matemáticos. Además, su significación se deriva forzosamente de la consideración intuitiva de esos conceptos: la incompleción de Gödel es significativa porque muestra cómo construir una oración *verdadera* que sin embargo es indemostrable; la parte (α) del teorema de Tarski es significativa porque sugiere inmediatamente que ningún predicado de LTGC+ es coextensional con el de *verdad*.

La razón última de que Tarski demostrara su Teorema I es que, como hemos visto, en 1933 él tenía *también*, como Gödel, razones filosóficas para desconfiar de la posibilidad de un concepto matemáticamente satisfactorio de verdad para oraciones de LTGC. Por tanto, el logro de Tarski adquiere especial relevancia si consideramos que Gödel no halló una forma matemáticamente satisfactoria de formular el hecho intuitivo más simple que subyace en sus resultados, a pesar de que por un tiempo compartió con Tarski la idea de que un predicado de verdad para LTGC o lenguajes similares era matemáticamente problemático (aunque por razones diferentes de las de Tarski). El logro de Tarski se basó, en una medida considerable, en su profundo análisis del concepto de verdad y de las condiciones suficientes para la posibilidad de definir predicados de verdad. Ese análisis es directamente responsable de su idea de formular el Teorema I en los términos en que está formulado.

Creo, pues, que la postura razonable es atribuir a Gödel la prioridad en el resultado intuitivo y a Tarski la prioridad en la formulación de un resultado matemático que expresa satisfactoriamente ese resultado intuitivo sin usar en la formulación concepto alguno de verdad. (El hecho de que Tarski usara métodos de Gödel (pero no el resultado intuitivo gödeliano, que desconocía) no afecta, naturalmente, a la cuestión de la prioridad.) Es muy apropiado llamar 'teorema de Tarski' a un resultado no semántico como el demostrado por Tarski, y esa es la práctica común en muchos manuales (por ejemplo, Lindström (1997), p. 17, Robbin (1969), p. 119,

Smullyan (1992), p. 104))[17]. Un resultado de este tipo puede atribuirse sin escrúpulos y con el respeto debido a Tarski. Ciertamente, decir que "el" teorema debe atribuirse a Gödel y no a Tarski (o viceversa) es simplificar indebidamente las cosas.

INSTITUTO DE INVESTIGACIONES FILOSÓFICAS, UNAM;
México, DF 04510, México
mariogt@servidor.unam.mx

BIBLIOGRAFIA

Carnap, R. (1934), *Logische Syntax der Sprache*, Julius Springer, Viena.

DeVidi, D. y G. Solomon (1995), "Tolerance and Metalanguages in Carnap's *Logical Syntax of Language*", *Synthese*, vol. 103, pp. 123-139.

—— (1999), "Tarski on 'Essentially Richer' Metalanguages", *Journal of Philosophical Logic*, vol. 28, pp. 1-28.

Feferman, S. (1988), "Kurt Gödel: Conviction and Caution", en S. Shanker (comp.), *Gödel's Theorem in Focus*, Croom Helm, Nueva York, pp. 96-114.

Gödel, K. (1931), "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", *Monatshefte für Mathematik und Physik*, vol. 38, pp. 173-198; también en Gödel (1986), pp. 144-194. Traducido al inglés como "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I", en Gödel (1986), pp. 145-195. (Referencias a la reimpresión y la traducción).

—— (1934), "On Undecidable Propositions of Formal Mathematical Systems", notas mimeografiadas preparadas por S. C.

[17] El resultado que estos manuales llaman 'teorema de Tarski' es la parte (β) del teorema original. Si estoy en lo cierto al subrayar la importancia de la parte (α), como he venido haciendo, entonces aplicar el nombre exclusivamente a la parte (β) oscurece uno de los méritos de Tarski.

Kleene y J. B. Rosser. Versión revisada y ampliada impresa en M. Davis (comp.), *The Undecidable*, Raven Press, Hewlett (Nueva York), 1965, pp. 39-74, y en Gödel (1986), pp. 346-371.

—— (1986), *Collected Works*, vol. I, Oxford University Press, Nueva York.

Grattan-Guinness, I. (1979), "In Memoriam Kurt Gödel: his 1931 Correspondence with Zermelo on his Incompletability Theorem", *Historia Mathematica*, vol. 6, pp. 294-304.

Lindström, P. (1997), *Aspects of Incompleteness*, Springer, Berlín.

Montague, R. (1957), Reseña de Wang (1952), *Journal of Symbolic Logic*, vol. 22, pp. 365-367.

Murawski, R. (1998), "Undefinability of Truth. The Problem of Priority: Tarski vs Gödel", *History and Philosophy of Logic*, vol. 19, pp. 153-160.

Robbin, J. W. (1969), *Mathematical Logic. A First Course*, Benjamin, Nueva York.

Smullyan, R. (1992), *Gödel's Incompleteness Theorems*, Oxford University Press, Nueva York.

—— (1994), *Diagonalization and Self-reference*, Oxford University Press, Nueva York.

Tarski, A. (1933), *Pojecie prawdy w jezykach nauk dedukcyjnych* (Sobre el concepto de verdad en lenguajes de las ciencias deductivas), *Travaux de la société des sciences et des lettres de Varsovie, Classe III*, no. 34, Varsovia. Traducido al inglés como parte de "The Concept of Truth in Formalized Languages", en Tarski (1956), pp. 152-267. (Referencias a la traducción inglesa).

—— (1935), "Der Wahrheitsbegriff in den formalisierten Sprachen", *Studia Philosophica*, vol. 1, pp. 261-405. Una traducción al alemán de Tarski (1933), con un nuevo "Nachwort" añadido. Traducido al inglés como "The Concept of Truth in Formalized Languages", en Tarski (1956), pp. 152-278. (Referencias a la traducción inglesa).

—— (1939), "On Undecidable Statements in Enlarged Systems of Logic and the Concept of Truth", *Journal of Symbolic Logic*, vol. 4, pp. 105-112.

—— (1944), "The Semantic Conception of Truth and the Foundations of Semantics", *Philosophy and Phenomenological Research*, vol. 4, pp. 341-376.

—— (1956), *Logic, Semantics, Metamathematics*, Oxford University Press, Oxford. 2da. ed.: Hackett, Indianapolis, 1983.

—— A. Mostowski y R. Robinson (1953), *Undecidable Theories*, North-Holland, Amsterdam.

Wang, H. (1952), "Truth Definitions and Consistency Proofs", *Transactions of the American Mathematical Society*, vol. 73, pp. 243-275.

—— (1987), *Reflections on Kurt Gödel*, M.I.T. Press, Cambridge (Massachusetts).

Wolenski, J. (1993), "Gödel, Tarski and the Undefinability of Truth", *Yearbook 1991 of the Kurt Gödel Society*, pp. 97-108. Reimpreso en Wolenski, *Essays in the History of Logic and Logical Philosophy*, Jagiellonian University Press, Cracovia, 1999 (Referencias a la reimpresión).

Abstract

This is the second part of a two-part paper devoted to a study of some historical, logical and philosophical issues arising from a reading of Tarski's celebrated "Wahrheitsbegriff" monograph. This second part concentrates on issues related to Tarski's "Wahrheitsbegriff" version of the theorem on the indefinability of truth. One of these issues is the correct exegesis of Tarski's claim that a truth definition cannot be constructed in a metalanguage if its order is not higher than that of the object language. Another issue is the correct contrast between Tarski's mathematical achievement in offering his version of the indefinability result and Gödel's achievement in his earlier discovery of another version of the result. The correct contrast must emphasize the fact that (contrary to frequent claims) Tarski's version does not use semantic notions (either defined or intuitive); no way towards a version of the indefinability theorem not employing intuitive semantic notions appears in Gödel, despite the fact that he, like Tarski, sought to find non-semantic versions of his results.