El concepto de verdad en la tradición analítica. Una mirada a las teorías principales

MARIO GÓMEZ TORRENTE

Instituto de Investigaciones Filosóficas, UNAM; Ciudad de México 04510, México
mariogt@unam.mx

Resumen: Se exponen en primer lugar algunas consideraciones introductorias sobre el concepto ordinario de verdad, los portadores de la propiedad de la verdad, y la paradoja del mentiroso. A continuación, se describen someramente algunas teorías tradicionales sobre la verdad, como la coherentista, la pragmatista, y algunas variantes de la correspondentista, estableciendo la distinción entre teorías correspondentistas sustantivistas e insustantivistas. El resto del artículo está dedicado a presentar de manera crítica las que quizá sean las dos teorías más importantes de la verdad en la tradición analítica, las teorías insustantivistas de Alfred Tarski (1933) y Saul Kripke (1975). En el caso de la teoría de Tarski, ésta se introduce presentando una definición tarskiana de un predicado de verdad para un lenguaje formal específico relativamente simple. Seguidamente se exponen las razones por las que esta definición es extensionalmente correcta para el lenguaje para el que se define, y cómo evade paradojas como la del mentiroso, en virtud del hecho de que el predicado definido no lo está para oraciones del metalenguaje en que se han dado las definiciones, de manera que el lenguaje al que se aplican no tiene su propio predicado de verdad. También se presentan críticas a la teoría tarskiana derivadas sobre todo de este último hecho. Nuestra presentación de la teoría de Kripke enfatiza este tipo de limitaciones, algunas de las cuales son superadas por esta teoría. La teoría de Kripke se presenta exponiendo su construcción de una interpretación consistente de un lenguaje que contiene su propio predicado de verdad, en la que este predicado no está definido para oraciones "no fundadas", incluidas oraciones como las que dan pie a las distintas versiones de la paradoja del mentiroso. Finalmente se describe una aparente limitación de la teoría kripkeana relacionada con la llamada "paradoja del mentiroso reforzado".

1. Preliminares sobre el concepto de verdad

¿Qué es la verdad? Hay una larga tradición de respuestas a esta pregunta filosófica. Aquí describiremos brevemente algunas de las respuestas ofrecidas dentro de la tradición analítica, con especial atención a dos de ellas, la ofrecida por la llamada "teoría semántica de la verdad" de Alfred Tarski y la ofrecida por la teoría de Saul Kripke. Pero antes distinguiremos la pregunta (y los tipos de respuestas que admite) de otras preguntas relacionadas con ella, y trataremos algunas otras cuestiones preliminares.

No es lo mismo preguntar qué es la verdad que preguntar cuál es la verdad (acerca de una determinada cuestión). Si preguntamos cuál es la verdad acerca de la muerte de Luis Donaldo Colosio no estamos haciendo una pregunta filosófica. Estamos preguntando, quizá, si la verdad es que lo mató un asesino solitario o que fue la víctima de una conspiración política. Si preguntamos qué es la verdad, en cambio, estamos preguntando cómo caracterizar la noción de verdad, cómo explicar esa noción de tal manera que el

concepto de verdad quede delimitado, y esta es una pregunta típicamente filosófica. Además, saber qué caracteriza a la verdad no necesariamente nos permitirá averiguar cuál es la verdad acerca de una determinada cuestión: como mencionaremos luego, algunos filósofos han propuesto que lo que caracteriza a la verdad es la correspondencia con los hechos; pero aceptar esta caracterización no nos ayuda mucho a saber cuál es la verdad acerca de la muerte de Luis Donaldo Colosio—para ello tendríamos que saber cuál de las hipótesis mencionadas, o de otras, se corresponde con los hechos.

No es lo mismo preguntar qué es la verdad que preguntar de qué se dice la verdad—o su contrario, la falsedad. Al preguntar de qué se dice la verdad y la falsedad estamos haciendo una pregunta filosófica, ciertamente. Pero no es la pregunta acerca de cómo caracterizar la noción de verdad, sino una pregunta preliminar a la tarea de caracterizar la noción de verdad. Al preguntar de qué se dice la verdad y la falsedad preguntamos qué tipo de cosas son susceptibles de ser verdades y falsedades.

Hay muchos tipos de cosas de las que decimos que son verdades o falsedades, verdaderas o falsas: ideas, creencias, conjeturas, afirmaciones, proposiciones, oraciones, etc. Este último tipo de cosas, las oraciones, han recibido una atención especial por parte de los filósofos analíticos que han estudiado la noción de verdad. Es importante hablar brevemente del sentido en que se dice de una oración que es verdadera o falsa. La idea fundamental que hay que apreciar es que una oración no es nunca verdadera o falsa *por sí sola*. Considérese la oración

(1.1) Ella tiene agujetas.

La oración (1.1) no es verdadera o falsa por sí sola, independientemente de quién sea la persona a la que nos estemos refiriendo con el pronombre personal ‘Ella’. En algunos casos, al usar (1.1) nos estaremos refiriendo a una mujer que tiene agujetas, en otros nos referiremos a una que no tiene, o quizá no nos estaremos refiriendo a nadie, quizá porque estamos pronunciando (1.1) por el mero placer de oír cómo suenan las palabras.

Este fenómeno es más general, no tiene que ver exclusivamente con la aparición en las oraciones de pronombres y otras palabras cuyo contenido no es “permanente” y depende fundamentalmente de las intenciones comunicativas explícitas de los hablantes en las distintas ocasiones de uso. Otra manera en que (1.1) no es verdadera o falsa por sí sola

tiene que ver con el hecho de que el contenido de un signo gráfico o sonoro cualquiera es arbitrario, lo cual se refleja a veces en el fenómeno de que una misma grafía o sonido tiene de hecho contenidos diferentes en idiomas o dialectos diferentes. Por ejemplo, en el español de España 'agujetas' significa aproximadamente "dolores en un músculo que no se ejercitaba y se acaba de ejercitar intensamente", mientras que en el español de México significa "cordones de las zapatillas". Incluso si ha quedado claro que con 'Ella' en un uso de (1.1) nos estamos refiriendo a una determinada mujer, puede que en ese uso la oración (1.1) sea verdadera en un sentido de 'agujetas' y falso en el otro.

De manera completamente general, lo que estos ejemplos parecen mostrar es que una oración no es verdadera o falsa por sí sola, sino únicamente en cuanto que expresa un contenido, a veces llamado proposición, verdadero o falso por sí solo, que queda fijado por las intenciones comunicativas y asignaciones arbitrarias de contenidos a signos gráficos, sonoros o de otro tipo. Sin embargo, a menudo decimos en contextos ordinarios y en filosofía que tal o cual oración es verdadera o falsa, pero sólo porque se sobreentiende cuál es el contenido de los signos de la oración. Algunas teorías filosóficas de la noción de verdad, y en particular las teorías de Tarski y Kripke, han sido formuladas dando por supuesto que la verdad o falsedad se dice de oraciones cuyo contenido se ha fijado previamente.

Consideremos la siguiente oración:

(1.2) (1.2) es falsa.

Haciendo las suposiciones obvias acerca del contenido de los signos de (1.2), (1.2) parece expresar la proposición de que la oración (1.2) es falsa. ¿Es esa proposición verdadera, o es falsa? Parece que un principio obvio acerca de la noción de verdad es que una oración es verdadera si y sólo si lo que dice es el caso, de manera que lo siguiente es una oración verdadera:

(1.3) '(1.2) es falsa' es verdadera si y sólo si (1.2) es falsa.

Pero entonces, por la implicación de izquierda a derecha, si (1.2) es verdadera, entonces es falsa y por tanto no es verdadera, lo cual es una contradicción. Por otro lado, por la implicación de derecha a izquierda, si (1.2) es falsa entonces es verdadera, y por tanto no es

falsa; de nuevo una contradicción. Tanto si suponemos que (1.2) es verdadera como si suponemos que es falsa llegamos a una contradicción, y como no hay otras opciones (por el principio de tercio excluso) parece que hemos llegado a una contradicción incondicional, pura y simple. A esta contradicción se la conoce desde la antigüedad como la "antinomia del mentiroso" o "paradoja del mentiroso". No es claro todavía, aun después de que hayan pasado miles de años desde que fuera formulada por primera vez, cuál es la razón exacta de que surja la paradoja ni cuál es su solución. A veces, aunque ni mucho menos siempre, las teorías sobre qué es la verdad han sido motivadas en parte por la paradoja del mentiroso; ejemplos de ello son las teorías de Tarski y Kripke, que describiremos más abajo.

Un posible diagnóstico del problema de la paradoja del mentiroso es que la oración (1.2) no expresa realmente un contenido o proposición, pues al fin y al cabo es sólo una oración, y por tanto no es realmente verdadera o falsa en sí misma. (Aunque a muchos les parece claro que (1.2) expresa una proposición, que es por tanto verdadera o falsa.) Así se bloquea el dilema del razonamiento de la paradoja. La pregunta inmediata es entonces si puede reproducirse una paradoja similar para proposiciones, no para oraciones. El intento empezaría seguramente considerando una oración como la siguiente:

(1.2') (1.2') expresa (en español) una proposición falsa.

¿Es la proposición presumiblemente expresada por la oración (1.2') verdadera, o falsa? Podríamos razonar así: si esa proposición es verdadera entonces esa proposición es falsa, lo cual es absurdo. Pero si esa proposición es falsa entonces es verdadera, lo cual es absurdo también. Es posible proponer que (1.2') no expresa una proposición, con lo que el nuevo razonamiento queda también bloqueado; pero queda la cuestión de cómo justificar que (1.2') no expresa una proposición. Más adelante veremos que la teoría de Kripke puede verse como proporcionando una justificación de la sugerencia de que las oraciones problemáticas no expresan proposiciones.

2. Teorías metafísicas y epistemológicas del concepto de verdad

Hay fundamentalmente dos tipos de maneras en que se ha querido caracterizar la noción de verdad en la tradición filosófica. Uno de ellos emplea nociones metafísicas, el otro emplea nociones epistemológicas.

Ya hemos mencionado la idea básica del primer tipo de caracterizaciones: la idea de que una proposición es verdadera cuando se corresponde con un hecho. Se trata de una idea asociada al nombre de Aristóteles—quizá erróneamente, pero no hay duda de que es una idea cara a la tradición escolástica y específicamente tomista inspirada por el Estagirita. La oposición principal a esta idea ha provenido generalmente de filósofos reacios a aceptar el valor explicativo o cognoscitivo de nociones percibidas como metafísicas. En particular, las nociones de correspondencia y de hecho han suscitado esta sospecha. A veces se ha sostenido, por ejemplo, que la noción de correspondencia es excesivamente oscura, en parte por la oscuridad misma de la noción de proposición, en parte por la de la relación entre las proposiciones verdaderas y los hechos que les corresponden. Otra objeción es que no tenemos una concepción de la noción de hecho según la cual todas las proposiciones que llamaríamos verdaderas se correspondan con un "hecho"; quizá—se ha dicho—tenemos una concepción de la noción de hecho según la cual es un hecho que la Torre Eiffel tiene 300 metros de altura, pero no según la cual haya un hecho que corresponda a la proposición verdadera de que 17 es un número primo, o a la proposición verdadera de que Cleopatra fue bella. Más adelante veremos que los proponentes de las teorías redundantista y semántica de la verdad han visto a veces sus propuestas como versiones no metafísicas de la idea correspondentista.

Una tendencia general de la modernidad fue la progresiva "epistemización" de los problemas metafísicos. Esa tendencia se manifestó también en el ámbito de la reflexión filosófica sobre la noción de verdad, con la aparición de caracterizaciones dadas en términos de nociones epistemológicas. Algunos de los filósofos que criticaron la naturaleza metafísica de la teoría correspondentista fueron también responsables de la formulación de una de las teorías epistemológicas de la verdad más conocidas, la llamada teoría coherentista. En nuestra descripción de esta teoría nos basaremos fundamentalmente en la versión que de ella dieron algunos positivistas lógicos. (Véase Hempel (1935).) Pero otras versiones de la idea son rastreables seguramente a filósofos de las tradiciones empirista e idealista. La idea fundamental puede enunciarse así: una proposición es verdadera cuando es "coherente" con un grupo de otras proposiciones que se delimitan de una manera especial; este grupo de proposiciones especiales puede ser, por ejemplo, el de las proposiciones acerca de "datos elementales de la experiencia" que los seres humanos están

dispuestos a aceptar. ¿Qué quiere decir ‘coherencia’? Dar una respuesta satisfactoria a esta pregunta es uno de los problemas básicos del filósofo con inclinaciones coherentistas. La idea intuitiva es, sin embargo, que son verdaderas las proposiciones no refutables por algún tipo de verdades empíricas “básicas”. Una dificultad típica de estas teorías es que, dados ciertos supuestos bastante naturales, ninguna proposición excepto las muy “básicas” es refutable por las “verdades empíricas básicas”. Es un hecho generalmente admitido que cualquier proposición mínimamente desligada de la experiencia inmediata es compatible con las “verdades empíricas básicas”, supuesto que uno esté dispuesto a llamar verdaderas a un sistema apropiado de proposiciones “coherentes” con aquella proposición y con las “verdades empíricas básicas”. Y sin embargo parece que debería haber proposiciones no básicas que sean falsas en un sentido absoluto.

Otro tipo de caracterizaciones de la noción de verdad en términos de nociones epistemológicas son las caracterizaciones que se suele llamar “pragmatistas”, asociadas a los filósofos norteamericanos de la escuela del mismo nombre. (Véase James (1906).) La idea general de estas caracterizaciones es que una proposición verdadera es una que los seres humanos, dadas ciertas condiciones más o menos idealizadas de su desempeño cognoscitivo, llegarán a creer o aceptar en un estadio más o menos idealizado de la historia cognoscitiva humana. Quizá no sea irrazonable admitir que todas las proposiciones con esta propiedad (suponiendo que la propiedad en cuestión haya quedado unívocamente caracterizada meramente con lo recién dicho) serán verdaderas en un sentido intuitivo—si bien no es en absoluto claro que no vaya a haber alguna proposición falsa que los seres humanos estén condenados a aceptar a causa de la naturaleza de los mecanismos cognoscitivos de que disponen. Pero es incluso más discutible que todas las proposiciones verdaderas tengan la propiedad enunciada por el pragmatista; es razonable pensar que muchas verdades están fuera del alcance cognoscitivo de los seres humanos.

Desde luego, esto es sólo el inicio del debate acerca de las teorías coherentista y pragmatista. Los proponentes de este tipo de teorías disponen de argumentos sofisticados para defenderse de este tipo de objeciones elementales, argumentos que ocasionalmente pretenden establecer la incoherencia de la noción intuitiva de verdad en que esas objeciones se apoyan.

3. La distinción entre sustantivismo e insustantivismo

Algunos filósofos han visto en la teoría correspondentista el germen de una teoría de la verdad en términos de conceptos no metafísicos, y por tanto menos objetable que la teoría correspondentista. Una idea que los ha guiado a menudo ha sido la siguiente. En el uso más típico, uno usa las expresiones 'es verdad' y 'es verdadera' cuando uno dice de una cierta oración u oraciones que son verdad o que son verdaderas, y cabe pensar que en esos casos uno en cierto sentido sólo está haciendo de una manera algo más enfática lo que podría hacer afirmando directamente esa oración u oraciones. A esta idea, explorada quizá por primera vez por F. P. Ramsey (véanse Ramsey (1927) y Strawson (1950)), se la conoce como "la teoría redundantista" de la verdad. Por ejemplo, según la idea redundantista, cuando uno afirma

(3.1) 'Los gatos ronronean' es verdadera

uno en algún sentido no está diciendo sino lo mismo que podría decir afirmando

(3.1') Los gatos ronronean.

La idea redundantista implica la deseable conclusión de que los bicondicionales de esta forma expresan algún tipo de equivalencia fuerte (quizá incluso una equivalencia conceptual o analítica en el caso de (3.2b)):

(3.2a) 'Los gatos ronronean' es verdadera si y sólo si los gatos ronronean;
(3.2b) Que los gatos ronronean es verdad si y sólo si los gatos ronronean.

Pero debe notarse que aceptar estas equivalencias es compatible con la aceptación de otro tipo de teorías. Lo que sí parece ser una consecuencia peculiar de la teoría redundantista es que las expresiones 'es verdad' y 'es verdadera' carecen de una función "representativa", no están por una propiedad genuina de las oraciones, sino que se usan meramente de una forma redundante, quizá para añadir cierto énfasis a la afirmación de oraciones particulares. Así, la teoría redundantista tiene la consecuencia de que la noción de verdad no es tanto una noción de la que haya que dar una caracterización como una noción extremadamente tenue, que no es susceptible de caracterización más que en un sentido trivial.

La conexión con la teoría correspondentista se enfatiza a veces por quienes añaden a la idea redundantista la opinión de que 'es verdad' o 'es verdadera' son expresiones

equivalentes en algún sentido fuerte a las expresiones ‘se corresponde con los hechos’ o incluso ‘es un hecho’, al menos en ciertos sentidos de estas últimas expresiones. No es irrazonable pensar que pueda darse una equivalencia fuerte de este tipo, quizá incluso analítica—mientras que sería difícil sostener que ‘es verdad’ o ‘es verdadera’ son analíticamente equivalentes a ‘es coherente con un grupo de otras oraciones (o proposiciones)’ o ‘es tal que los seres humanos, dadas ciertas condiciones más o menos idealizadas de su desempeño cognoscitivo, llegarán a creerla o aceptarla en un estadio más o menos idealizado de la historia cognoscitiva humana’. De todos modos, parece más o menos claro que si uno sostiene la idea redundantista y al mismo tiempo sostiene que ‘es verdad’ o ‘es verdadera’ son expresiones analíticamente equivalentes a las expresiones ‘se corresponde con los hechos’ o ‘es un hecho’, esto sólo puede ser así en algún sentido especialmente débil, probablemente redundantista, de estas últimas expresiones—pues al afirmar una oración particular, sin usar concepto alguno de verdad, uno no parece estar predicando ninguna propiedad de la oración.

Pero la idea redundantista parece claramente objetable. Quizá la idea resulta plausible en el caso de (3.1) y (3.1’), pero no lo es en otros casos. Si yo afirmo

(3.3) La primera oración que pronunció Sor Juana Inés de la Cruz es verdadera

no estoy por ello haciendo algo que podría hacer afirmando la oración de la que hablo—no tengo ni idea de cuál pueda ser. De la misma manera, si afirmo

(3.4) Todos los teoremas demostrados por los matemáticos son verdaderos

no estoy por ello haciendo algo que podría hacer afirmando individualmente todas las oraciones matemáticas de las que hablo—hacerlo sería simplemente imposible.

En vista de ejemplos como (3.3) y (3.4) no parece posible sostener que las expresiones ‘es verdad’ y ‘es verdadera’ tienen una función enfática y realmente redundante; también parecen hacer difícil usar la motivación redundantista para sostener la tesis de que esas expresiones no expresan una propiedad de un cierto tipo de oraciones. Pero autores no redundantistas han buscado otras maneras de defender una tesis sólo ligeramente más débil, a saber, la tesis de que ‘es verdad’ y ‘es verdadera’ no expresan una propiedad “metafísica” o “sustantiva” de las oraciones (o proposiciones) que la poseen,

sino una propiedad "insustantiva". Es por ello que a las teorías de la verdad descritas en la sección anterior se las conoce como "sustantivistas", mientras que a las teorías en cuestión se las conoce como "insustantivistas" (*deflationary* en inglés). (En general, se habla también de la teoría redundantista como una teoría insustantivista, por razones obvias.)

Es razonable ver a las teorías kripkeana y tarskiana de la verdad, que nos ocuparán a continuación, como teorías insustantivistas de la verdad. Esta característica, que se apreciará cuando describamos las teorías y veamos el tipo de nociones en términos de las cuales delimitan el ámbito de las oraciones verdaderas, explica en buena medida la preeminencia de estas teorías en la tradición analítica. Otras virtudes que explican esta preeminencia son (i) el no verse afectadas por la dificultad generada para el redundantismo por oraciones como (3.3) y (3.4), (ii) el ofrecer un cierto tipo de "soluciones" para la paradoja del mentiroso y, fundamentalmente, (iii) el proporcionar caracterizaciones puramente matemáticas de las nociones de oración verdadera y de satisfacción. (En el caso de la teoría tarskiana, hay un rasgo adicional que explica su preeminencia especial, incluso por encima de la de la teoría kripkeana, y en particular su uso como piedra angular para desarrollar de una manera rigurosa la teoría matemática clásica de los modelos: la teoría tarskiana respeta el postulado clásico de que todas las oraciones de un lenguaje han de tener un valor de verdad.) Veremos por qué las teorías tarskiana y kripkeana tienen estas características (i) a (iii), y también en qué sentido delimitan el ámbito de las proposiciones verdaderas. Pero antes habremos de describir las teorías.

4. Las definiciones tarskianas de la verdad

La teoría semántica de la verdad fue formulada por A. Tarski a finales de los años 20 y principios de los años 30 del siglo XX. (Véanse Tarski (1944) para una exposición del propio Tarski en español, y Gómez Torrente (2001a) y (2001b) para una exposición y discusión histórica y filosófica de la presentación original de Tarski en los años 30; véase también Barrio (2014b) para otra exposición en español.) Lo que hizo Tarski fue ofrecer un *método* para dar *definiciones* de predicados de verdad para lenguajes formales particulares a cuyas expresiones se les han asignado significados concretos. (La restricción a lenguajes formales no es infranqueable; no hay obstáculos de principio a la formulación de definiciones tarskianas para lenguajes no formales, con tal que su sintaxis se pueda

especificar de una manera más o menos precisa.) Tarski no dio una definición de la noción de verdad para proposiciones en general, ni siquiera dio una definición de un predicado único de verdad para oraciones en general, ni siquiera de un predicado relacional de la forma '*O* es verdadera en *L*' que sea satisfecho por pares oración-lenguaje. Dado un lenguaje formal particular con significado *L*, el método de Tarski nos indica meramente cómo dar una definición de un predicado monádico 'es *OV* en *L*' que se aplica intuitivamente precisamente a las oraciones verdaderas de *L*.

Nótese que al decir que 'es *OV* en *L*' se aplicará intuitivamente precisamente a las oraciones verdaderas de *L* no hemos dicho que 'es *OV* en *L*' vaya a ser equivalente en un sentido fuerte a 'es verdadera' o siquiera a 'es verdadera en *L*', mucho menos que vaya a ser conceptual o analíticamente equivalente a alguna de estas expresiones. La manera como la teoría tarskiana de la verdad delimita el ámbito de la verdad es relativamente poco ambiciosa: la teoría da meramente un método para construir un predicado 'es *OV* en *L*' que es *coextensional* con (es decir, que se aplica exactamente a las mismas cosas que) el predicado intuitivo 'es verdadera en *L*'.

La definición de 'es *OV* en *L*' se formula en un *metalenguaje* apropiado para *L*, es decir, en un lenguaje por medio del cual se pueden decir suficientes cosas *acerca de L*. En un metalenguaje tarskiano para *L*, y por tanto en el *definiens* de la definición de 'es *OV* en *L*', aparecerán sólo tres tipos de términos: (a) términos de la sintaxis de *L*, en particular expresiones para nombrar a todas las expresiones de *L*, y expresiones que denoten ciertas operaciones definidas sobre expresiones de *L*, como la operación de concatenación de expresiones de *L*; (b) términos lógicos, conjuntísticos, y matemáticos en general: símbolos para conectivas y cuantificadores y recursos para hablar de nociones como pertenencia, inclusión, funciones, secuencias finitas e infinitas, cardinalidad de conjuntos, etc.; (c) por último, deberán aparecer también los términos específicos de *L* o traducciones suyas al metalenguaje. La definición de 'es *OV* en *L*' contendrá exclusivamente términos de los grupos (a), (b) y (c). Es por esta razón que la teoría de Tarski se ha visto a menudo como una teoría insustantivista de la verdad, pues los términos de los grupos (a), (b) y (c) no parecen nombrar propiedades que hayan parecido filosóficamente "sustantivas" o "metafísicas" (salvo si hay términos "sustantivos" en el grupo (c), y por tanto en *L*).

Dar una formulación general y abstracta de su método para definir el predicado ‘es *OV* en *L*’ en el metalenguaje apropiado para un lenguaje *L* cualquiera de los que Tarski tiene en mente sería una tarea algo ardua y, lo que es peor, la descripción resultante no sería muy iluminadora. Por ello Tarski mismo ilustró su método describiendo su aplicación a un lenguaje particular especialmente simple. Nosotros haremos lo mismo, y de hecho el lenguaje que nos servirá de ejemplo, al que llamaremos *L1*, será estructuralmente muy parecido al usado por Tarski. *L1* es un lenguaje de primer orden; en lo que sigue, daremos por supuesta la familiaridad del lector y la lectora con este tipo de lenguajes formales, así como con ciertas nociones conjuntísticas elementales.

Los signos primitivos de *L1* son el cuantificador universal (‘$\forall$’), el signo de conjunción (‘$\wedge$’), el signo de negación (‘$\neg$’), paréntesis (‘(’, ‘)’), un predicado diádico (‘A’), la letra ‘x’, y un acento subíndice (‘ˌ’) para generar un número infinito de variables por posposición a ‘x’ (para abreviar, usaremos la notación ‘x_n’ para la variable en que la letra ‘x’ va seguida de *n* acentos subíndices). El recorrido de las variables en la interpretación deseada de *L1* es el conjunto de todas las personas. El significado deseado de ‘A’ es la relación que se da entre dos personas cuando la primera ama a la segunda. Las interpretaciones de los otros signos lógicos son las normales. Las fórmulas atómicas son las de la forma ‘Ax_kx_l’. Las fórmulas complejas se obtienen por las operaciones de negación, disyunción (rodeada por paréntesis) y cuantificación universal con respecto a las variables.

En el metalenguaje para *L1* hay, como dijimos, nombres para todas las expresiones de *L1* (es decir, para las series finitas formadas por signos primitivos de *L1*). Esto se consigue incluyendo en ese metalenguaje nombres para los signos primitivos y obteniendo nombres para las otras expresiones por aplicaciones sucesivas de la función binaria de concatenación. Por ejemplo, nos podemos referir a la expresión ‘$\forall x_ıAx_ıx_ı$’ en el metalenguaje tarskiano para *L1* por medio de la expresión ‘la expresión formada concatenando el cuantificador universal con la primera variable con la letra a mayúscula con la primera variable con la primera variable, en ese orden’. Usaremos aquí el conocido artificio de las semicomillas debido a Quine para abreviar expresiones de este tipo. Con este artificio podemos abreviar, por ejemplo, la expresión ‘el cuantificador universal’ por medio de la expresión ‘$\ulcorner\forall\urcorner$’; la expresión ‘la expresión formada concatenando el

cuantificador universal con la primera variable con la letra a mayúscula con la primera variable con la primera variable, en ese orden', por medio de la expresión '$\ulcorner\forall x_1 A x_1 x_1\urcorner$', y la expresión 'la expresión formada concatenando el cuantificador universal con la *n*-ésima variable con la letra a mayúscula con la *n*-ésima variable con la *n*-ésima variable, en ese orden', por medio de la expresión '$\ulcorner\forall x_n A x_n x_n\urcorner$'.

Es importante subrayar la idea básica en virtud de la cual el predicado 'es *OV* en *L1*', una vez definido, será coextensional con el predicado intuitivo de verdad para oraciones de *L1*. Para ello hay que enunciar la famosa 'convención T' de Tarski. (La 'T' es porque 'truth' empieza por 't', y el inglés es el idioma en que más se ha hablado de la convención. En el polaco de los primeros textos de Tarski la convención lleva el nombre 'P', en el alemán el nombre 'W' y en castellano estrictamente hablando debería llevar el nombre 'V'.) La convención T es una definición enunciable en el *metametalenguaje* de *L1* (similares definiciones, o "convenciones T" se enunciarán en los metametalenguajes de otros lenguajes para los que queramos definir un predicado coextensional con el predicado intuitivo de verdad), y que intuitivamente da una condición suficiente para que el predicado 'es *OV* en *L1*' sea coextensional con el predicado intuitivo de verdad para oraciones de *L1*. La convención T para *L1*, por ejemplo, dirá que

> (T) una definición formalmente correcta de un símbolo 'es *OV* en *L1*' es una *definición adecuada de la verdad* si y sólo si tiene como consecuencias todas las oraciones que se obtienen a partir de la expresión '*x* es *OV* en *L1* si y sólo si *p*' reemplazando el símbolo '*x*' por un nombre "por concatenación" de una oración cualquiera de *L1* y el símbolo '*p*' por la expresión que es la traducción de aquella oración al metalenguaje.

(Nótese que los bicondicionales de que habla la convención T son similares a los bicondicionales de la forma de (3.2a).)

¿Cuál es el motivo por el que podemos pensar que la convención T enuncia una condición suficiente para que 'es *OV* en *L1*' sea coextensional con el predicado intuitivo de verdad para las oraciones de *L1*? El motivo lo da un argumento intuitivo como el siguiente. Abreviemos por medio de '*OV*' un predicado definido 'es *OV* en *L1*' cuya definición satisface la convención T. Supongamos primero que '*p*' es una oración de *L1* a la que se aplica el predicado '*OV*' (independientemente de si la oración metalingüística "*OV*('*p*')" es

demostrable); en otras palabras, supongamos que *OV*(‘*p*’). Sabemos que “*OV*(‘*p*’) si y sólo si *p*” es demostrable (es una consecuencia de una simple definición), y por tanto que es verdadera; así, por nuestro supuesto, podemos concluir que *p*, y por tanto que ‘*p*’ es verdadera en el sentido intuitivo. Por otro lado, supongamos que ‘*p*’ es verdadera en el sentido intuitivo; entonces *p*, y así, por el mismo razonamiento que antes, *OV*(‘*p*’).

Este razonamiento intuitivo no aparece en Tarski, y sin duda no lo habría considerado un razonamiento matemáticamente satisfactorio. Pero da una idea del tipo de consideraciones informales que sin duda justificaron para él el *desiderátum* de que una definición del predicado deseado ‘es *OV* en *L1*’ cumpliera con la convención T. Lo que estamos buscando puede resumirse, pues, así: buscamos un predicado definible con ayuda exclusivamente de los términos de los grupos (a), (b) y (c) del metalenguaje y que cumpla con la convención T (y del que pueda demostrarse que lo hace).

Tarski menciona que la idea que surge inmediatamente es la de dar una “definición recursiva” del predicado ‘es *OV* en *L1*’. Si esto es posible, la manera de hacerlo sería dar condiciones para la verdad de un cierto tipo de oraciones básicas, luego dar condiciones para la verdad de oraciones complejas en términos de las condiciones para la verdad de oraciones menos complejas, y por último definir el predicado ‘es *OV* en *L1*’ como aplicándose a aquellas oraciones que o son del tipo básico o son complejas pero la verdad “les ha sido transmitida” por la verdad o falsedad de las menos complejas. La idea se ve más claramente para un lenguaje aún más simple que *L1*. Digamos que *L2* es el lenguaje con signos primitivos ‘=’, ‘¬’, ‘0’, ‘1’, ‘2’, ‘3’, ‘4’, ‘5’, ‘6’, ‘7’, ‘8’ y ‘9’, cuyas fórmulas atómicas son las de la forma ‘*n*=*m*’ (con *n* y *m* numerales decimales), y cuyas fórmulas complejas se forman anteponiendo ‘¬’ a fórmulas menos complejas. Entonces podemos definir un predicado ‘es *OV* en *L2*’ apropiado para *L2* de la siguiente manera:

> una oración de *L2* es *OV* en *L2* si y sólo si (α) es de la forma ‘*n*=*n*’ (este es el tipo de oraciones verdaderas “básicas”) o (β) es de la forma ‘¬*O*’ y *O* no es verdadera.

Así enunciada, esta definición es indeseablemente circular (pues ‘es verdadera’ aparece en el *definiens*), pero es fácil transformarla en una definición no circular por medio de un truco matemático debido a Frege y Dedekind:

> una oración *O* de *L2* es *OV* en *L2* si y sólo si *O* pertenece a *todo* conjunto *C* de oraciones que (α) contiene a las oraciones de la forma '*n*=*n*' y que (β) siempre que no contiene a una oración *P* contiene a la oración '¬*P*'.

Sin embargo, esta manera de proceder no es factible en el caso de lenguajes como *L1*, pues en ellos hay oraciones complejas que no se forman a partir de *oraciones* menos complejas, sino de fórmulas con variables libres (la oración (falsa) '$\forall x_1 \forall x_{11}(Ax_1x_{11} \vee Ax_{11}x_1)$' es un ejemplo), y no se ve cómo especificar ni un conjunto básico de oraciones verdaderas ni un proceso recursivo de "transmisión" de la propiedad de la verdad en términos de la verdad o falsedad de oraciones menos complejas. Gran parte del mérito de Tarski se debe a que se diera cuenta de que, en sus propias palabras,

> se presenta la posibilidad de introducir un concepto más general que sea aplicable a cualquier fórmula, pueda ser definido recursivamente, y que, cuando se aplique a oraciones, nos lleve directamente al concepto de verdad. Estos requisitos los cumple la noción de la *satisfacción de una fórmula dada por objetos dados*.

La noción de satisfacción es una noción familiar: por ejemplo, el número 3 satisface la ecuación '*x*-2=1'; Frida Kahlo y Diego Rivera satisfacen (no importa en qué orden) la fórmula de *L1* 'Ax_1x_{11}'; Toluca, Morelia y Puebla satisfacen, en ese orden, la "fórmula" '*x* se halla entre *y* y *z*'. Tarski define recursivamente una versión abstracta de este concepto para *L1*: la noción de satisfacción de una fórmula cualquiera *F* por una secuencia *infinita h* (con dominio los enteros positivos, o lo que viene a ser lo mismo, las variables de *L1*, y recorrido incluido en la clase de las personas)—y no una noción de satisfacción para fórmulas y objetos, pares de objetos, tríos de objetos, etc. Intuitivamente, *F* es satisfecha por *h* cuando *F* es "verdadera" si interpretamos cada variable libre x_n que aparezca en *F* como si fuera un nombre de la persona *h*(*n*). La definición es la siguiente:

> Una secuencia infinita de personas *h* satisface una fórmula *F* si y sólo si *h* y *F* son tales que o bien (α) existen números naturales *k* y *l* tales que $F=\ulcorner Ax_kx_l \urcorner$ y h_k ama a h_l; o (β) hay una fórmula *G* tal que $F=\ulcorner \neg G \urcorner$ y *h* no satisface la fórmula *G*; o (γ) hay fórmulas *G* y *H* tales que $F=\ulcorner (G \wedge H) \urcorner$ y *h* satisface *G* y *h* satisface *H*; o finalmente (δ) hay un número natural *k* y una fórmula *G* tales que $F=\ulcorner \forall x_k G \urcorner$ y toda secuencia infinita de personas que difiera de *h* a lo sumo en el lugar *k*-ésimo satisface la fórmula *G*.

De nuevo esta definición recursiva es circular estrictamente hablando, pues emplea el predicado ‘satisface’. Pero de nuevo puede ponerse en forma no circular por medio del truco de Frege y Dedekind. La definición no circular es la siguiente:

> Una secuencia infinita de personas *h* satisface una fórmula *F* si y sólo si el par <*h*, *F*> pertenece a toda relación *R* entre secuencias y fórmulas tal que (α) si existen números naturales *k* y *l* tales que $G=\ulcorner Ax_k x_l \urcorner$ y g_k ama a g_l, entonces <*g*, *G*> pertenece a *R*; (β) si <*g*, *G*> no pertenece a *R*, entonces <*g*, $\ulcorner \neg G \urcorner$> pertenece a *R*; (γ) si <*g*, *G*> pertenece a *R* y <*g*, *H*> pertenece a *R*, entonces <*g*, $\ulcorner (G \wedge H) \urcorner$> pertenece a *R*; (δ) para todo número natural *k* y fórmula *G*, si toda secuencia infinita de personas *f* que difiera de una secuencia *g* a lo sumo en el lugar *k*-ésimo es tal que <*f*, *G*> pertenece a *R*, entonces <*g*, $\ulcorner \forall x_k G \urcorner$> pertenece a *R*.

La definición de satisfacción tiene como consecuencia que el que una secuencia *h* satisfaga una fórmula *F* o no sólo depende de lo que *h* asigne a (los subíndices de) las variables que aparezcan libres en *F*. En otras palabras, si la secuencia *h* satisface la fórmula *F*, y la secuencia infinita de personas *g* es tal que para todo *k*, $h_k=g_k$ si v_k es una variable libre de *F*, entonces la secuencia *g* también satisface la fórmula *F*. (Se habla de esto como un resultado o lema de *coincidencia*.) Esto quiere decir que si *F* es una oración —una fórmula sin variables libres— entonces o toda secuencia satisface *F* o ninguna lo hace. Además, la definición de satisfacción se ha construido de forma que las oraciones intuitivamente verdaderas son las del primer tipo, las satisfechas por toda secuencia, y por tanto es posible definir de esta forma el predicado ‘es *OV* en *L1*’:

> *O* es *OV* en *L1* si y sólo si *O* es una oración de *L1* y toda secuencia infinita de personas satisface *O*.

5. Virtudes de la teoría tarskiana de la verdad

Tarski señaló que la comprobación de que esta definición es apropiada se haría demostrando rigurosamente que es una “definición adecuada de verdad” en el sentido de la convención T. Ello requeriría la formalización de la metateoría, tarea ardua donde las haya que Tarski no estaba dispuesto a realizar. Sin embargo, el hecho es intuitivamente claro y Tarski se conformó con mostrar cómo se demostrarían en el metalenguaje algunos

bicondicionales de los mencionados en la convención T. He aquí un ejemplo de cómo se haría esto con el bicondicional correspondiente a la oración de *L1* '$\forall x_1 A x_1 x_1$', o sea

> ⌜$\forall x_1 A x_1 x_1$⌝ es *OV* en *L1* si y sólo si para toda persona *a*, *a* ama a *a*:
>
> ⌜$\forall x_1 A x_1 x_1$⌝ es *OV* en *L1* si y sólo si para toda secuencia *h*, *h* satisface '$\forall x_1 A x_1 x_1$'
>
> si y sólo si para toda *h*, para toda *g* que difiere de *h* a lo sumo en lo que asigna a 1, *g* satisface '$A x_1 x_1$'
>
> si y sólo si para toda *h*, para toda *g* que difiere de *h* a lo sumo en lo que asigna a 1, *g*(1) ama a *g*(1)
>
> si y sólo si para toda persona *a*, *a* ama *a* (pues dada una secuencia *h*, toda persona es asignada a 1 por alguna secuencia que difiere de *h* en lo asignado a 1).

Al satisfacer la convención T, una definición tarskiana del predicado de verdad será, como vimos, extensionalmente correcta. Quizá no es exagerado decir que, al indicar un sentido notablemente claro en que una caracterización "insustantivista" de la noción de verdad es indisputablemente extensionalmente correcta, la teoría semántica de la verdad constituye uno de los mayores logros en la investigación filosófica sobre esta noción.

Podemos apreciar también ahora las razones por las que la teoría semántica tiene las características (i) a (iii) mencionadas al final de la sección 3:

(i) No se ve afectada por la dificultad generada para el redundantismo por oraciones como (3.3) y (3.4). La razón es que, a pesar de ser "insustantivista" en un sentido más o menos claro, la teoría semántica no propone que la expresión 'es verdadera' no exprese una propiedad, y, además, los predicados definidos que muestra cómo construir son predicados genuinos, que expresan propiedades de un cierto tipo, si bien propiedades "insustantivas". En el supuesto de que tengamos un predicado 'es *OV* en *L*' en un metalenguaje para un lenguaje *L* que contenga la primera oración pronunciada por Sor Juana Inés de la Cruz (y que incluya en dicho metalenguaje la descripción 'la primera oración pronunciada por Sor Juana Inés de la Cruz') y que contenga todos los teoremas matemáticos (e incluya en dicho metalenguaje el predicado 'es un teorema matemático'), cosa para la cual no hay obstáculos de principio, podremos decir sin problemas que la primera oración pronunciada por Sor Juana es *OV* en *L* y que todos los teoremas matemáticos son *OV* en *L*.

(ii) Ofrece un cierto tipo de "solución" para la paradoja del mentiroso. La razón es que el predicado 'es *OV* en *L*' es siempre un predicado de un lenguaje diferente al lenguaje de las oraciones a las que se aplica. En el caso de la oración (1.2) es fácil ver cómo este hecho bloquea la derivación de la paradoja para 'no es *OV* en *L*' (que expresaría "es falsa en *L*"): no es posible formar una oración del lenguaje para el que hemos definido 'es *OV* en *L*' que diga de sí misma que no es *OV* en *L*, pues 'no es *OV* en *L*' no es un predicado de ese lenguaje; por otro lado, es ciertamente posible formar una oración *O* del metalenguaje que diga de sí misma que no es *OV* en *L*, pero al ser una oración del metalenguaje es simplemente verdadera y no paradójica, pues no hay razón de que valga para ella el bicondicional análogo a (1.3) ('*O* es *OV* en *L* syss *O* no es *OV* en *L*'), que es esencial para la derivación de la paradoja; ese tipo de bicondicionales sólo valen para oraciones del lenguaje para el que se ha definido 'es *OV* en *L*', no para oraciones del metalenguaje.

Mencionemos aquí que en su trabajo sobre la verdad Tarski aplica el razonamiento de la paradoja del mentiroso en una prueba puramente matemática de que no es posible que un lenguaje de a lo sumo el mismo orden que un lenguaje *L* (y que pueda funcionar como metalenguaje de *L,* en el sentido de que pueda expresar suficientes cosas sobre la sintaxis de *L*) contenga un predicado de verdad para *L.* El ejemplo habitual es el lenguaje usual para la aritmética de primer orden y sus extensiones. Como Gödel mostró (y también Tarski, de forma independiente), estos lenguajes pueden hablar de su propia sintaxis: las expresiones de un lenguaje *L* de este tipo pueden ponerse en una correspondencia biunívoca con un subconjunto de los naturales (el de los números de Gödel de las expresiones de *L*), y para cada predicado sintáctico relevante puede definirse un predicado aritmético satisfecho precisamente por los números correspondientes a las expresiones que satisfacen el predicado sintáctico original. Usando resultados debidos en esencia a Gödel, es posible probar que, para cada fórmula (con una variable libre) $F(x)$ de un lenguaje que extienda al de la aritmética, existe una oración *O* del mismo lenguaje verdadera si y sólo si el número de Gödel de *O* satisface $F(x)$. Supongamos que "verdad para *L*" fuera expresable en *L* por medio de una fórmula $V(x)$, esto es, que $V(x)$ es satisfecha por todos y sólo los números de Gödel de oraciones verdaderas de *L*; por la proposición anterior habría una oración de *L*, *O*, verdadera si y sólo si el número de Gödel de *O* satisface $\neg V(x)$; *O* "diría de sí misma" que es falsa. *O* es o bien verdadera o bien no lo es (siempre en la interpretación deseada de *L*);

si es verdadera, su número de Gödel satisface $\neg V(x)$ y por tanto *O* no es verdadera; si no es verdadera, su número de Gödel satisface $\neg V(x)$, y por tanto es verdadera. Así, *O* es verdadera si y sólo si no lo es; la contradicción nos permite concluir que *L* no puede contener su propio predicado de verdad.

(iii) Por último, la teoría semántica proporciona una caracterización puramente matemática de las nociones de oración verdadera y de satisfacción, lo cual se ve claramente al inspeccionar los recursos en términos de los que se dan las definiciones tarskianas, que sólo incluyen esencialmente recursos matemáticos (los recursos sintácticos pueden matematizarse, como acabamos de mencionar). Como dijimos, quizá esta característica es la principal responsable de la preeminencia de la teoría semántica entre los lógicos e, indirectamente, entre los filósofos, especialmente entre los filósofos analíticos.

6. Algunas críticas a la teoría de Tarski y la motivación de la teoría kripkeana

A pesar de su preeminencia en los contextos donde se ha requerido un uso científico del concepto de verdad, en particular en el ámbito de la teoría de modelos en lógica matemática, la teoría semántica de la verdad no ha carecido de críticas. La raíz fundamental de todas ellas son las diferencias entre los predicados definidos por el método tarskiano y el predicado intuitivo de verdad. Una de esas diferencias es, como ya mencionamos, que son meramente extensionalmente equivalentes, pero no son analíticamente equivalentes, ni siquiera necesariamente equivalentes. Otra de esas diferencias es que no es posible decir con un predicado tarskiano—que siempre está restringido a un lenguaje particular—cosas que parecen poder decirse en virtud de la "universalidad" del predicado intuitivo (un ejemplo es la falsedad 'Todas las oraciones de todos los lenguajes son verdaderas'). Las críticas de Kripke a la teoría de Tarski tienen que ver con estas limitaciones expresivas de los predicados tarskianos.

Observemos que, si quisiéramos representar en el lenguaje natural, por medio de predicados definidos según las ideas de Tarski, el razonamiento de la antinomia del mentiroso usando la oración (1.2), al predicar 'no es verdadera' (que sería la forma tarskiana de expresar 'es falsa') de (1.2) en el razonamiento, este 'no es verdadera' no podría ser el mismo predicado que el que aparece usado (en la forma 'es falsa') en (1.2); este último pertenecería a un lenguaje o "capa del lenguaje" "inferior" a la del predicado

‘es verdadera’ que predicamos de (1.2). Así, si el lenguaje natural funcionara según el modelo que parecería poder extraerse de las ideas de Tarski, deberíamos verlo como conteniendo muchos (de hecho, seguramente infinitos) predicados diferentes de verdad, todos ellos gráfica y fonéticamente idénticos, eso sí. (En realidad, Tarski no veía sus ideas como un modelo del funcionamiento del lenguaje natural, aunque fue natural para muchos conjeturar que sí apuntaban a un modelo plausible del concepto de verdad en el lenguaje natural, con muchos predicados diferentes pero gráfica y fonéticamente idénticos.) Al principio tendríamos un primer predicado de verdad que predicaríamos (o cuya negación predicaríamos) de oraciones que no contuvieran ningún predicado de verdad; después, un segundo predicado de verdad que predicaríamos (o cuya negación predicaríamos) de las oraciones que contuvieran a lo sumo el primer predicado de verdad, y así sucesivamente.

Una de las motivaciones de Kripke (1975) es ofrecer un intento de solución a la antinomia del mentiroso que supere algunas objeciones que genera el modelo del funcionamiento del concepto de verdad en el lenguaje natural que acabamos de describir—Kripke lo llama “el punto de vista ortodoxo”, presumiblemente porque muchos en su época pensaban que ese modelo era plausible (¡aunque no Tarski!). Veamos algunas de las objeciones que expone el mismo Kripke. En primer lugar, aunque parece intuitivamente claro que toda oración en que aparezca el predicado de verdad debe de tener un “nivel” en el sentido explicado en el párrafo anterior, este nivel puede depender no sólo de la forma de la oración, como se supone, según Kripke, desde el punto de vista ortodoxo, sino también de ciertos hechos empíricos relativos a ella. Si Biden dice que la oración

(6.1) Todas las afirmaciones de Trump sobre la elección de Georgia son falsas

es verdadera, el “nivel” de este ‘es verdadera’ puede ser muy alto, dependiendo de las iteraciones de predicaciones de verdad o falsedad que haya hecho Trump en sus afirmaciones; puede no tratarse del segundo ‘es verdadera’ de la lista infinita del párrafo anterior. Por ejemplo, Trump puede haber dicho ‘Jones no dice la verdad cuando dice que Smith dijo la verdad cuando afirmó que la elección de Georgia fue legal’).

Otra de las objeciones que expone Kripke es la siguiente. Supongamos que Jones afirma (6.1), y que Trump afirma a su vez

(6.2) Todas las afirmaciones de Jones sobre la elección son falsas.

Entonces, en particular, Jones quiere afirmar que (6.2) es falsa, y Trump quiere afirmar que (6.1) es falsa; para ello, el nivel de 'es falso' en (6.1) ha de ser mayor que el de 'es falso' en (6.2), y viceversa, pero esto es imposible. Desde el punto de vista ortodoxo, al menos una de las dos oraciones estaría mal formada o no tendría sentido; sin embargo, podemos a menudo asignar valores de verdad a (6.1) y (6.2) sin ambigüedad: si al menos una afirmación de Jones sobre la elección es verdadera, (6.2) será falsa; si todas las otras afirmaciones de Trump sobre la elección son también falsas, (6.1) será verdadera. Pero no siempre podríamos decidirnos sobre el valor de verdad de (6.1) y (6.2): considérese el caso de que lo único que hubiesen afirmado Jones y Trump sobre la elección hubiesen sido, respectivamente, (6.1) y (6.2). El punto de vista ortodoxo elimina estos casos problemáticos barriendo de paso los casos en que (6.1) y (6.2) no presentan problemas.

La solución de Kripke a la antinomia del mentiroso pasa por proponer que el lenguaje natural puede verse como conteniendo únicamente un predicado de verdad. Kripke propone evitar la paradoja postulando que ciertas oraciones que contienen el predicado de verdad, entre ellas las paradójicas, no son ni verdaderas ni falsas, por no tener condiciones de verdad claras; en otras palabras, el predicado de verdad no está completamente definido, hay "huecos" en cuanto al valor de verdad (*truth-value gaps*). Ello bloquea la aplicación del principio de tercio excluso en el argumento para obtener la paradoja. Kripke caracteriza el conjunto de las oraciones que contendrán el predicado de verdad y que serán verdaderas o falsas mediante la siguiente parábola: supongamos que queremos explicar el significado de 'es verdadera' a alguien que es en general un hablante competente del español, pero que no conoce el significado de ese predicado; si le decimos que una oración es verdadera cuando se dan las condiciones que nos permitirían afirmarla, el hablante podrá descubrir sucesivamente el valor de verdad de oraciones como "'La tierra gira alrededor del sol' es verdadera", " " 'La tierra gira alrededor del sol' es verdadera" es verdadera", etc., que en última instancia podrá reducir a las condiciones de verdad de oraciones cuyo significado ya conoce; pero no podrá asignar un valor de verdad, por

ejemplo, a oraciones autorreferenciales en que se predique de sí mismas la verdad o la falsedad. Verbigracia, no podrá asignar un valor de verdad a

(6.3) (6.3) es verdadero,

pues al buscar las condiciones que le permitirían afirmar (6.3) volverá siempre al punto de partida. Lo mismo le ocurriría con las oraciones paradójicas como (1.2), y con ejemplos de referencias cruzadas (como en el caso de que lo único que hubiesen afirmado Jones y Trump sobre la elección hubiesen sido, respectivamente, (6.1) y (6.2)). A las oraciones para las que el hablante del ejemplo (o nosotros mismos) podría encontrar un valor de verdad, eliminando predicados del tipo 'es verdadera', Kripke las llama "fundadas" (*grounded*).

La solución kripkeana a la paradoja del mentiroso, que hemos esbozado en lo esencial, no es la contribución más importante de Kripke a esta cuestión (de hecho, soluciones muy similares habían sido propuestas con anterioridad al artículo de Kripke). Lo más importante es que Kripke muestra cómo extender cualquier lenguaje interpretado (que pueda hablar de su propia sintaxis) a otro lenguaje interpretado que contiene su propio predicado de verdad y en el que hay "huecos" en los valores de verdad, o sea en el que el predicado de verdad no está completamente definido. En concreto, los correlatos en la interpretación de las oraciones no fundadas, y por tanto los de las paradójicas, no satisfarán ni el predicado de verdad ni su negación. Naturalmente, el uso de Tarski del principio de tercio excluso en la prueba de que un lenguaje no puede contener su propio predicado de verdad presuponía que la interpretación del lenguaje considerado era clásica, esto es, que todos los predicados del lenguaje estaban completamente definidos por la interpretación—que cualquier objeto del dominio de la interpretación o bien satisface un predicado dado o su negación. A continuación, expondremos sucintamente el método de Kripke y el modelo a que da lugar. (En esta exposición de nuevo daremos por supuesta la familiaridad del lector y la lectora con varios aspectos de los lenguajes formales de primer orden, así como con ciertas nociones conjuntísticas elementales, aunque algo más avanzadas que las usadas en las definiciones tarskianas. Véase también la exposición en Teijeiro y Szmuc (2014).) Concluiremos viendo que este modelo no está sujeto a las objeciones que Kripke ponía al

punto de vista ortodoxo, pero parece estar sujeto a otra objeción a su capacidad de modelar todas las intuiciones sobre el predicado de verdad en el lenguaje natural, la objeción generada por las llamadas "oraciones del mentiroso reforzadas".

7. El método kripkeano para definir el predicado de verdad

Sea $\{P_1,...,P_n, f_1,...,f_m, c_1,...,c_p\}$ el vocabulario no lógico de un lenguaje (finito) de primer orden *L*, en cuyo vocabulario lógico estén el cuantificador universal ('$\forall$'), el signo de conjunción ('$\wedge$'), el signo de negación ('$\neg$'), paréntesis ('(', ')'), el predicado diádico de identidad ('='), la letra 'x', y un acento subíndice (',') para generar un número infinito de variables por posposición a 'x'. (Como de costumbre, $P_1,...,P_n$ son los predicados de *L*, $f_1,...,f_m$ son sus símbolos de función, y $c_1,...,c_p$ son sus constantes individuales.) Diremos que una interpretación **A** para ese vocabulario y por tanto para *L* es *no clásica* si es un conjunto de la forma

$$\mathbf{A}=<A, <<P_1^{\mathbf{A}1}, P_1^{\mathbf{A}2}>,..., <P_n^{\mathbf{A}1}, P_n^{\mathbf{A}2}>>, <f_1^{\mathbf{A}},...,f_m^{\mathbf{A}}>, <c_1^{\mathbf{A}},...,c_p^{\mathbf{A}}>>,$$

donde A es un conjunto no vacío, $P_i^{\mathbf{A}i}$ es siempre un subconjunto de A^k (cuando k es la adicidad de P_i), $P_i^{\mathbf{A}1} \cap P_i^{\mathbf{A}2} = \varnothing$, $f_i^{\mathbf{A}}$ es siempre una función de A^k (cuando k es la adicidad de f_i) en A, y $c_i^{\mathbf{A}}$ es siempre un elemento de A. (Así, si para todo i=1,...,n, si P_i es k-ádico, $P_i^{\mathbf{A}1} \cup P_i^{\mathbf{A}2} = A^k$, **A** es en esencia una interpretación clásica). Hablaremos de $P^{\mathbf{A}1}$ como la *extensión* de un predicado P, y de $P^{\mathbf{A}2}$ como su *antiextensión.*

Podemos definir para *L,* una interpretación no clásica **A**, y cada secuencia h que sea una asignación de objetos de A al conjunto de las variables de *L*, una función de denotación h* del conjunto de los términos de *L* en A, análoga a la definida para interpretaciones clásicas. Kripke usa la lógica trivaluada fuerte de Kleene para tratar los predicados no completamente definidos; siguiendo los esquemas de Kleene podemos definir recursivamente una función **A*** que asigne a cada formula de *L* un valor de entre v, f, i (leídos "verdadero", "falso", "indeterminado") mediante una secuencia h, de la siguiente manera:

$\mathbf{A}^*_h(\ulcorner t_1=t_2 \urcorner)$ (donde t_1 y t_2 son términos de *L*) es v si $h^*(t_1)=h^*(t_2)$; f si $h^*(t_1)\neq h^*(t_2)$; nunca es i.

$\mathbf{A}^*_h(\ulcorner Pt_1,...,t_k \urcorner)$ (donde P es un predicado k-ádico de *L* y $t_1,...,t_k$ son términos de *L*) es v si $<h^*(t_1),...,h^*(t_k)> \in P^{\mathbf{A}1}$; es f si $<h^*(t_1),...,h^*(t_k)> \in P^{\mathbf{A}2}$; es i si $<h^*(t_1),...,h^*(t_k)> \notin P^{\mathbf{A}1} \cup P^{\mathbf{A}2}$.

$\mathbf{A}^*_h(\ulcorner \neg F \urcorner)$ (donde *F* es una fórmula de *L*) es v si $\mathbf{A}^*_h(F)$ es f; es f si $\mathbf{A}^*_h(F)$ es v; es i si $\mathbf{A}^*_h(F)$ es i.

$\mathbf{A}^*_h(\ulcorner F \wedge G \urcorner)$ (donde *F* y *G* son fórmulas de *L*) es v si $\mathbf{A}^*_h(F) = \mathbf{A}^*_h(G) =$ v; es f si $\mathbf{A}^*_h(F) =$ f o $\mathbf{A}^*_h(G) =$ f; es i en otro caso.

$\mathbf{A}^*_h(\ulcorner \forall x_n F \urcorner)$ (donde *F* es una fórmula de *L*) es v si para toda secuencia j que difiera de h a lo sumo en el valor de x_n, $\mathbf{A}^*_j(F) =$ v; es f si hay una secuencia j que difiere de h a lo sumo en el valor de x_n y tal que $\mathbf{A}^*_j(F) =$ f; es i en otro caso.

Esta semántica permite probar fácilmente un lema de coincidencia para oraciones de *L*.

Vayamos con la construcción de Kripke. Partamos de una cierta interpretación clásica **A** de *L*. Supongamos además que *L* tiene suficientes recursos para hablar de su sintaxis (*L* puede ser el lenguaje usual de primer orden para la aritmética; también podemos imaginar que *L* es algún fragmento formalizado del lenguaje natural sin el predicado 'es verdadero'). Sea *L'* el lenguaje que resulta de añadir a *L* un predicado monádico 'V' (que desempeñará el papel de 'es verdadero'). Sea $\mathbf{A}(S_1,S_2)$ la interpretación no clásica para *L'* definida como sigue: $\mathbf{A}(S_1,S_2)$ coincide con **A** en el universo y en la interpretación de los signos de *L*; además, S_1 es $V^{\mathbf{A}(S1,S2)1}$ y S_2 es $V^{\mathbf{A}(S1,S2)2}$. Sea $[O]$ el "código" de una oración *O*, es decir, el objeto que le corresponde en la interpretación (en el caso de la aritmética su número de Gödel; en el caso del lenguaje natural, la oración misma). Sean entonces

S_1'={$[O]$: *O* es una oración de *L'* y $\mathbf{A}(S_1,S_2)^*(O)$=v}

S_2'={$[O]$: *O* es una oración de *L'* y $\mathbf{A}(S_1,S_2)^*(O)$=f} $\cup$

$\{a \in A$: a no es el código de una oración de *L'*$\}$.

Si queremos interpretar 'V' como "verdad en $\mathbf{A}(S_1,S_2)$", debemos exigir que $S_1=S_1$' y $S_2=S_2$', esto es, que la extensión de 'V' sean (los códigos de) las oraciones verdaderas en $\mathbf{A}(S_1,S_2)$ y que su antiextensión sean (los códigos de) las oraciones falsas (más las cosas que no son (códigos de) oraciones). Sea ϕ la función que asigna a cada par de subconjuntos disjuntos de A, $<S_1,S_2>$, el par $<S_1',S_2'>$ (en la notación de más arriba). Un punto fijo de ϕ (esto es, un par $<S_1,S_2>$ de subconjuntos disjuntos de A para el que $\phi(<S_1,S_2>)=<S_1,S_2>$ $(=<S_1',S_2'>)$) proporcionará por tanto una interpretación apropiada del predicado 'V'. Así, probar la existencia de puntos fijos de ϕ es probar la existencia de una interpretación de *L'* bajo la cual *L'* contiene su propio predicado de verdad. El principal resultado de Kripke es la prueba de que existen puntos fijos de ϕ.

Para cada ordinal α definimos por recursión transfinita una estructura no clásica $\mathbf{A}_\alpha$ como sigue (si $\mathbf{A}_\beta= \mathbf{A}(S_1,S_2)$, ponemos $S_1=S_{1\beta}$ y $S_2=S_{2\beta}$):

$\mathbf{A}_0= \mathbf{A}(\varnothing,\varnothing)$;

si $\alpha=\beta+1$ y $\mathbf{A}_\beta= \mathbf{A}(S_1,S_2)$, $\mathbf{A}_\alpha= \mathbf{A}(S_1',S_2')$;

si α es un ordinal límite, $\mathbf{A}_\alpha= \mathbf{A}(\cup_{\beta<\alpha}S_{1\beta}, \cup_{\beta<\alpha}S_{2\beta})$.

Hay un paralelo entre la generación recursiva de estas estructuras y el proceso imaginario mediante el que el hablante de nuestra parábola asignaba sucesivamente valores de verdad a nuevas oraciones que contenían el predicado 'es verdadero': en un primer momento, la interpretación del predicado está completamente indeterminada (en el nivel 0) y sucesivamente van apareciendo en su extensión y su antiextensión nuevas oraciones.

Digamos que $<S_1,S_2>\leq<S_1^\dagger,S_2^\dagger>$ ($<S_1^\dagger,S_2^\dagger>$ extiende a $<S_1,S_2>$) si y sólo si $S_1\subset S_1^\dagger$ y $S_2\subset S_2^\dagger$. Entonces ϕ es monótona respecto del orden $\leq$: si $<S_1,S_2>\leq<S_1^\dagger,S_2^\dagger>$, entonces $\phi(<S_1,S_2>)\leq\phi(<S_1^\dagger,S_2^\dagger>)$ (la prueba es una inducción sobre la complejidad de las oraciones de *L'*). Por la monotonía de ϕ, es fácil ver por inducción ordinal que si $\beta<\alpha$, la

interpretación de 'V' en $\mathbf{A}_\alpha$ extiende a la interpretación de 'V' en $\mathbf{A}_\beta$. Ahora no es difícil ver que hay α tal que $\phi(<S_{1\alpha},S_{2\alpha}>)=<S_{1\alpha},S_{2\alpha}>$, o sea que hay puntos fijos de ϕ. Esto se ve observando que habrá un ordinal α tal que $<S_{1\alpha},S_{2\alpha}>=<S_{1\alpha+1},S_{2\alpha+1}>$: si para todo α o bien $S_{1\alpha}$ está incluido propiamente en $S_{1\alpha+1}$ o bien $S_{2\alpha}$ está incluido propiamente en $S_{2\alpha+1}$, entonces el conjunto de oraciones de *L'* en $S_{1\omega_1} \cup S_{2\omega_1}$ sería no numerable, pero sólo hay un número numerable de oraciones en *L'*. La existencia de un punto fijo de ϕ garantiza que *L'* puede interpretarse de modo que 'V' sea un predicado de verdad para *L'*.

Siguiendo con la metáfora anterior, un punto fijo sería el análogo del momento en que el hablante imaginario ya no puede dar más valores de verdad a oraciones del español; fuera de la extensión y la antiextensión para 'V' que suministra el punto fijo quedarán las oraciones no fundadas, y entre ellas las paradójicas. Se puede ahora dar una definición formal precisa de lo que es una oración fundada (de *L'*): es una oración que es verdadera o falsa en la interpretación que suministra el menor punto fijo de ϕ. Si *O* es una oración fundada, el menor ordinal α para el que $\mathbf{A}_\alpha$ da un valor de verdad a *O* proporciona una medida del nivel de *O*, en el sentido intuitivo de 'nivel' del que habíamos hablado. Es fácil ver, por inducción ordinal, que las oraciones de *L'* en las que aparecen autopredicaciones de verdad o falsedad (por ejemplo, las oraciones de la aritmética de las que hablábamos al final de la sección 5, o la oración (6.3) y las oraciones autorreferenciales paradójicas como (1.2) en el caso del lenguaje natural) no serán ni verdaderas ni falsas en ninguna interpretación $\mathbf{A}_\alpha$; así, no serán fundadas.

Veamos cómo se resuelven en el modelo recién descrito los problemas que Kripke detectaba en el "punto de vista ortodoxo" sobre la paradoja del mentiroso. En primer lugar, en el caso de oraciones como (6.1), la interpretación inicial (**A**, en la construcción anterior) proporciona la extensión del predicado 'ser una afirmación de Trump sobre la elección de Georgia'; si en ella hay alguna oración no fundada, (6.1) será no fundada; si todas las oraciones en ella son fundadas, (6.1) también lo será, y pertenecerá a la extensión o la antiextensión de 'V' en alguna $\mathbf{A}_\alpha$ con α suficientemente grande; en cualquier caso, el nivel de (6.1) (en el sentido preciso mencionado en el párrafo anterior) dependerá de **A** de un modo esencial. Lo mismo ocurrirá en el caso de la referencia cruzada de (6.1) y (6.2): que sean fundadas dependerá de las extensiones de 'ser una afirmación de Trump sobre la

elección de Georgia' y de 'ser una afirmación de Jones sobre la elección', proporcionadas por **A**, de la que también dependerá el nivel de (6.1) y (6.2) si son fundadas. En el ejemplo anterior, si una de las afirmaciones de Jones sobre la elección es fundada y verdadera, de nivel α, (6.2) será falsa de nivel $\alpha+1$; si además todas las afirmaciones de Trump distintas de (6.2) son falsas, entonces (6.1) será verdadera y de un nivel superior a $\alpha+1$.

También podemos ver que la teoría de Kripke cumple con los desiderátums (i) a (iii) de la sección 3. En cuanto a (i), no se ve afectada por la dificultad generada para el redundantismo por oraciones como (3.3) y (3.4), pues, al igual que la teoría de Tarski, la teoría de Kripke propone que el predicado de verdad expresa una propiedad de un cierto tipo, si bien una razonablemente "insustantiva". De nuevo un predicado kripkeano 'V' para un lenguaje apropiado *L* podrá expresar que la primera oración pronunciada por Sor Juana es V y que todos los teoremas matemáticos son V. Por otro lado, la teoría evidentemente ofrece un cierto tipo de solución para la paradoja del mentiroso, como vimos hace un par de párrafos, según requería (ii). Por último, según requería (iii), la teoría de Kripke proporciona una caracterización meramente conjuntística, y por tanto puramente matemática, de la noción de oración verdadera para un lenguaje, lo cual de nuevo comprobamos inspeccionando los recursos utilizados en la construcción de más arriba.

¿Quiere esto decir que la teoría de Kripke es claramente la teoría correcta de la verdad, o el modelo correcto del funcionamiento del predicado de verdad en el lenguaje natural, o al menos la teoría o el modelo correcto dados los desiderátums insustantivistas? Hay al menos una objeción que podría usarse para poner esto en duda, creada por las llamadas "oraciones del mentiroso reforzado". Una oración del mentiroso reforzado es, en la versión más simple y típica, una oración que afirma de sí misma que no es verdadera, a diferencia de una oración del mentiroso normal, que afirma de sí misma que es falsa (el caso de (1.2), recordemos); por ejemplo,

(7.1) (7.1) no es verdadera

es una oración del mentiroso reforzado.

Si aceptamos el principio de bivalencia de la lógica clásica, (7.1) es equivalente a

(7.2) (7.1) es falsa.

Si suponemos, por contra, que hay oraciones sin valor de verdad, ni verdaderas ni falsas, (7.1) y (7.2) no son equivalentes. (7.1) es intuitivamente paradójica: si es verdadera, entonces no lo es; si no es verdadera, entonces es verdadera, pues afirma precisamente que no es verdadera.

En los modelos con huecos en cuanto al valor de verdad, las oraciones paradójicas como (7.1) carecen de valor de verdad. Por ejemplo, consideremos el modelo proporcionado por un punto fijo como el construido inductivamente por Kripke para el lenguaje de la aritmética de primer orden con un predicado de verdad, 'V'. Si n* es el nombre estándar (numeral) de n y n es el número de Gödel de ⌜¬Vn*⌝, entonces n no está ni en la extensión ni en la antiextensión de 'V' en el modelo del punto fijo, y así ⌜¬Vn*⌝ no es ni verdadera ni falsa en el modelo. (Esto puede verse mostrando por inducción ordinal que n no pertenece a la extensión ni a la antiextensión de 'V' en ninguna de las estructuras $\mathbf{A}_\alpha$.) El problema resulta entonces de la siguiente observación: si ⌜¬Vn*⌝ no es ni verdadera ni falsa, entonces no es verdadera, pero esto parece ser precisamente lo que afirma ⌜¬Vn*⌝, y por tanto ⌜¬Vn*⌝ será verdadera después de todo; en el lenguaje natural, el argumento paralelo procedería así: si '(7.1) no es verdadera' no es ni verdadera ni falsa, entonces '(7.1) no es verdadera' no es verdadera, pero entonces (7.1) no es verdadera, y así '(7.1) no es verdadera' será verdadera después de todo.

Naturalmente, esta contradicción no puede reproducirse formalmente, pues, como señala Kripke, el lenguaje objeto (aquel para el que se construye el modelo) no puede expresar que una oración suya es o falsa o carece de valor de verdad; es decir, no puede expresar que una oración no es verdadera, en el sentido en que decimos en el metalenguaje (por ejemplo, en el argumento anterior que llevaba a la contradicción) que no es verdadera. Ello se debe a la interpretación de la negación en la lógica trivaluada de Kleene: que ⌜¬Vt⌝ sea verdadera en una estructura no clásica **A** quiere decir que ⌜Vt⌝ es falsa, es decir, que el número (de Gödel) nombrado por t pertenece a la antiextensión de 'V' en **A**, o sea que la oración cuyo número de Gödel es el número denotado por t es falsa; no quiere decir que esta oración es o falsa o carece de valor de verdad.

Así, las oraciones del mentiroso reforzado parecen poner de manifiesto una limitación en los medios de expresión de los lenguajes interpretados mediante una semántica trivaluada: el predicado 'V' del lenguaje objeto y el predicado 'es verdadero' del metalenguaje no expresan lo mismo, pues es verdadero que una oración del mentiroso reforzado como ⌜¬Vn*⌝ no es verdadera, según la semántica intuitiva del metalenguaje, pero suponer que la traducción obvia de esta afirmación metalingüística es verdadera en el modelo trivaluado, es decir, suponer que el número de Gödel de ⌜¬Vn*⌝ está en la extensión de 'V', implica una contradicción. Una de las motivaciones de las construcciones de modelos trivaluados es sin embargo mostrar que un predicado del lenguaje objeto puede ser interpretado como aplicándose a todo lo que en el metalenguaje podemos establecer como verdadero; por tanto, las oraciones del mentiroso reforzado cuestionan el que esto sea posible (y refutan de hecho que ello se haya conseguido en el caso del modelo de Kripke). En definitiva, las oraciones del mentiroso reforzado ponen en duda que la semántica trivaluada sea una maqueta fiel y lo más rica posible del comportamiento del predicado de verdad del lenguaje natural.

Hay muchas otras teorías matemáticas (y por tanto "insustantivistas") de la verdad en la tradición analítica reciente, que han buscado aprovechar las lecciones que dejaron las teorías de Tarski y Kripke en la construcción de modelos que superen los problemas que estas teorías enfrentan. (Sobre algunas de estas teorías, en español pueden verse Buacar y Picollo (2014), Martínez (2014), Rosenblatt y Pailos (2014) y Tajer (2014).) Pero también hay varias teorías de otros tipos. La situación es fluida y compleja, e intentar describirla panorámicamente pero con un detalle satisfactorio es una tarea complicada donde las haya, que habrá de quedar para otra ocasión.

Referencias

Barrio, E. A. (comp.) (2014a), *La Lógica de la Verdad*, Eudeba, Buenos Aires.

Barrio, E. A. (2014b), "Definiciones tarskianas de la verdad", en Barrio (2014a), 25-73.

Barrio, E. A. (2014c) (comp.), *Paradojas, Paradojas y más Paradojas*, College Publications, Londres.

Buacar, N. M. y L. M. Picollo (2014), "La teoría revisionista de la verdad", en Barrio (2014a), 117-186.

Gómez Torrente, M. (2001a), "Notas sobre el *Wahrheitsbegriff*, I", *Análisis Filosófico* 21, 5-41.

Gómez Torrente, M. (2001b), "Notas sobre el *Wahrheitsbegriff*, II", *Análisis Filosófico* 21, 149-185.

Hempel, C. G. (1935), "La teoría de la verdad de los positivistas lógicos", en Nicolás y Frápolli (1997), 481-493.

James, W. (1906), "Concepción de la verdad según el pragmatismo", en Nicolás y Frápolli (1997), 25-43.

Kripke, S. (1975), "Esbozo de una teoría de la verdad", en Nicolás y Frápolli (1997), 109-143.

Martínez, J. (2014), "La paradoja del mentiroso", en Barrio (2014c), 11-26.

Nicolás, J. A. y M. J. Frápolli (comps.) (1997), *Teorías de la Verdad en el Siglo XX*, Tecnos, Madrid.

Ramsey, F. P. (1927), "La naturaleza de la verdad", en Nicolás y Frápolli (1997), 265-279.

Rosenblatt, L. y F. Pailos (2014), "Paracompletitud sofisticada", en Barrio (2014a), 187-248.

Strawson, P. F. (1950), "Verdad", en Nicolás y Frápolli (1997), 281-307.

Tajer, D. (2014), "Dialeteísmo. Una teoría contradictorioa de la verdad", en Barrio (2014a), 249-292.

Tarski, A. (1944), "La concepción semántica de la verdad y los fundamentos de la semántica", en Nicolás y Frápolli (1997), 65-108.

Teijeiro, P. y D. E. Szmuc (2014), "Teoría de puntos fijos de Kripke", en Barrio (2014a), 75-116.