Tesis Doctoral

Estudio de la paradoja del mentiroso y afines

Serafín Benito Santos
Licenciado en Ciencias Físicas

Departamento de Lógica, Historia y Filosofía de la Ciencia
Facultad de Filosofía
Universidad Nacional de Educación a Distancia
Año 2007

Departamento:

Lógica, Historia y Filosofía de la Ciencia.

Facultad:

Filosofía (U.N.E.D.).

Título de la tesis:

Estudio de la paradoja del mentiroso y afines.

Autor y titulación:

Serafín Benito Santos, Licenciado en Ciencias Físicas.

Directora de la tesis:

Dra. D^{A}. Amparo Díez Martínez.

A mis padres.

Agradecimientos: A Amparo Díez por su dedicación, sabias orientaciones y paciencia mostradas en la dirección de la tesis, así como en el trabajo de investigación. A Feli, mi esposa, por su comprensión y tolerancia. A todos los profesores de los cursos de doctorado, que realicé con placer —Luis Vega, con quien además inicié el trabajo de investigación, Pilar Castrillo, Enrique Alonso y Amparo Díez—. También, mi agradecimiento a algunos autores —como Russell, Wittgenstein o Martin Gardner— que, mucho antes, aumentaron con sus libros mi gusto por la lógica y las paradojas.

Índice general

1

Introducción: el mentiroso, una paradoja no resuelta

El objetivo del presente trabajo es el estudio de la paradoja del mentiroso y afines. Quizá por ser una paradoja ya conocida en la antigua Grecia —una versión de la paradoja se atribuye a Epiménides[1] y otra a Eubúlides[2]—, sencilla de formular y, al mismo tiempo, resistente a una solución satisfactoria, se ha convertido en la más famosa de las paradojas de autorreferencia.

Fue Bertrand Russell, en el contexto de los intentos de formalización de la matemática de finales del siglo XIX y comienzos del XX, quien puso de manifiesto que diversas paradojas revelaban fallos básicos en principios aparentemente indiscutibles. Como señala Ferreirós (1999, p. 307), hasta entonces, aunque eran conocidas bastantes, no se consideraban un peligro serio para la formalización de las matemáticas. Hacia 1900, a pesar de que las paradojas del mayor cardinal y del mayor ordinal habían sembrado dudas sobre los fundamentos de la teoría de conjuntos, predominaba la idea de que el origen de los problemas estaba en las sofisticadas nociones de la teoría de conjuntos transfinitos y, por tanto, se pensaba que una revisión técnica de estas nociones permitiría solucionarlos.

La paradoja de Russell cambió la situación dado que no empleaba complejas nociones de la teoría de conjuntos de Cantor. Se basaba en nociones muy simples

[1] Vidente cretense del siglo VI a. C., autor de escritos religiosos y poéticos. Se le atribuye la afirmación "todos los cretenses son mentirosos". Es interesante observar que el que esta afirmación sea o no paradójica depende de hechos empíricos. Concretamente si algún cretense no es mentiroso, la afirmación es simplemente falsa.

[2] Filósofo megárico (siglo IV a. C.) nacido en Mileto. Su versión de la paradoja toma la forma: "un hombre dice que está mintiendo, ¿es cierto lo que dice o falso?"

(conjunto, relación de pertenencia, negación y todo). Como es bien sabido, en una carta que Russell escribió en 1902 a Frege, le indicó que su paradoja suponía "una dificultad" en el sistema axiomático con el que el lógico alemán pretendía fundamentar la aritmética. La reacción de Frege no dejó lugar a dudas, la paradoja era importante, no solo porque mostraba la inconsistencia de los fundamentos de su aritmética sino porque incluso "los únicos fundamentos posibles de la aritmética, parecen desvanecerse".[3] La mala fortuna quiso que Frege conociera la inconsistencia de su fundamentación de la aritmética justo cuando el segundo volumen de su *Grundgesetze der Arithmetik* (*Las leyes básicas de la aritmética*) estaba en prensa. No obstante, tuvo tiempo para añadir un apéndice a su obra discutiendo el descubrimiento de Russell. En este apéndice se aprecian dudas que afectan al programa logicista en su conjunto.

Las reacciones a la nueva situación planteada por la paradoja de Russell fueron diversas. El propio Russell al constatar que tanto esa paradoja como otras paradojas de autorreferencia no estaban relacionadas con las ideas de número y cantidad llegó a la conclusión de que lo que había que revisar no eran esas ideas sino la propia lógica en sus aspectos más básicos.[4] Tal vez por esta estrecha relación entre las paradojas de autorreferencia y principios lógicos considerados de sentido común, Russell prefería el término *contradicciones* para referirse a las paradojas. Sin embargo, esto no significa que pensara que no tenían solución. De hecho no se desanimó en su defensa del logicismo e ideó la teoría de los tipos para solucionar las *contradicciones*.

Un término probablemente más adecuado que el de contradicción es*antinomia*. Aunque, con seguridad, habrá diferentes interpretaciones del mismo, lo podríamos caracterizar como el resultado de llegar a dos conclusiones mutuamente contradictorias, partiendo de premisas aparentemente verdaderas y mediante razonamientos lógicos de apariencia impecable. Este término ha sido usado por Poincaré, Zermelo y Fraenkel.

El término más utilizado en la actualidad es el de paradoja. Si hubiera que señalar una diferencia con antinomia habría que decir que en la paradoja se pone más énfasis en que la corrección del argumento que lleva a contradicción es solo

[3] Tomado de un extracto de la carta de Frege a Russell del 22 de junio de 1902, citado en Garciadiego (1992b, p. 157).

[4] Véase por ejemplo, Whitehead y Russell (1927, p. 60).

aparente, aunque no se haya descubierto qué es lo que realmente falla para llegar a dicha contradicción.

En *Principia Mathematica* (en adelante, PM), Whitehead y Russell exponen siete *contradicciones* que conviene señalar brevemente.

- Del mentiroso: Alguien dice "estoy mintiendo", ¿es cierto lo que dice o falso? Quizás, la forma más frecuente de presentar actualmente el problema de esta paradoja es la siguiente: llamemos oración del mentiroso a "esta oración es falsa", ¿es falsa o no es falsa la oración del mentiroso? Parece indudable que o bien es falsa o bien no lo es, pero cualquier posibilidad lleva a la contraria.

- De Russell: Sea w la clase de todas las clases que no son miembros de sí mismas. ¿Es w miembro de sí misma? Como en la paradoja anterior, cualquier posible respuesta conduce a una contradicción.

- Sea T la relación entre dos relaciones R y S que se da cuando R no tiene la relación R con S. ¿Tiene T la relación T con T? Cualquier posible respuesta conduce a una contradicción.

- De de Burali-Forti. Se sabe que toda serie bien ordenada (como es el caso de la serie de todos los ordinales) tiene un número ordinal y que el ordinal de una serie de números ordinales es una unidad mayor que el mayor ordinal de la serie. ¿Cuál es el ordinal de la serie de todos los ordinales? Si lo llamamos Ω, la serie de todos los ordinales, que incluirá Ω, tendrá, como mínimo, el número ordinal $\Omega + 1$, luego Ω no es el ordinal de la serie de todos los ordinales.

- De Berry. "The least integer not nameable in fewer than nineteen syllables", es decir, el menor entero que no se puede nombrar, en inglés, con menos de diecinueve sílabas. ¿Existe? Por un lado, parece obvio que de todos los números que no se pueden nombrar en inglés con menos de diecinueve sílabas habrá uno que será el menor. Por otro, la frase entrecomillada muestra que dicho número se puede nombrar en inglés con dieciocho sílabas, lo cual es una contradicción.

- Del menor ordinal indefinible. ¿Existe? Se sabe que el número de posibles definiciones es menor que el número de ordinales transfinitos por lo que hay ordinales indefinibles. Puesto que los ordinales forman una serie bien

ordenada, debe haber un ordinal indefinible que sea el menor. Sin embargo, es definible mediante la expresión "el menor ordinal indefinible".

- Paradoja de Richard. Sea E la clase de todos los números decimales que pueden ser definidos mediante un número finito de palabras. E es numerable porque el conjunto de series finitas de palabras lo es. Sea N el número cuya enésima cifra es $(p+1)$ *mód* 10, siendo p la enésima cifra del enésimo número de E. ¿Pertenece N a E? No, porque al menos se diferencia en una cifra de cualquier número de E.[5] Sí porque N ha sido definido mediante un número finito de palabras.

Todas estas paradojas, según Russell, proceden de la violación de lo que denominó el principio del círculo vicioso: *lo que conlleva la totalidad de una colección no debe ser un miembro de la colección* o *ninguna totalidad puede contener miembros solamente definibles en términos de sí misma*. Como consecuencia, el argumento de una función no puede involucrar la propia función,[6] lo cual conduce a una jerarquía de funciones y proposiciones. Esa jerarquía tiene su desarrollo formal en la teoría de los tipos (ramificada).

De acuerdo con la teoría de los tipos, la paradoja del mentiroso se explica del siguiente modo. "Estoy mintiendo" supone que "hay una proposición p tal que afirmo p y p es falsa". Pero si p es una proposición de orden n, una proposición en la que p aparezca como variable aparente es de orden mayor que n. Por tanto, p no puede ser igual a "hay una proposición p tal que afirmo p y p es falsa". De esta forma desaparece la autorreferencia y, con ella, la contradicción.

Independientemente de las críticas que se pueden hacer a la solución de Russell a las paradojas, hay que reconocerle valiosos méritos. No solamente ofrece una solución formal, la teoría de los tipos, sino también un fundamento filosófico: el principio del círculo vicioso. Además es el primer intento serio de solucionar un grupo importante de paradojas, con la virtud de que una misma teoría se aplica a la solución de paradojas que más tarde han sido clasificadas en distintos grupos.

La clasificación más conocida de las paradojas de las que estamos tratando se debe a Ramsey (1925):

[5] Realmente esto no es suficiente pues, por ejemplo, $0,4 = 0,3999...$ aunque sus cifras no son iguales. Hay varias alternativas para que la diferencia en una cifra garantice que los números son diferentes; por ejemplo, tomar $(p+2)$ *mód* 10 como enésima cifra del número N.

[6] Si una expresión que denota un valor involucra la propia función, involucra la totalidad de sus argumentos y, por el principio del círculo vicioso, no debe formar parte de esa totalidad, es decir, no debe ser argumento de la función.

- Contradicciones lógicas. Las de Russell, Burali-Forti y la de la relación T entre dos relaciones, expuesta anteriormente. Solo precisan términos lógicos o matemáticos, como clase y número. Estas paradojas también han recibido el calificativo de sintácticas o conjuntistas.

- Contradicciones epistemológicas. Las del mentiroso, Berry, el menor ordinal indefinible, Richard y la de Weyl.[7] No son puramente lógicas y pueden ser debidas a ideas defectuosas respecto al pensamiento y al lenguaje. Actualmente se suelen denominar paradojas semánticas.

En cierto sentido, se puede decir que la paradoja del mentiroso es más puramente semántica que las de Berry, Richard o el menor ordinal indefinible, ya que no se ve afectada por conceptos matemáticos como número u ordinal. Resulta pues un buen prototipo para estudiar las paradojas semánticas. Consecuentemente, debe quedar claro que centrarse en el estudio de la paradoja del mentiroso no supone buscar una solución exclusiva para esa paradoja. Lo deseable sería que una solución a la misma abriera las puertas a una solución general de las paradojas semánticas e incluso a una solución general válida tanto para estas como para las conjuntistas.

Los dos autores más influyentes en el estudio de la paradoja del mentiroso han sido Tarski y Kripke.

Tarski se encuentra con la paradoja cuando afronta el problema de la definición de verdad Tarski (1933, 1969, 1999). Tras analizar los supuestos que conducen a la contradicción del mentiroso, llega a la conclusión de que una definición de verdad consistente solo es posible renunciando a utilizar para ello un lenguaje semánticamente universal o semánticamente cerrado.[8] Consecuentemente, distingue el lenguaje del que se habla (lenguaje objeto) y el lenguaje con el que hablamos del primero (metalenguaje). La definición de verdad que buscamos se expresa en el metalenguaje y se refiere al lenguaje objeto.

Así, si "X" es el nombre de una oración del lenguaje objeto, las oraciones "X es verdadera" y "X es falsa" pertenecerán al metalenguaje con el que hablamos del lenguaje objeto, pero no al propio lenguaje objeto.

[7] Normalmente llamada de Grelling: un adjetivo es heterológico si no se aplica a sí mismo. ¿Es heterológico el adjetivo heterológico?

[8] Un lenguaje que "contiene, además de sus expresiones, los nombres de dichas expresiones, así como los términos semánticos como el término "verdadero" para referirse a las oraciones de este lenguaje", Tarski (1999, p. 14).

La solución, o más bien disolución, resultante para la antinomia del mentiroso es sencilla. Si "L" denota una oración de un lenguaje, la oración "L es falsa" no pertenece a ese lenguaje sino a su metalenguaje y, consiguientemente, la oración denotada por "L" no puede ser "L es falsa".

La similitud con la solución de Russell es clara, pues, en ambos casos, hay una jerarquía que impide la autorreferencia.

No obstante las soluciones de Russell y Tarski comparten serios problemas que las hacen insatisfactorias (véase el apartado 2.1, p. 21 y ss.).

Kripke (1975) muestra que el predicado 'verdadero' es expresable en un lenguaje formal suficientemente rico como para expresar en él su propia sintaxis. Para ello admite *huecos* de valor de verdad, es decir, oraciones que no son verdaderas ni falsas ya que, de no ser así, el teorema de Tarski de indefinibilidad de la verdad lo haría imposible. Entre las oraciones que no son verdaderas ni falsas se encuentran las paradójicas. Una ventaja importante de la solución sugerida por Kripke es que, frente a la jerarquía de predicados 'verdadero$_1$', 'verdadero$_2$'..., solo hay un predicado 'verdadero' lo que está más en concordancia con el uso de este predicado en el lenguaje natural. A pesar de ello, la solución de Kripke, como él mismo reconoce, no se libra de la necesidad de usar un metalenguaje que está por encima del lenguaje formal que él construye, lo cual supone que su lenguaje formal no es verdaderamente universal. Y lo que es peor, como veremos (apartado 2.2, p. 28 y ss.) las paradojas reaparecen en ese metalenguaje.

Es por tanto fácilmente entendible que, después de Kripke, las propuestas de solución de las paradojas semánticas hayan seguido apareciendo. Indiquemos sucintamente algunas de las más destacables (incluidas algunas anteriores a la propuesta de Kripke):

- La propuesta de van Fraassen (1970) está basada en la teoría de las presuposiciones de Strawson que, a su vez, recoge una idea de Frege.[9] Se entiende que:

  A presupone B ssi A es verdadero o falso solo cuando B sea verdadero

  Para van Fraassen la oración del mentiroso ("lo que ahora digo es falso") presupone una contradicción, por lo que no es verdadera ni falsa.

[9] Frege (1984, p. 36).

- Otra interesante propuesta, recogida en Martin (1970b), es la del propio Martin (1970a). Se basa en que los predicados tienen rangos de aplicabilidad (RA). Si a no pertenece al rango de aplicabilidad de F, entonces ni Fa ni $\sim Fa$ son semánticamente correctas, por lo que ninguna tiene valor de verdad. Un ejemplo sería "la virtud es triangular". En el caso de una oración autorreferencial, la determinación de la corrección semántica es algo más complejo. En primer lugar, se establece:

  [M] RA(verdadero) = RA(falso) = conjunto oraciones con valor de verdad

  Después, se establece la prueba de corrección semántica para las oraciones autorreferenciales:

  [SCA] Fa es semánticamente correcta si y solo si la referencia demostrativa del término sujeto de la oración $\in RA(F)$

  Para poder afirmar que una oración tiene valor de verdad es necesario, según Martin, que pase la prueba de corrección semántica. Pero la oración del mentiroso ("esta oración es falsa"), que llamaré (L), no pasa la prueba. Para ello, la referencia demostrativa de su término sujeto, también (L), debería pertenecer a RA(falso). Pero, según [M], para que (L) pertenezca a RA(falso), (L) debe tener un valor de verdad. Surge así lo que Martin denomina un "regreso infinito" porque para que (L) pase la prueba debe tener un valor de verdad, es decir, debe haber pasado previamente la prueba.

  Así pues, no hay manera de que la oración del mentiroso pase la prueba, es decir, se trata de una oración semánticamente incorrecta, una oración sin valor de verdad.

- Para McGee y otros la noción de verdad hereda la vaguedad de los predicados no semánticos. "Juan es alto" puede no ser completamente cierto ni completamente falso y, consiguientemente, lo mismo puede decirse de "'Juan es alto' es verdadero". Habría pues tres tipos de oraciones: claramente verdaderas, claramente falsas y las que no son una cosa ni otra. Por supuesto, (L) pertenecería a este tercer tipo.

- Frente a las propuestas anteriores que admiten oraciones que no son verdaderas ni falsas, otros, como Graham Priest, han defendido soluciones *inconsistentes* en que las oraciones paradójicas son a la vez verdaderas y falsas.

- Un tipo de propuesta que ha tenido considerable aceptación en las últimas décadas se basa en considerar que la extensión del término "verdadero", como la de "aquí" o "ahora", depende del contexto en que se emite la oración que lo contiene. Entre los defensores de propuestas de este tipo podemos mencionar a Burge, Charles Parsons, Barwise y Etchemendy (1987) y Simmons (1993).

- Tomaré a Hansson (1978) como ejemplo de una interpretación de las oraciones (directa o indirectamente) autorreferenciales basada en ecuaciones. Para él, interpretar una oración es responder a la pregunta acerca de qué proposición satisface la condición que la oración establece. En el caso de una oración autorreferencial de la forma "esta oración es P" la condición que establece es "x=P(x)". Como en una ecuación matemática puede haber una, ninguna o varias soluciones. En el primer caso estamos ante una oración no paradójica, en el segundo ante una paradoja del tipo de (L), en el tercero ante una del tipo de (TT).[10]

- Gupta, Herzberger y Belnap defienden una teoría de la revisión de la verdad. Un resumen informal y muy conciso de esta teoría tal como aparece en el capítulo 4 de Gupta y Belnap (1993) es el siguiente. En primer lugar se destaca que el comportamiento patológico del concepto de verdad es análogo al comportamiento de los conceptos que aparecen en definiciones circulares. Por tanto, se centran en este tipo de definiciones para las cuales, igual que para las que no son circulares, aceptan que la definición fija completamente el significado de lo definido. Sin embargo, generalmente, una definición circular no determina la extensión de lo definido. En cambio, sí se puede determinar esa extensión una vez que hacemos una hipótesis sobre la extensión de lo definido. De esta forma hipotética es como Gupta y Belnap entienden el significado de una definición circular. Esta no proporciona una extensión de lo definido pero sí una *regla de revisión*, una regla que, aplicada a una extensión hipotética nos proporciona un mejor (o igualmente buen) candidato

[10] "(TT)" es el nombre de la oración "esta oración es verdadera" a la que podemos denominar oración del veraz (*truth teller sentence* en la terminología anglosajona).

para la extensión de lo definido. Aplicando reiteradamente la regla de revisión a todas las hipótesis posibles se tiene una secuencia de extensiones. En las *oraciones* ordinarias, cualquiera que sea la hipótesis inicial, la extensión de lo definido acaba siendo la misma; en las patológicas esto no sucede. Por ejemplo la aplicación reiterada de la regla de revisión a $L =_{def}$ *L es falsa* daría las series (T, F, T, F...)[11] y (F, T, F, T...) que no se estabilizan en ningún valor; por tanto, L es patológica.

- Skyrms (1970) llega a afirmar que el principio de la sustitución de los idénticos es incorrecto en el sentido de que al aplicarlo a premisas verdaderas no garantiza conclusiones verdaderas.[12] Para ello parte de la paradoja del mentiroso reforzado que él expresa mediante la identidad "$a =$'$\sim Ta$"'. Además considera que '$\sim Ta$' no es verdadero ni falso, por lo que es verdadero

$$\sim T \text{ '} \sim Ta\text{'}$$

De aquí, por el principio de la sustitución de los idénticos, se obtendría

$$\sim Ta$$

Luego, aplicado a una premisa verdadera, este principio nos ha llevado a una conclusión que no es verdadera (ni falsa). Según Skyrms, dicho principio es válido en el sentido, débil, de que, aplicado a una premisa verdadera, garantiza que la conclusión no será falsa por lo que, si nos restringimos a una lógica bivalente, el principio funciona en el sentido fuerte. La explicación de que no funcione en el ejemplo anterior la basa en que el principio nos lleva de una oración no autorreferencial a una autorreferencial carente de significado.

Después de tantos intentos, continúa la sensación de que la paradoja del mentiroso (y las semánticas en general) no está satisfactoriamente resuelta. Da la impresión de que, si hay quien recurre a aceptar oraciones verdaderas y falsas a la vez y quien recurre a afirmar que el principio de la sustitución de los idénticos aplicado a premisas verdaderas no garantiza conclusiones verdaderas, la situación es un tanto

[11] Aquí 'T' representa 'verdadero' y 'F', 'falso'.

[12] Por supuesto también es incorrecto en entrecomillados, en contextos modales y epistémicos, pero el autor aclara que no se refiere aquí a estos casos.

desesperada. A esa impresión pueden contribuir también las interpretaciones de Tarski que afirman que el lenguaje natural es inconsistente u opiniones como la de Herzberger acerca de que hay conceptos semánticos que el lenguaje natural no puede expresar.

Putnam (2000a) va más allá al afirmar que no habrá una solución a las paradojas en el sentido de "a point of view that simply makes all appearance of paradox go away".[13] Unas páginas después lo corrobora con las siguientes palabras:

> If you want to say something about the Liar, in the sense of being able to finally answer the question "Is it meaningful or not? And if it is meaningful, is it true or false? Does it express a proposition or not? Does it have a truth value or not? And which one?" then you will always fail. [...] the totality of our desires with respect to how a truth predicate should behave in a semantically closed language, in particular our desire to be able to say, without paradox, of an arbitrary sentence in such a language that it is true, or that it is false, or that it is neither true nor false, cannot be adequately satisfied.[14]

Sin embargo, las argumentaciones de Putnam no son una demostración rigurosa de que paradojas como la del mentiroso no puedan resolverse. Si hubiera tal demostración el problema podría considerarse resuelto, aunque de un modo negativo. Como no es así, tiene sentido seguir indagando sobre el problema buscando aportaciones que puedan ser clarificadoras del mismo. En este trabajo pretendo iniciar mis indagaciones partiendo de un análisis de los problemas que aquejan a las principales propuestas de solución. Espero que este análisis sirva para establecer los requisitos que debe cumplir una solución satisfactoria, las pruebas que debe superar y que no superan otras *soluciones*. Posteriormente, intentaré caracterizar el problema con la ayuda de lenguajes formales y formular algunas propuestas positivas que superen los requisitos antes mencionados.

[13] Putnam (2000a, p. 6).
[14] Ibíd. p. 14.

2

Principales propuestas de solución y sus problemas

2.1. Soluciones jerárquicas de Russell y Tarski

La primera versión de la teoría de los tipos la expone Russell en el apéndice B de su obra de 1903 *The principles of mathematics*. Se basa en dos postulados: 1º/ toda función proposicional tiene un rango de significado, es decir el rango al que deben pertenecer los argumentos para que la función tenga un valor; 2º/ un tipo es el rango de significado de una función proposicional. A partir de aquí se genera una jerarquía de tipos que comienza por los *individuos* (objetos que no son rangos) y continúa con las clases de individuos, las clases de clases de individuos, etc. Esta es, de modo muy sucinto, la que después se ha denominado teoría simple de los tipos.

La teoría simple resuelve la paradoja de Russell pero tiene problemas con otras como la del mentiroso. La función proposicional "p es verdadero" tiene, en principio, como rango de significado todas la proposiciones, por lo que todas las proposiciones pertenecerían al mismo tipo. Sin embargo, dadas una función proposicional cuyos argumentos son individuos y otra cuyos argumentos son clases de individuos, las proposiciones que se obtienen al sustituir los argumentos por valores, pertenecerán a tipos distintos.

En PM, se solucionan algunas dificultades de la teoría simple de tipos mediante la aplicación del principio del círculo vicioso. Aplicado a funciones proposicionales, puede expresarse diciendo que los valores de cualquier función proposicional que

tenga una variable ligada están excluidos del ámbito de valores posibles de esa variable. Por ejemplo, "(x) fx" no puede ser argumento de la función f. Todo esto conduce a una jerarquía de significados de las palabras "verdadero" y "falso" que así, sin más, son ambiguas. El razonamiento[1] se basa en: 1º/ dada una función proposicional hay siempre una proposición (verdadera o falsa) que afirma todos sus valores; 2º/ teniendo en cuenta tanto el punto 1º/ como que todas las proposiciones no son falsas, si "p es falsa" fuese una función proposicional, podríamos decir que "(p) p es falsa" es una proposición falsa; 3º/ pero, por el principio del círculo vicioso no es posible decir "{(p) p es falsa} es falsa"; 4º/ por eso, dicho principio nos lleva a la conclusión de que "p es falsa" no es una función proposicional: "the word "false" really has different meanings, appropriate to propositions of different kinds"[2]. Si a la proposición fa, obtenida de la función fx, le es aplicable el predicado *verdadera de orden 1*, a "(x) fx" no le será aplicable, por el principio del círculo vicioso, pero sí podremos decir que es *verdadera o falsa de orden 2*.

Más adelante, Russell y Whitehead desarrollan una jerarquía de funciones proposicionales (y proposiciones) en distintos *órdenes*. El principio del círculo vicioso nos lleva a ella. En una función de primer orden sus variables, libres o ligadas, corresponden a individuos. Una función es de orden $n+1$ si alguna de sus variables, libre o ligada, es una función de orden n y ninguna de orden mayor que n. Como, por otra parte, el tipo de una función viene determinado por el tipo de sus argumentos, se puede considerar que dentro de cada tipo hay diversos órdenes.[3] De ahí que a esta forma, más compleja, de la teoría de los tipos, se la haya denominado teoría ramificada de los tipos.

Los comentarios a la teoría de los tipos han sido numerosos. El que más unanimidad suscita es la crítica de que es demasiado restrictiva: no solamente evita las *peligrosas* paradojas sino también otras partes de las matemáticas inocuas y, a veces, importantes como la prueba de la infinitud de los números naturales o la definición de un número real mediante las cortaduras de Dedekind.

Los propios autores de PM reconocen que:

[1] Véase el apartado III, capítulo II de la Introducción a PM.

[2] Ibíd., p. 42.

[3] Una función de orden $n+k$ en la que k es el mayor orden de sus argumentos, se dice que es de orden n, relativo al de sus argumentos. Así entre las funciones con los mismos argumentos se distinguen distintos órdenes.

> It is possible that the use of the vicious-circle principle, as embodied in the above hierarchy of types, is more drastic than it need be[4]

La solución de Russell para este problema consiste en introducir el axioma de reducibilidad, por el que, dada una función cualquiera siempre hay una función predicativa[5] formalmente equivalente.

Las críticas al axioma de reducibilidad se pueden calificar de abrumadoras. De acuerdo con Ferreirós (1999, p. 349), ha sido considerado inaceptable por Weyl, Wittgenstein, Hilbert, Ramsey, Gödel, Waismann y Quine. El mismo Russell se sentía incómodo con la introducción del axioma. Dos de las principales razones en contra de él son las siguientes. En primer lugar, echa por tierra la pretensiones logicistas de PM pues, como señaló Ramsey, ese axioma es "a genuine proposition, whose truth or falsity is a matter of brute fact, not of logic".[6] De otra parte, al ser reducible cualquier función a una función de orden 1, la distinción de órdenes dentro de cada tipo resulta superflua lo que, en buena medida, convierte también en superfluo el principio del círculo vicioso del que deriva.

Entre otras críticas sufridas por la teoría de los tipos de PM, merece destacarse que el sistema no se libra realmente de la autorreferencia. Como ejemplo,

> hay que asegurar una afirmación como «todas las variables cuantificadas de nivel n tienen como recorrido las propiedades de nivel n», lo cual no es otra cosa que una afirmación general sobre las propiedades de todos los niveles bajo la cual cae la propia afirmación.[7]

Diversos lógicos intentaron elaborar nuevas versiones de la teoría de los tipos que mejorasen la versión de PM. Para Ramsey era fundamental la distinción entre paradojas lógicas y epistemológicas. Las primeras se resolvían mediante el principio de que una función proposicional no puede ser, significativamente, argumento de sí misma. Con otras palabras, se resuelven mediante la teoría simple de tipos. Las segundas, al no ser puramente lógicas, deben explicarse en función de ideas erróneas respecto al pensamiento y al lenguaje. ¿Cómo, si no, entender que el axioma de reducibilidad no reproduzca las paradojas al evitar la distinción entre funciones elementales (predicativas en PM) y no elementales?

[4] Whitehead y Russell (1927, p. 60).
[5] Una función es predicativa si es de orden 1 (relativo al de sus argumentos).
[6] Ramsey (1925, pp. 174-5).
[7] Lorenzo (1998, p. 165).

> This is not, however, the case, owing to the peculiar nature of the contradictions in question; for, as pointed out above, this second set of contradictions are not purely mathematical, but all involve the ideas of thought or meaning, in connection of which equivalent functions [...] are not interchangeable; for instance, one can be meant by a certain word or symbol, but not the other, and one can be definable, and not the other.[8]

La idea de que la teoría simple de tipos era suficiente para evitar las paradojas en un sistema lógico fue ampliamente aceptada.

En cuanto a las paradojas semánticas es fundamental destacar el estudio de Tarski del concepto de verdad en los lenguajes formalizados, que le permitió proponer una solución a dichas paradojas.

Tarski se plantea el problema de encontrar una definición satisfactoria de verdad, es decir, que sea materialmente adecuada y formalmente correcta. Una definición de la verdad será materialmente adecuada cuando todas las equivalencias de la forma [T]:

[T] X es verdadera si y solo si p

se sigan de ella. Se entiende que para convertir el esquema [T] en una equivalencia concreta, "p" debe ser sustituido por una oración y "X" por el nombre de esa oración.

El problema de la corrección formal consiste, esencialmente, en describir la estructura formal del lenguaje en el que daremos la definición. Para descubrir las condiciones que ha de cumplir un lenguaje en el que establecer una definición de verdad, Tarski analiza la antinomia del mentiroso. El análisis se puede resumir del siguiente modo: sea "L" el nombre de la oración "L es falsa"; de la aplicación del esquema [T] resulta la contradicción

L es verdadera si y solo si L es falsa

Es pues claro que el lenguaje nos conducirá a esta contradicción si permite construir una oración L que afirme su propia falsedad. Así ocurre en los lenguajes universales (como el inglés o el español) o semánticamente cerrados, es decir,

[8] Ramsey (1925, p. 191).

lenguajes que contienen, además de sus expresiones: a) los nombres de dichas expresiones; b) predicados semánticos, como "verdadero", "nombre", "designación", para referirse a las oraciones de este lenguaje. En consecuencia

> tenemos que utilizar dos lenguajes diferentes al discutir el problema de la definición de la verdad y, más generalmente, cualquier tipo de problema en el campo de la semántica. El primero de estos lenguajes es el lenguaje "del que se habla" y que es el objeto de esta discusión; la definición de la verdad que buscamos se aplica a las oraciones de este lenguaje. El segundo es el lenguaje "con el que hablamos" del primer lenguaje, y en términos del que nos gustaría, en concreto, construir la definición de verdad para el primer lenguaje. Denominaremos al primer lenguaje "*lenguaje objeto*" y al segundo "*metalenguaje*".[9]

Tarski añade que los términos "lenguaje objeto" y "metalenguaje" tienen únicamente un sentido relativo. La definición de verdad será relativa a un lenguaje determinado; si queremos definir la verdad en su correspondiente metalenguaje, tendremos que utilizar un nuevo metalenguaje de un nivel más elevado para hacerlo. Nos encontramos así con toda una jerarquía de lenguajes.

También estudia las características que debe reunir un metalenguaje. Puesto que en el metalenguaje pretendemos expresar una definición materialmente adecuada de la verdad y esta debe tener como consecuencias todas las equivalencias del esquema [T], en el que "p" debe reemplazarse por una oración cualquiera del lenguaje objeto y "X" por el nombre de esa oración; el metalenguaje debe contener (una traducción de) toda oración del lenguaje objeto y un nombre para cada una de esas oraciones. Tarski (1983a, pp. 210-1) distingue tres clases de expresiones en el metalenguaje: (1) expresiones de carácter lógico general; (2) expresiones con el mismo significado que las constantes del lenguaje a discutir, lo que permitirá traducir toda oración del lenguaje al metalenguaje; (3) expresiones específicas del metalenguaje de carácter descriptivo-estructural que denotan signos y expresiones del lenguaje objeto, clases y secuencias de tales expresiones y relaciones existentes entre ellas; este tercer tipo de expresiones permiten asignar un nombre individual a las correspondientes expresiones del lenguaje objeto. Utilizando expresiones de las clases (1) y (2) así como signos y expresiones del lenguaje objeto, introduce los demás términos *descriptivo-estructurales* (función oracional, oración, axioma,

[9] Tarski (1999, p. 15).

oración demostrable...) mediante definiciones. Finalmente, define el concepto de verdad basado en el de satisfacción.

> Una vez que se ha obtenido la definición general de satisfacción, nos damos cuenta de que es aplicable automáticamente también a las funciones predicativas especiales que no contienen variables libres, por ejemplo, a las oraciones. Resulta que para una oración solo son posibles dos casos: o bien todos los objetos satisfacen una oración, o bien ningún objeto satisface dicha oración. De esta forma, llegamos a la definición de verdad y falsedad diciendo que *una oración es verdadera si todos los objetos la satisfacen y es falsa si ningún objeto la satisface*[10]

Tarski destaca que esta definición solo es posible si el metalenguaje es *esencialmente más rico* que el lenguaje objeto. Que no se cumpla esta condición supone que sería posible la interpretación del metalenguaje en el lenguaje objeto y, por tanto, reconstruir en aquel lenguaje la antinomia del mentiroso. Así ocurriría por ejemplo, en el lenguaje de la aritmética o un lenguaje que lo incluya, en los que sería de aplicación el conocido teorema de la indefinibilidad de la verdad.

Aunque las principales conclusiones de Tarski en lo referente a la definición de la verdad en los lenguajes formalizados son indiscutibles desde el punto de vista técnico, no parece que dichas conclusiones supongan una teoría de la verdad satisfactoria ni, en particular, que solucionen el problema de las paradojas en el lenguaje natural. Algunas de las principales críticas recibidas son:

- La jerarquía de lenguajes, y la correspondiente jerarquía del concepto de verdad, que sirve para evitar la paradoja, no tiene una justificación independiente de su utilidad a este respecto. Antes al contrario, en los lenguajes naturales solo hay una palabra "verdadero" y no una secuencia de *verdaderos* de distintos niveles.

- Hay aplicaciones globales de "verdadero" que la teoría de Tarski no puede representar. Ejemplo: "toda proposición es verdadera o no".

- Las oraciones no son o no paradójicas exclusivamente en función de su significado sino que pueden serlo en función de hechos empíricos.[11]

[10] Tarski (1999, p. 18).

[11] Volveremos a esta observación al estudiar la propuesta de Kripke.

- La condición de adecuación material excluye las teorías de la verdad en que las oraciones pueden no ser verdaderas ni falsas. Si en el esquema [T] sustituimos "p" por una oración no verdadera ni falsa y "X" por el nombre de esa oración, obtenemos la afirmación de una equivalencia entre "X es verdadera", que es falsa, y "p" que no es verdadera ni falsa.

Por otra parte, es interesante observar las importantes similitudes entre la solución de Russell y la de Tarski de la paradoja del mentiroso. Por distintos caminos el concepto de verdad queda jerarquizado en ambas propuestas y, en ambos casos, la frase (L), "esta oración es falsa", no puede expresarse porque (L) y "(L) es falsa" son de distinta clase (Russell diría que son de distinto orden; y Tarski, que pertenecen a distintos lenguajes). Church ha comparado las soluciones de ambos[12] y ha llegado a la conclusión de que la solución de Russell a las paradojas semánticas es un caso especial de la solución de Tarski.

Merecen destacarse, para terminar, algunos problemas comunes a las soluciones de Russell y Tarski y a las jerárquicas en general:

- Son demasiado restrictivas. Prohíben o limitan excesivamente la autorreferencia. Esto resulta especialmente claro en la solución de Russell que no permitiría muchas autorreferencias *inocuas*. Por ejemplo, tomemos la definición, "Y $=_{def}$ el mayor de los habitantes de Madrid". Al definir Y en términos de una totalidad que contiene al propio Y, esta definición viola el principio del círculo vicioso cuando, en realidad, no es una definición problemática.

- La simple afirmación de que la jerarquía existe, o cualquier afirmación sobre la totalidad de niveles de la jerarquía,[13] está fuera de ella. Entonces, ¿qué lugar ocupa esa afirmación? Si la respuesta es que ese tipo de afirmaciones solo es posible en el lenguaje natural, la solución jerárquica desplaza el problema de las paradojas a este lenguaje, pero no lo resuelve. La respuesta alternativa, a saber, que ese tipo de afirmaciones es inexpresable, tampoco es satisfactoria, porque entonces no se podría exponer en qué consiste la teoría ni hacer afirmaciones tan sencillas como "toda proposición es verdadera o falsa".

[12] "Comparison of Russell's Resolution of the Semantical Antinomies with that of Tarski", 1976. Reimpreso en Martin (1984).

[13] Por ejemplo, "toda proposición es verdadera o no es verdadera".

- La jerarquización del predicado verdadero (o del predicado falso) en niveles (verdadero$_1$, verdadero$_2$...) es problemática y artificiosa. Como señala Kripke (1975, p. 58):

  > Surely our language contains just one word 'true', not a sequence of distinct phrases $\ulcorner \text{true}_n \urcorner$, applying to sentences of higher and higher levels.

quien, en las páginas siguientes, muestra la dificultad e incluso imposibilidad de determinar el nivel del predicado "falso" en ciertas oraciones cuyo valor de verdad, dados ciertos hechos, es claro. Siguiendo las ideas de Gödel, podemos señalar otra dificultad en la jerarquización del predicado verdadero, considerando la oración "para cada número natural n, existe un predicado *verdadero*$_n$". En una teoría jerárquica esto es verdadero pero no se puede decir que sea *verdadero*$_k$ para ningún número natural k, como exigiría la propia teoría jerárquica. Tampoco sirve permitir números transfinitos como subíndices de *verdadero*, porque el razonamiento anterior seguiría siendo válido cambiando "número natural" por "número finito o transfinito".

2.2. Propuesta de Kripke

La imposibilidad de definir la verdad, de acuerdo con el esquema [T] de Tarski, en un metalenguaje que no sea *esencialmente más rico* que el lenguaje objeto supone una limitación muy importante a tener en cuenta en ulteriores elaboraciones de una teoría de la verdad satisfactoria y que, por tanto, dé explicación del concepto de verdad en lenguajes universales como el lenguaje natural.

El más influyente de los intentos de escapar a esa limitación está expuesto por Kripke en su "Outline of a Theory of Truth".[14] De acuerdo con la mayoría de las alternativas a la solución de Tarski, entre las que destacan las de Bas van Fraassen y Robert L. Martin recogidas en Martin (1970b), Kripke sostiene que la solución a las paradojas semánticas pasa por reconocer la necesidad de admitir oraciones sin valor de verdad, entre las que se encuentran las oraciones paradójicas.

Otro aspecto al que Kripke concede gran importancia es que no hay ninguna característica sintáctica o semántica que determine si una oración es paradójica o no. Así, la oración enunciada por Jones:

[14] Kripke (1975).

(1) La mayoría de las afirmaciones de Nixon sobre el Watergate son falsas

está bien formada y no es, en principio, problemática. Pero supóngase que la mitad de las afirmaciones de Nixon sobre el Watergate son verdaderas y la otra mitad falsas sin contar su afirmación siguiente:

(2) Todo lo que dice Jones sobre el Watergate es verdadero

y que, además, la única oración de Jones sobre el Watergate es (1). En este caso, (1) y (2) resultan paradójicos.[15] Queda claro, por tanto, que unos hechos desfavorables pueden hacer que unas oraciones de apariencia inocua resulten paradójicas.

Kripke también critica la jerarquía de lenguajes de Tarski por su dificultad de ser aplicada al lenguaje natural. No hay problemas si tomo una oración como "la nieve es blanca" y le atribuyo un valor de verdad usando el predicado "verdadero$_1$"; el predicado "verdadero$_2$" puedo aplicarlo a oraciones que contienen "verdadero$_1$", etc. Pero consideremos la siguiente oración, enunciada por Dean:

(3) Todas las afirmaciones de Nixon sobre el Watergate son falsas

Para empezar, no es posible asignar un nivel a esta oración en función de sus características sintácticas o semánticas, dado que depende de hechos empíricos relacionados con las afirmaciones de Nixon sobre el Watergate. La dificultad de asignar un nivel a la oración (3) se convierte en imposibilidad si Nixon afirma algo sobre las afirmaciones de Dean, porque la mutua referencia supondría que la afirmación de Nixon es de un nivel superior a la de Dean y viceversa. Y esto ocurrirá aunque el valor de verdad de ambas afirmaciones sea claro. Por ejemplo, si todo lo que Nixon ha dicho sobre el Watergate es falso y además dice que

(4) Todo lo que Dean diga sobre el Watergate es falso

y Dean ha hecho alguna afirmación cierta sobre el Watergate, diferente a (3), es claro que (4) es falsa y (3) es verdadera.

Un concepto clave para Kripke es el de oración fundamentada (*grounded sentence*). El problema de determinar el valor de verdad de una oración en la que aparece la noción de verdad se intenta resolver reduciéndolo a otra oración. Por ejemplo, el problema de determinar el valor de verdad de una oración de la forma

[15] Si (1) es verdadero, (2) ha de ser falso, de donde (1) resulta ser falso; etc.

"A es falsa" se reduce fácilmente a determinar el valor de verdad de A. Pero puede que en A vuelva a aparecer la noción de verdad y el proceso deba continuar. Si al final del proceso encontramos una oración en que no interviene el concepto de verdad (como "la nieve es blanca"), la oración original está fundamentada y será verdadera o falsa. Sin embargo, hay casos en que es fácil ver que el proceso no termina nunca. Tal es el caso de las oraciones

(TT) (TT) es verdadera

(L) (L) es falsa

que, aunque se siguen considerando significativas, no expresan una proposición.

Kripke introduce un esquema semántico para manejarse con predicados parcialmente definidos. Dado un dominio, D, un predicado $P(x)$ se interpreta mediante un par de subconjuntos de D disjuntos, (S_1, S_2). S_1 es la extensión de $P(x)$, y S_2 es la antiextensión. $P(x)$ es verdadero si el elemento denotado por x pertenece a S_1, falso si pertenece a S_2 e indefinido en otro caso.

Para manejar las conectivas lógicas, considera adecuada la "lógica trivaluada fuerte" de Kleene.[16] Resulta, por ejemplo, que la disyunción "(la nieve es blanca) $\vee$ (L)" es verdadera. Esto podría suponer un problema si (L) se considerase no significativa, pues la disyunción anterior, presumiblemente, no sería significativa. Pero el problema no surge, porque Kripke considera que una oración bien formada es significativa independientemente de que resulte paradójica o no.

La plasmación formal de las ideas anteriores se atiene a lo siguiente. Se parte de un lenguaje de primer orden, L, interpretado mediante un dominio no vacío, D, que tiene un conjunto finito (o infinito enumerable) de predicados interpretados mediante relaciones totalmente definidas. Se supone también que las expresiones de L se pueden codificar en D. Al añadir un nuevo predicado, $T(x)$, L se convierte en $\mathcal{L}$. La interpretación de L que resulta de interpretar $T(x)$ mediante una extensión S_1 y una antiextensión S_2, se denomina $\mathcal{L}(S_1, S_2)$.

Kripke genera una jerarquía de interpretaciones que solo se diferencian en la extensión y antiextensión del predicado $T(x)$. $\mathcal{L}_0 = \mathcal{L}(\Lambda, \Lambda)$, siendo Λ el conjunto vacío. Dado un entero, α, si $\mathcal{L}_\alpha = \mathcal{L}(S_{1,\alpha}, S_{2,\alpha})$, entonces $\mathcal{L}_{\alpha+1} = \mathcal{L}(S_{1,\alpha+1}, S_{2,\alpha+1})$, donde $S_{1,\alpha+1}$ es el conjunto de (los códigos de) las oraciones verdaderas de $\mathcal{L}_\alpha$ y

[16] Aunque señala que se pueden usar igualmente otros esquemas para manejar los huecos de verdad (*true gaps*).

$S_{2,\alpha+1}$ es el conjunto de elementos de D que no son (códigos de) oraciones de $\mathcal{L}_\alpha$ o son (códigos de) oraciones falsas de $\mathcal{L}_\alpha$.

Seguidamente, demuestra que, para cualquier α, la interpretación de $T(x)$ en $\mathcal{L}_{\alpha+1}$ extiende la interpretación de $T(x)$ en $\mathcal{L}_\alpha$, es decir, $S_{1,\alpha} \subseteq S_{1,\alpha+1}$ y $S_{2,\alpha} \subseteq S_{2,\alpha+1}$. Por tanto,

> *the predicate $T(x)$ increases, in both its extension and its antiextension, as α increases. More and more sentences get declared true or false as α increases; but once a sentence is declared true or false, it retains its truth value at all higher levels.*[17]

Incluso ampliando la jerarquía a niveles transfinitos, sigue siendo cierto que la extensión y la antiextensión de $T(x)$ no decrecen cada vez que se aumenta α. Pero lo que es más interesante es que se puede probar que habrá un nivel ordinal, σ, tal que $(S_{1,\sigma}, S_{2,\sigma}) = (S_{1,\sigma+1}, S_{2,\sigma+1})$, o, lo que es lo mismo, tal que $(S_{1,\sigma}, S_{2,\sigma})$ es un punto fijo. Y esto significa que la interpretación de $T(x)$ en $\mathcal{L}_\alpha$ es una interpretación especialmente adecuada si queremos que $T(x)$ se interprete como "x es verdadero".

Aquellas oraciones que tienen un valor de verdad en el menor punto fijo, $\mathcal{L}_\alpha$, son las fundamentadas. Toda oración fundamentada, A, tiene un nivel que es el menor ordinal, α, tal que A tiene un valor de verdad en $\mathcal{L}_\alpha$.

A continuación Kripke muestra las ventajas de su teoría. Es fácil determinar que oraciones como (TT) o (L) no están fundamentadas. El nivel de una oración depende de hechos empíricos. Es fácil distinguir las oraciones no fundamentadas paradójicas de las que no lo son: las paradójicas no tienen valor de verdad en ningún punto fijo; las no paradójicas, como (TT), tienen un valor de verdad en un punto fijo si tienen un valor de verdad en algún nivel anterior, aunque para ello la jerarquía de interpretaciones de $\mathcal{L}$ no deba comenzar en $\mathcal{L}(\Lambda, \Lambda)$. Etc.

Las virtudes de la teoría propuesta por Kripke son meritorias, pero el propio autor reconoce alguna de sus limitaciones. Si bien se consigue un lenguaje con su propio predicado de verdad, no se trata de un lenguaje verdaderamente universal:

> First, the induction defining the minimal fixed point is carried out in a set-theoretic metalanguage, not in the object language itself. Second, there are assertions we can make about the object language which we

[17] Kripke (1975) en Martin (1984, p. 68) (cursiva en el original).

> cannot make in the object language. For example, Liar sentences are *not true* in the object language [...] but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate [...] The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us.[18]

Por último, también reconoce que nociones introducidas por su teoría como "fundamentada", "paradójica", etc. pertenecen al metalenguaje y no al lenguaje objeto. Curiosamente, esta situación es considerada aceptable por Kripke argumentando que esos conceptos no se encuentran en el lenguaje natural antes de que los filósofos piensen en su semántica.

Seguramente el principal inconveniente de la teoría de Kripke, que él mismo reconoce, es la necesidad de ascender a un metalenguaje para referirse a su propia jerarquía de interpretaciones de $\mathcal{L}$. Si bien se consiguen mejoras respecto a la propuesta de Tarski, no acaba siendo una teoría satisfactoria del concepto de verdad en un lenguaje verdaderamente universal. Las paradojas se han resuelto en el lenguaje pero, en el metalenguaje, pueden reaparecer. Como ocurre, en general, en las teorías de la verdad que admiten oraciones que no son verdaderas ni falsas (llamémosle indefinidas), queda sin resolver la "paradoja del mentiroso reforzada":[19]

(SL) (SL) no es verdadera

Es fácil ver que si la oración (SL) es verdadera, entonces '(SL) no es verdadera' es falsa, de donde, (SL) es falsa; si (SL) es falsa o indefinida, entonces no es verdadera, pero, eso es lo que afirma (SL), y por tanto, es verdadera. En cualquier caso se llega a contradicción. En la propuesta de Kripke, (SL) tiene que expresarse en el metalenguaje dado que '$\sim$ T(x)' significa, en su lenguaje interpretado, 'x es falsa', mientras que, en (SL), la expresión 'no es verdadera' quiere decir 'es falsa o indefinida' y el predicado 'ser indefinido' sólo puede expresarse en el metalenguaje. En definitiva, la paradoja del mentiroso reforzada aparece en el metalenguaje.

Incluso la paradoja simple del mentiroso también aparece en el metalenguaje. Suele considerarse que dicha paradoja se resuelve diciendo que la oración (L) no es

[18] Ibíd., pp. 79-80.

[19] En la terminología anglosajona, *strengthened liar paradox.*

verdadera ni falsa y después se añade que, sin embargo, la paradoja del mentiroso reforzada no puede solucionarse con los *huecos de verdad*. Pero desde el momento en que aceptamos que hay oraciones no verdaderas ni falsas y usamos un lenguaje donde esto sea expresable ¡la paradoja simple del mentiroso vuelve a aparecer! Si suponemos que (L) es indefinida (ni verdadera ni falsa), entonces (L) no es falsa, y como (L) consiste en "(L) es falsa", resulta (L) falsa (contradicción) y, por tanto, verdadera (nueva contradicción). Kripke soluciona la paradoja del mentiroso simple en su lenguaje $\mathcal{L}_\sigma$ donde no es expresable el predicado 'ni verdadero ni falso', pero en el metalenguaje en que explica su teoría sí aparece este concepto lo que hace que la paradoja simple del mentiroso vuelva a presentarse.

Aunque considero suficientemente argumentado que la teoría de Kripke no ofrece una solución satisfactoria, merecen ser referidas algunas otras críticas.

Entre los distintos puntos fijos correspondientes a distintas interpretaciones del lenguaje $\mathcal{L}$, Kripke destaca varios como especialmente interesantes (mínimo, máximo, intrínseco, el mayor intrínseco), pero no nos dice cuál corresponde a la interpretación correcta.[20] Tampoco se decanta por un esquema de evaluación de las conectivas lógicas.

Simmons (1993, nota 25, p. 194) también señala que hay oraciones intuitivamente fundamentadas que no están en el punto fijo mínimo, como son las oraciones fundamentadas del metalenguaje en que Kripke expone su teoría.

A Gupta (1982, parte IV) se deben interesantes críticas a la teoría de Kripke. Sin embargo, prefiero no enunciarlas debido a que las considero más discutibles que las anteriores.

2.3. Otras propuestas

En este apartado pretendo continuar el análisis de las dificultades que presentan diversas propuestas de solución. Fundamentalmente, me centraré en las esbozadas en la introducción, con la intención, allí manifestada, de que este análisis sirva para establecer los requisitos que debe cumplir una solución satisfactoria, las pruebas que debe superar y que no superan, en su totalidad, ninguna de estas propuestas.

[20] De acuerdo con Simmons (1993, nota 25, pp. 193-4), solamente el punto fijo mínimo captura la intuición del concepto de verdad.

2.3.1. Propuesta de van Fraassen[21]

La relación de presuposición entre dos oraciones A y B consiste en:

A presupone B ssi A es verdadero o falso solo cuando B sea verdadero

Van Fraassen sostiene que hay casos no triviales de presuposición por lo que se deben aceptar oraciones que no son verdaderas ni falsas. Por ejemplo, la oración "el rey de Francia es calvo" presupone que el rey de Francia exista; por tanto, de no existir, la oración "el rey de Francia es calvo" no será verdadera ni falsa.

Mediante lo que denomina *supervaluación* una oración de un lenguaje será verdadera (o falsa) en relación a un modelo, M, si lo es bajo la valuación clásica[22] de todos los modelos totales que son una extensión de M. Esto tiene dos consecuencias importantes: 1ª/ una supervaluación permite oraciones sin valor de verdad; 2ª/ una oración o un argumento en el lenguaje es válido en todas las supervaluaciones si y solo si es válido en todas las valuaciones clásicas, y por tanto, mediante la lógica clásica puede expresarse todo razonamiento *en* el lenguaje.

Van Fraassen introduce en el lenguaje formal un predicado, T, que pretende expresar *verdadero*. "A es falso" se expresa mediante $T(\sim A)$.

Veamos a continuación qué ocurre con las oraciones del mentiroso (L) y del mentiroso reforzada (SL).[23] Aunque van Fraassen no lo hace explícito, se entiende que (L) se expresará mediante $T(\sim L)$ y (SL) mediante $\sim T(SL)$.

(L) presupone una contradicción[24] por lo que no es verdadera ni falsa. Además, no hay problema en evaluar $T(L)$ como falsa.

Por supuesto, (SL) tampoco es verdadera ni falsa. Pero ahora $T(SL)$ tampoco puede ser verdadera ni falsa. Si fuera verdadera, (SL) también lo sería,[25] pero como, por definición, (SL) es $\sim T(SL)$, tendríamos a la vez $T(SL)$ y $\sim T(SL)$. Si fuera falsa, es decir si $T(\sim T(SL))$, $\sim T(SL)$ sería verdadera, que, por definición de (SL), es como decir que (SL) sería verdadera.

Así pues, no se puede expresar en el lenguaje que (SL) no es verdadera, lo cual, como señala Parsons,[26] muestra la incapacidad de T para expresar el concepto

[21] van Fraassen (1970).

[22] Una valuación clásica es la asignación a cada oración de T (*true*) o F (*false*) en concordancia con las tablas de verdad clásicas de las conectivas lógicas (not, or, etc.).

[23] Van Fraassen las denomina A y B respectivamente.

[24] Tanto si es verdadera como si es falsa se deduce una contradicción.

[25] Debido a que, por supuesto, dada una oración A, de $T(A)$ se deduce A.

[26] Parsons (1974) en Martin (1984, p. 12).

de verdadero tal como se emplea en el metalenguaje. En realidad no es preciso recurrir a la paradoja del mentiroso reforzada para mostrar esta incapacidad. Si quisiéramos expresar en el lenguaje que (L) no es falsa lo haríamos mediante $\sim T(\sim L)$ pero, por definición, $T(\sim L)$ es (L) por lo que deduciríamos $\sim (L)$. Es decir, $\sim (L)$ se evaluaría como verdadera y (L) como falsa.

El intento de solucionar este problema se basa en clasificar las oraciones en diferentes *tipos-valor*. Para ello, dada una oración, A, se define $T^n(A)$ como A, si $n = 0$, o como $T(T^{n-1}(A))$, si n es un número natural. Una oración es de tipo-valor n (*value-type* n) cuando n es el menor entero tal que $T^n(A)$ es verdadera o falsa.

Si $T^n(A)$ es verdadera, también lo será A y entonces A será de tipo-valor cero. Si A es de tipo-valor n, mayor que cero, A no es verdadera en un sentido extendido que vendrá expresado por: $\sim T^n(A)$ es verdadera para algún $n > 0$.

Pero con este nuevo sentido se puede volver a crear una paradoja representando en el lenguaje la oración A: "esta oración no es verdadera en sentido extendido", es decir, "existe n entero, $n \geq 0$, tal que $\sim T^n(A)$". Resulta ahora que $T^n(A)$ no puede ser verdadera ni falsa para ningún n. Es obvio que no puede ser verdadera porque también lo sería A. $T^n(A)$ tampoco puede ser falsa para algún n, porque entonces, se tendría que existe n tal que $\sim T^n(A)$ y eso es justo lo que expresa A.

El hecho de que $T^n(A)$ no puede ser de tipo-valor n para ningún n entero, lo expresa van Fraassen diciendo que A es de tipo-valor ω. Se da cuenta de que el problema de la no universalidad de su lenguaje formal no se ha resuelto y que los intentos por resolverlo sólo consiguen desplazarlo. Al final de su artículo reconoce que:

> the nontruth of a sentence of value-type ω cannot be expressed in the formal languages constructed [...] There are ways to remedy this, but the remedy usually provides us with the resources for constructing further paradoxes, which yield new limitations on our means of expression.[27]

[27] van Fraassen (1970, p. 23).

2.3.2. Propuesta de Martin[28]

El planteamiento de Martin es estudiar qué condiciones deben cumplir las oraciones para tener valores de verdad, con la intención de aplicar los resultados de ese estudio a la paradoja del mentiroso.

Su idea básica es que todo predicado, F, tiene un rango de aplicabilidad, $RA(F)$. Por definición, dada una oración, Fa,

[SC] Fa es semánticamente correcta si y solo si $a \in RA(F)$

Las oraciones semánticamente incorrectas no tienen valor de verdad.

Los otros cuatro pilares de la solución de Martin son:

- Si abreviamos *verdadero* mediante T y *falso* mediante F:

  [M] $RA(T) = RA(F) =$ conjunto de oraciones con valor de verdad

  Una alternativa sería *RA(T) = RA(F) = conjunto de todas las oraciones*. Pero, claramente, la escogida por Martin es más coherente con una solución basada en categorías o rangos de aplicabilidad. Sin embargo, tiene una consecuencia importante: si A no tiene valor de verdad, tampoco lo tienen $T(A)$, $F(A)$, $\sim A$ (la negación es electiva, no exclusiva), $T(\sim A)$, etc.

- Distinción entre referencia propia y referencia demostrativa. Nos interesa esta última, la de una expresión que se usa para identificar un objeto.

- Prueba para determinar si una oración autorreferencial de la forma Fa es semánticamente correcta:

  [SCA] Fa es semánticamente correcta si y solo si la referencia demostrativa del término sujeto de la oración $\in RA(F)$

  Martin establece otra prueba para las oraciones no autorreferenciales.

- Para poder afirmar que una oración tiene valor de verdad es necesario que pase la prueba de corrección semántica.

[28] Martin (1970a).

Consideremos la oración del mentiroso: L = *"esta oración es falsa"*. Para que pase la prueba de corrección semántica, la referencia demostrativa del término sujeto, es decir, L, debe pertenecer a $RA(falso)$ lo cual, según $[M]$, significa que L debe ser una oración con valor de verdad y, por tanto, semánticamente correcta. En definitiva, al someter L a la prueba de corrección semántica surge la necesidad de que, previamente, L pase dicha prueba. Esta regresión infinita supone que L nunca superará la prueba, es decir, según el criterio de Martin, no es semánticamente correcta y, por ello, carece de valor de verdad.

En mi opinión Martin debería justificar el criterio por el que, si no es posible probar que una oración es semánticamente correcta, considera probado que no lo es. Mientras no lo justifique podemos afirmar que se trata de un criterio *ad hoc* para que L resulte sin valor de verdad.

Veamos ahora el análisis de la oración del mentiroso reforzada, SL. Dado que Martin utiliza una negación electiva, SL no puede tener la forma "esta oración no es verdadera" (se necesitaría una negación exclusiva o renunciar a $[M]$) sino SL = *"esta oración es falsa o sin valor de verdad"*. Aquí surge el problema de determinar el rango de aplicabilidad de un predicado compuesto de la forma "P o Q". Sin mayor justificación, Martin propone (p. 96) que $RA(P\ o\ Q) = RA(P) \cap RA(Q)$. Como consecuencia, $RA(falso\ o\ sin\ valor\ de\ verdad) = RA(falso)$. Al aplicar a SL la prueba de corrección semántica, aparece la misma regresión infinita que cuando se aplicó a L, por lo que se concluye que SL también es semánticamente incorrecta.

Naturalmente, el criterio $RA(P\ o\ Q) = RA(P) \cap RA(Q)$ es criticable por tratarse de un criterio forzado para que, a pesar de que la oración "SL es sin valor de verdad" es verdadera, "SL es falsa o sin valor de verdad" carezca de valor de verdad. Máxime cuando el propio Martin, poco después, defiende un criterio incompatible con esta situación:

> for disjunctions [...] it seems more natural to regard the truth of one disjunct as sufficient for the truth of the disjunction (i.e. even if the other is without truth-value)[29]

si bien matizado por una nota al pie en la que recoge una sugerencia de Donellan para abandonar o al menos restringir este principio.

[29] Ibíd., p. 99.

En la sección III, el autor presenta la sintaxis y la semántica de un lenguaje formalizado en el que se pueden construir oraciones que atribuyen propiedades sintácticas o semánticas a otras expresiones del lenguaje.

En la sección IV reconoce el problema principal de su propuesta de solución, la falta de universalidad de ese lenguaje formal:

> The English predicate, 'is not a true sentence', where the 'not' is understood to be a merely "excluding negation", has not direct counterpart in this language.[30]

Este problema era ya claro desde el momento en que se estableció$[M]$ y se escogió la negación electiva, en consonancia con la idea de una solución basada en categorías. Puesto que si A no tiene valor de verdad, tampoco lo tendrá $T(A)$ ni $T(\sim A)$ y, consecuentemente, ni la oración falsa "A es una oración verdadera" vendrá expresada por $T(A)$ ni la oración verdadera "A no es una oración verdadera" vendrá expresada por $\sim T(A)$.

2.3.3. Propuesta de McGee

McGee defiende que el predicado "verdadero" es un predicado vago. Esta vaguedad significa, según su propuesta, que habrá oraciones claramente verdaderas, claramente falsas e indeterminadas (*unsettled*). Sin embargo, no se quiere caer en una semántica con tres valores de verdad (aparecerían problemas similares a los de otras propuestas de ese tipo, como las de Kripke o Martin) para lo que se propone caracterizar las oraciones indeterminadas del siguiente modo: "'unsettled' means, not "neither true nor false" but "either true or false but it is not settled which"'[31]

La versión de la oración del mentiroso reforzada con que se tiene que enfrentar esta propuesta es

(MG) (MG) no es claramente verdadera

En principio, si suponemos que (MG) es indeterminada podemos decir que es claramente verdadero que (MG) es indeterminada y, por tanto, que es claramente verdadero que (MG) no es claramente verdadera. Por definición de (MG), esto significa que (MG) es claramente verdadera, lo que contradice el supuesto inicial.

[30] Ibíd., p. 104.
[31] McGee (1989, p. 535).

Sin embargo, McGee argumenta que este razonamiento no es correcto, dado que de la hipótesis de que una oración es indeterminada no se sigue que sea claramente verdadero que la oración es indeterminada.[32] Téngase en cuenta que

> A sentence is definitely true if our linguistic conventions, together with the nonsemantic facts, insure that it is true. Definite truth is itself a vague notion. That our conventions do not insure a sentence' truth does not imply that there are conventions that forbid our making conventions that would insure its truth. So the fact that a sentence $\ulcorner\phi\urcorner$ is not definitely true does not entail that it is definitely true that $\ulcorner\phi\urcorner$ is not definitely true.[33]

Dado que suponer que (MG) es claramente verdadera o que es claramente falsa conducen a contradicción, McGee parece sugerir que (MG) es indeterminada pero está indeterminado que lo sea. Lo que realmente afirma es que su teoría deja el estatus de la oración completamente abierto.

Creo que esta propuesta de solución no es correcta. Como hemos visto, para McGee, que no esté determinado p significa que p es verdadero o falso pero no está establecido si es una cosa u otra. Por tanto, si no está establecido que (MG) sea indeterminada, "(MG) es indeterminada" es verdadera o falsa y no está establecido si es una cosa u otra. Pero esto no es cierto. "(MG) es indeterminada" no puede ser falsa porque entonces (MG) no sería indeterminada y, por tanto, sería claramente verdadera o claramente falsa. Pero cualquiera de estos dos supuestos lleva a contradicción. Al no poder ser falsa, hemos asegurado, hemos establecido, que "(MG) es indeterminada" es verdadera,[34] lo que nos permite afirmar, en contra de McGee, que es claramente verdadero que (MG) es indeterminada.

Desde otro punto de vista, se puede alegar que la solución anterior no es una verdadera solución. Por un lado afirma que (MG) es indeterminada y, por otro, resulta que no está determinado si lo anterior es verdadero o falso. Si esto no lo queremos ver como contradictorio tendremos que aceptar que afirmar "(MG) es indeterminada" es vacío de contenido, porque, del mismo modo que si no afirmamos nada, queda abierta la posibilidad de que "(MG) es indeterminada" sea

[32] McGee (1990, p. 7).

[33] McGee (1989, p. 537-8).

[34] No se olvide que toda oración que no es falsa es verdadera y viceversa (aunque no pueda decirse lo mismo si cambiamos "falsa" por "claramente falsa" y "verdadera" por "claramente verdadera").

falso. En definitiva, la solución de McGee, si no es contradictoria, se reducirá a no decir nada con contenido semántico sobre (MG) y, por supuesto, eso no puede considerarse una solución. Precisamente el problema de la oración del mentiroso reforzada es que parece que no se puede decir nada sobre su veracidad sin caer en contradicción.

Otras críticas interesantes a las propuestas de McGee pueden verse en Simmons (1993, pp. 72-8) y Mills (1995).

2.3.4. Propuestas *inconsistentes*

Los defensores de este tipo de propuestas, entre los que podemos señalar a Rescher, Brandom, Chiara y Priest, defienden que la noción de verdad es inconsistente y la oración (L) del mentiroso es a la vez verdadera y falsa. Eso no significa que la noción de verdad sea irrelevante o vacua. Por el contrario, es fundamental que la noción de verdad satisfaga la condición de adecuación material de Tarski:

> [...] all instances of the T-scheme, $T\ulcorner\phi\urcorner \leftrightarrow \phi$, should be provable. For this does indeed characterise, at least in a weak sense, our naive notion of truth[35]

Los argumentos para defender este tipo de propuestas son variados. Destacaré dos. Uno es que la condición de Tarski aplicada al lenguaje natural conduce a contradicción, puesto que el lenguaje natural es semánticamente cerrado. Pero esas contradicciones no deben asustarnos. De hecho, podemos comprobar que en la vida ordinaria de las personas se aceptan inconsistencias sin, por ello, aceptar que todo es verdadero y falso a la vez (a pesar de que, desde el punto de vista lógico clásico, de una contradicción se deduce cualquier proposición). Por ejemplo, las leyes de un país son inconsistentes en diversos aspectos sin que eso signifique que el sistema legal en su conjunto sea inservible o trivial. En el campo de la lógica formal se trata pues de que:

> [...] the received formal logical theory must be changed. [...] However, it is important to notice that the received theory is not blindly destroyed. [...] we must come to accept some formulas of the form $A \wedge \neg A$, since some of these are indeed true.[36]

[35] Priest (1984, p. 155).
[36] Ibíd., p. 154.

El otro argumento es que no parece posible una solución a las paradojas semánticas dentro de una lógica consistente:

> The reasons for supposing the logical paradoxes to be true contradictions are at least two-fold. The major reason is that all attempts to treat them as anything else have been singularly unsuccessful, or at any rate a good deal less successful than the present proposal. Nor is this merely an inductive argument. For there are substantial reasons why no consistent approach to the paradoxes can work.[37]

En cuanto a la oración L del mentiroso, de la condición de adecuación de Tarski resulta que si L es verdadera, L es falsa y viceversa. Además, si, para defender una solución consistente, decimos de L que no es verdadera ni falsa, tenemos que L no es falsa, justo lo contrario de lo que dice L, luego L es falsa, etc. Las teorías que no aceptan inconsistencias no tienen fácil la solución de la paradoja. En cambio, si aceptamos que los principios que rigen el concepto de verdad son inconsistentes, y no por ello desechables, tenemos una razón para aceptar que haya oraciones verdaderas y falsas a la vez y, precisamente, esta posibilidad permite encontrar una solución a la paradoja del mentiroso que de otra forma no parece posible; al menos, sin tener que apelar a sofisticadas o rebuscadas nociones lógicas.

Se pueden hacer diversas críticas a las propuestas inconsistentes. En primer lugar, habrá quien considere no aceptable una lógica que acepte inconsistencias (inconsistente, paraconsistente...) y, por tanto, la solución a las paradojas que se deriva de ella. La segunda crítica es que la solución de afirmar que L es a la vez verdadera y falsa parece descaradamente *ad hoc*. Sin embargo, hay buenas razones en defensa de las lógicas *inconsistentes*, lo cual haría más cuestionable dicha crítica. Pero dejemos las críticas globales a las lógicas *inconsistentes* y pasemos a analizar la solución concreta que ofrecen a la paradoja del mentiroso.

En estas lógicas habrá oraciones verdaderas y no falsas (TnF), oraciones falsas y no verdaderas (FnT), oraciones verdaderas y falsas (TF) y, posiblemente, oraciones no verdaderas ni falsas (nTnF). Es claro que L no puede ser no falsa, porque, si L no fuese falsa, tendríamos que "L es falsa" es falsa y, por ende, que L es falsa. También es claro que, al ser L falsa, "L es falsa" es verdadera y, consiguientemente, L es verdadera. Esta sería la prueba de que L es verdadera y falsa. Como vemos, aunque estemos en una lógica inconsistente, el razonamiento se basa en que L no

[37] Ibíd., p. 153.

puede ser falsa y no falsa a la vez. Más en concreto, las oraciones pueden ser de uno de los cuatro tipos enunciados al principio (TnF, FnT, TF o nTnF) pero, se entiende, no de varios de esos tipos a la vez.

¿Qué ocurre con la oración del mentiroso reforzada (*SL = "esta oración no es verdadera"*)? Es fácil ver que SL no puede ser TnF, FnT, TF ni nTnF porque al suponer que es de uno de esos cuatro tipos se llega a la conclusión de que es de otro tipo diferente. En realidad, basta darse cuenta de que si suponemos SL verdadero, resulta no verdadero y viceversa.

Un intento de superar la paradoja del mentiroso reforzada, por analogía a la forma en que se superó la del mentiroso simple, consistiría en afirmar que SL es a la vez verdadero, no verdadero, falso y no falso. Pero esto significaría admitir que los cuatro tipos TnF, FnT, TF y nTnF no son excluyentes lo cual daría lugar a la aparición de 16 tipos (cada uno correspondería a un subconjunto de {TnF, FnT, TF, nTnF}). En esta situación, "no ser verdadero" admitiría dos interpretaciones: pertenecer a uno de los tipos que contiene FnT o nFnT o no pertenecer a ninguno de los tipos que contienen TnF o TF. Si suponemos que los 16 tipos que tenemos ahora son excluyentes, bastaría tomar esta segunda interpretación para que SL siguiera sin solución. Porque si SL es verdadera (lo cual significaría que pertenece a uno de los tipos que contienen TnF o TF), SL no es verdadera (lo cual significaría que no pertenece a uno de los tipos que contienen TnF o TF) y viceversa.

Naturalmente, si consideramos nuestros 16 tipos no excluyentes solo vamos a complicar las cosas. Aparecerán 2^{16} tipos y existirá una interpretación de "no es verdadero" bajo la cual SL seguirá sin solución.

Como vemos las soluciones inconsistentes no superan la paradoja del mentiroso reforzada. Además, como señalan Gupta y Belnap (1993, p. 14 y ss.), tienen la nada deseable consecuencia de que no solo las oraciones paradójicas sino todas las que no lo son resultan verdaderas y falsas a la vez. Para ello consideran la oración

(C) Si (C) es verdadera entonces X

donde X puede ser una oración no paradójica cualquiera. (C) es la conocida como paradoja de Curry o de Löb. Si aceptamos a la vez la condición de Tarski y que "(C)" es un nombre de la oración "Si (C) es verdadera entonces X", podemos demostrar X:

	(1)	(C) es verdadera ↔ [Si (C) es verdadera entonces X] ; Condición de Tarski aplicada a (C)	
2	(2)	(C) es verdadera	; Hipótesis
2	(3)	Si (C) es verdadera entonces X	; 1, 2
2	(4)	X	; 2, 3
	(5)	Si (C) es verdadera entonces X	; 2, 4 (eliminación hipótesis)
	(6)	(C) es verdadera	; 1, 5
	(7)	X	; 5, 6

Así puede demostrarse cualquier proposición y su contraria, con lo que la inconsistencia contamina oraciones que, en principio no eran problemáticas. Ahora todas las oraciones son verdaderas y falsas y, como consecuencia, los conceptos verdadero y falso quedan desprovistos de contenido.

2.3.5. Propuestas basadas en un concepto de verdad sensible al contexto

Este tipo de soluciones se basa en considerar que la extensión del término "verdadero", como la de "aquí" o "ahora", depende del contexto en que se emite la oración que lo contiene. Como muestra de este tipo de propuestas comentaré la de Barwise y Etchemendy (1987), quizás la más influyente de ellas.

Inspirándose en la visión del concepto de verdad de Austin (1950), Barwise y Etchemendy (1987) consideran que la proposición expresada por un enunciado (por la emisión de una oración declarativa), no solo depende del significado de la oración sino también de la *situación* a la que el enunciado se refiere. Así la oración "Clara tiene el as de corazones" puede dar lugar a un enunciado verdadero o a uno falso dependiendo de la situación en que se enuncie. Pero no se trata de la situación global del mundo, es decir, el enunciado sería falso si se produce en una sala donde ninguno de los jugadores se llama Clara aunque en otro lugar del mundo haya una persona llamada Clara que tenga en sus manos el as de corazones.

En una proposición p, hay pues dos componentes: una situación (*About(p)*) y un tipo (*Type(p)*, determinado por el significado de la oración). La proposición p es verdadera si la situación *About(p)* es del tipo *Type(p)*. En caso contrario, la proposición se considera falsa.

Barwise y Etchemendy realizan un desarrollo formal de estas ideas comenzando por la definición de un lenguaje formal muy sencillo y modelando las proposiciones de acuerdo con las ideas anteriores.

Las proposiciones son de la forma

$$\{s; T\}$$

donde s es una situación y T un tipo.

Una situación se modela como un conjunto de *estados de cosas* (*states of affairs*), siendo un estado de cosas un vector de una de las formas

$$< H, a, c; i >$$
$$< Tr, p; i >$$
$$< Bel, a, p; i >$$

donde a es un persona, c un naipe y p una proposición. $i \in \{0, 1\}$. Informalmente, $< H, a, c; 1 >$ significa que a tiene el naipe c; $< Tr, p; 1 >$, que p es verdadera y $< Bel, a, p; 1 >$, que a cree p. Si cambiamos la polaridad, $i = 0$, la relación (H, Tr o Bel) no se da. Así $< H, a, c; 0 >$ significará que a no tiene el naipe c, etc.

Un tipo es de la forma

$$[\sigma]$$

donde σ es un estado de cosas, o de una de las formas

$$[\wedge X]$$
$$[\vee X]$$

para algún conjunto de tipos, X.

Con este modelado de las proposiciones, se dice que una proposición $\{s; T\}$ es verdadera si y solo si s es del tipo T, es decir,

[T1] $\{s; [\sigma]\}$ es verdadera ssi $\sigma \in s$

[T2] $\{s; [\wedge X]\}$ es verdadera ssi $\{s; T\}$ es verdadera para todo $T \in X$

[T3] $\{s; [\vee X]\}$ es verdadera ssi $\{s; T\}$ es verdadera para algún $T \in X$

¿Cómo analizar la oración del mentiroso? En primer lugar conviene resaltar nuevamente que una oración no determina por sí sola una proposición. Debemos ver la

oración como una *función proposicional* en que los valores posibles del argumento son situaciones:

> Let's distinguish between the meaning of a sentence and the propositional content of a statement made with it. Intuitively, the former should be a propositional function, something that gives us a proposition when supplied with the situation the proposition is about, while the latter would be such a proposition. Thus a sentence can be ambiguous in terms of propositional content without having two separate meanings, without expressing two distinct propositional functions.[38]

Por tanto, una misma oración puede expresar distintas proposiciones sin cambiar su significado, simplemente porque cambie la situación a la que se aplica.

Los autores definen una función, Val, que hace corresponder a cada oración una proposición paramétrica. Un parámetro es siempre s, una situación. Otros parámetros posibles son p y q_i (proposiciones) que corresponden, respectivamente, a las expresiones demostrativas $this$ y $that_i$ del lenguaje formal.

En el lenguaje formal, la oración del mentiroso es

$$\downarrow \neg True(this) \tag{2.1}$$

La función Val asocia a esta oración la proposición paramétrica p que es la única solución de la ecuación

$$p = \{s; [< Tr, p; 0 >]\} \tag{2.2}$$

El hecho de que haya una única solución a esta ecuación es demostrable dentro de la teoría de hiperconjuntos. Barwise y Etchemendy utilizan esta teoría de conjuntos, que sustituye el axioma de fundación de la teoría clásica de conjuntos (ZFC) por el de antifundación. Este axioma permite que la relación de pertenencia sea circular, por lo cual, la teoría de hiperconjuntos es más adecuada para modelar proposiciones circulares como las del mentiroso.

Si llamamos f_s a la solución de la ecuación (2.2) tendremos:

$$f_s = \{s; [< Tr, f_s; 0 >]\} \tag{2.3}$$

Es decir, tenemos una proposición del mentiroso para cada situación s.

[38] Barwise y Etchemendy (1987, p. 138).

Hay situaciones que deberíamos descartar porque no pueden corresponderse con el *mundo real*, por ejemplo, aquellas que son incoherentes por contener un estado de cosas y su dual.[39] Para que una situación pueda corresponderse con el mundo real debe tratarse de una *situación posible*, es decir, aquella situación s que cumple:

- Si un estado de cosas pertenece a s, su dual no pertenece a s
- Si $< Tr, p; 1 > \in s$, entonces p es verdadera
- Si $< Tr, p; 0 > \in s$, entonces p es falsa

Consideremos, finalmente, la proposición del mentiroso, f_s, para una situación posible, s. De acuerdo con [T1] (p. 44) y (2.3), f_s es verdadera ssi $< Tr, f_s; 0 > \in s$. Pero, como s es una situación posible, $< Tr, f_s; 0 > \in s$ implica que es f_s falsa. Así pues, f_s no puede ser verdadera y, por tanto, es falsa.

La paradoja ha desaparecido. A cambio, el hecho semántico de que f_s es falsa no puede pertenecer a s (porque entonces f_s sería verdadera). Sin embargo, ¡ese hecho semántico puede pertenecer a otra situación posible, s', más amplia! Basta con tomar

$$s' = s \cup \{< Tr, f_s; 0 >\} \tag{2.4}$$

El procedimiento puede reiterarse indefinidamente puesto que la proposición del mentiroso para s'

$$f_{s'} = \{s'; [< Tr, f_{s'}; 0 >]\} \tag{2.5}$$

es falsa, hecho que se *diagonaliza* fuera de s' pero que puede encontrarse en

$$s'' = s' \cup \{< Tr, f_{s'}; 0 >\} \tag{2.6}$$

Esta secuencia creciente de situaciones justifica que se considere que estamos ante una solución jerárquica. Por supuesto, no hay una situación universal, pero ello no parece preocupar demasiado a Barwise y Etchemendy, quienes consideran este hecho como una conclusión que debemos aceptar, del mismo modo que, para evitar la paradoja de Russell, aceptamos que no existe el conjunto de todos los conjuntos.

[39] El dual de un estado de cosas es el que resulta de cambiar únicamente su polaridad.

> [...] no actual situation can contain all the facts of the world. For no matter how comprehensive we take an actual situation w to be, it must at least omit the first-class fact that f_w is false. Thus, just as the Russell construction shows us that there cannot be a universal set, the Liar construction shows that the situations propositions can be about fall short of universality.[40]

Sin duda, la propuesta que he intentado describir en sus aspectos fundamentales, es atractiva. Para algunos puede tener el atractivo filosófico derivado de un concepto de verdad sensible al contexto, para otros el de evitar la paradoja del mentiroso y afines.

No obstante, también hay aspectos de esta propuesta que, cuanto menos, calificaría de poco agradables.

En primer lugar, cabe argumentar que si el contexto al que se refiere un enunciado es conocido, puede hacerse explícito en el enunciado. Retomemos el ejemplo de la oración "Clara tiene el as de corazones" emitida en una sala, S, donde hay unos jugadores jugando a las cartas, pero ninguno se llama Clara. Según Barwise y Etchemendy el enunciado es falso aunque haya, en otro lugar del mundo, una persona llamada Clara que tenga el as de corazones. Pero nada nos impide incluir el contexto en la oración y expresar el mismo enunciado con una oración que ya no depende de ese contexto: la oración "en la sala S hay una persona llamada Clara que tiene el as de corazones".

De hecho, algo similar hacen los autores cuando establecen [T1]:

[T1] $\{s; [\sigma]\}$ es verdadera ssi $\sigma \in s$

Aceptan implícitamente que, dados un *estado de cosas*, σ, y una *situación* s, la oración "$\sigma \in s$" expresa una proposición por sí misma, y no, como su propia teoría establece, una función proposicional que se convierte en verdadera o falsa según qué situación tome como argumento.

En segundo lugar, como los propios autores reconocen (ibíd., p. 176), cuando afirmamos algo, no esta claro qué situación corresponde a esa afirmación. Sin embargo, se defienden, normalmente no es preciso que los límites exactos de la situación estén bien determinados. Lo que no especifican es cuáles son los aspectos de una situación que deben estar bien determinados y, lo más importante, si esos

[40] Ibíd., p. 155.

aspectos quedan determinados siempre que se enuncia una oración. Si nos fijamos en oraciones como la del veraz o la del mentiroso, que solo se refieren a sí mismas, todo parece indicar que no. Su propia naturaleza nos sugiere que su significado, si lo tienen, no depende del contexto. Da la sensación de que el único motivo por el que se hace que la proposición expresada por una oración autorreferencial sea sensible al contexto es porque, de ese modo, se altera la autorreferencia que provocaba la paradoja. La interpretación informal de la proposición f_s, ya no es "f_s es falsa" sino "(f_s es falsa) $\in$ s".[41]

Consideremos la oración del veraz, que, en la formalización de Barwise y Etchemendy es:

$$t_s = \{s; [< Tr, t_s; 1 >]\} \tag{2.7}$$

Es fácil ver que, para algunas situaciones posibles es verdadera y para otras falsa (ibíd., teorema 7 en p. 133). Ahora bien, imaginemos que se reúne un grupo de personas y una dice "esta oración es verdadera". ¿Hay forma de determinar, siquiera, si la situación a la que se refiere ese enunciado es de las que hacen t_s verdadera o de las que la hacen falsa? Claramente, no.

La pertenencia de un estado de cosas a una situación parece que debiera ser una cuestión que se pudiera decidir empíricamente. Mas, en el caso anterior, no se ve la manera de decidir si $< Tr, t_s; 1 >$ pertenece a s o no. Tampoco en el caso de la oración del mentiroso se decide, empíricamente, si $< Tr, f_s; 0 >$ pertenece a la situación posible s o no. Por el contrario, la no pertenencia de $< Tr, f_s; 0 >$ a s, es algo que, según vimos, se establece como una conclusión lógica.

Una última crítica a la propuesta de Barwise y Etchemendy (aunque no a todas las propuestas *contextuales*) es que produce una solución jerárquica de la paradoja del mentiroso.[42] Como ocurre en las soluciones jerárquicas las afirmaciones sobre esa jerarquía suelen estar fuera de ella. Por ejemplo, como hemos visto, para cada situación, s, hay una proposición del mentiroso, f_s, y toda proposición del mentiroso es falsa. Llamemos r a la situación que corresponde a esta afirmación. Entonces, si la proposición formada por la situación r y (el tipo correspondiente a) la afirmación anterior es verdadera, deben pertenecer a r todos los estados de cosas

[41] De este modo se rompe la simetría de la interpretación convencional que es la que nos induce a pensar que no hay más razones para que la oración del mentiroso sea falsa que para que sea verdadera. Así se explica la antiintuitiva conclusión de que la proposición del mentiroso es ahora falsa.

[42] Simmons (1993) propone una solución contextual no jerárquica que no analizamos en este trabajo.

que *dicen* que toda proposición del mentiroso es falsa. Es decir, para toda situación s, se debe cumplir $< Tr, f_s; 0 > \in r$. En particular, se tendría $< Tr, f_r; 0 > \in r$, pero como $f_r = \{r; [< Tr, f_r; 0 >]\}$, eso significaría que f_r es verdadera. Lo que contradiría que toda proposición del mentiroso es falsa.

En definitiva, la afirmación de que toda proposición del mentiroso es falsa no puede expresarse, sin contradicción, dentro del lenguaje que los autores utilizan para solucionar la paradoja del mentiroso. Es pues, pertinente repetir la pregunta que hicimos al criticar las soluciones jerárquicas: ¿qué lugar ocupa esa afirmación? Y podemos repetir la contestación: si la respuesta es que ese tipo de afirmaciones solo es posible en el lenguaje natural, la solución jerárquica desplaza el problema de las paradojas a este lenguaje, pero no lo resuelve. La respuesta alternativa, a saber, que ese tipo de afirmaciones es inexpresable, tampoco es satisfactoria, porque entonces no se podría exponer en qué consiste la teoría.

2.3.6. Las oraciones autorreferenciales como ecuaciones

Hansson (1978) considera que:

> To interpret a sentence is thus to find the answer to the question: which proposition (possible fact, state of affairs) satisfies the condition that the sentences poses? It is illuminating to compare this question with a corresponding one for equations: which number x satisfies the condition posed by the equation?[43]

En muchos casos, la condición que establece una oración es trivial. Por ejemplo, la oración "Juan come una manzana" establece una condición que puede expresarse mediante la ecuación $x =$ *Juan come una manzana*. Igual que para la ecuación numérica $x = 10$, la solución es obvia. Sin embargo, cuando la condición es impredicativa la situación es diferente. La condición será de la forma $x = f(x)$ que, dependiendo de la función f, puede tener una, ninguna o varias soluciones. Y esto es tan cierto, insiste Hansson, en las ecuaciones proposicionales como en las numéricas.

Para interpretar la oración del mentiroso tendremos que responder a la pregunta de qué proposición satisface la condición $x = x$ *es falsa*. Puesto que las

[43] P. 382.

proposiciones solo pueden ser verdaderas o falsas, es fácil ver que ninguna proposición puede satisfacer la condición. Así pues la oración del mentiroso no expresa ninguna proposición.

Por el contrario, la ecuación $x = x$ *es verdadera* correspondiente a la oración del veraz admite infinitas soluciones y, consecuentemente, dicha oración también es paradójica.

El recurso a las ecuaciones proposicionales para interpretar las oraciones autorreferenciales (o, más en general, a sistemas de ecuaciones proposicionales para interpretar conjuntos de oraciones que se referencian entre sí) es, sin duda, interesante. La formación de esas ecuaciones a partir de las oraciones es sencilla y directa. Las ecuaciones reflejan la estructura de sus correspondientes oraciones y, por ello, pueden facilitar un análisis de las mismas. En la ecuación proposicional la paradoja se disuelve, porque lo que antes era una contradicción se convierte ahora, simplemente, en la inocua constatación de que hay una ecuación sin solución.

Este es también el análisis de Wen (2001), otro defensor de interpretar las paradojas semánticas mediante ecuaciones. Para él la oración del mentiroso no muestra un simple objeto AF (donde A simboliza "esta oración" y F simboliza "es falsa") sino una relación referencial $A := AF$, es decir, A refiere a AF. De modo bastante natural surge un interesante análisis de la oración del mentiroso:

> It is the assumption of existence of a sentence given that satisfies the Liar equation $X := XF$ that causes contradiction in the Liar paradox. In other words, there can be no sentence given that says, of itself, that it is false.[44]

Lo que queda confuso en la propuesta de Lan Wen y, sobretodo, en la de Hansson es la naturaleza verdadera de las oraciones con referencias a oraciones. En una solución basada en ecuaciones hay dos alternativas: a) la oración es un modo de representar una ecuación proposicional; b) la oración no representa a la ecuación sino a la solución de esa ecuación, siempre que haya una y solo una.

Creo que la primera alternativa es descartable. La oración y la ecuación proposicional son entidades de naturaleza diferente. Las proposiciones las expresamos mediante oraciones no mediante ecuaciones proposicionales, las cuales, al tener variables libres, nunca expresan una proposición. Consideremos el siguiente par

[44] Ibíd., p. 45.

de oraciones: "Roma es la capital de Italia"; "la anterior oración es falsa". Las ecuaciones que esta secuencia de oraciones establece podemos escribirlas del siguiente modo: $x =$ *Roma es la capital de Italia*; $y = x$ *es falsa*. La segunda oración, vaya por caso, no es un modo de representar la ecuación $y = x$ *es falsa*, como lo pone de manifiesto el hecho de que la oración es falsa, en tanto que la ecuación, en todo caso, se convierte en verdadera cuando asignamos a x e y las proposiciones expresadas por la primera y la segunda oraciones, respectivamente.

La única alternativa razonable es que la oración represente la solución de la ecuación proposicional que tiene asociada. Para ello es necesario que esa ecuación tenga una única solución. Si no se cumple este requisito, como en las oraciones del mentiroso y del veraz, simplemente diremos que la oración no expresa proposición alguna. De acuerdo con este planteamiento, una oración autorreferencial tendría una representación formal del tipo

$$(\iota x)\,(x = f(x)) \tag{2.8}$$

entendiendo que x es una variable proposicional y que si $x = f(x)$ no tiene una única solución, a (2.8) no corresponde ninguna proposición.

Dado que ninguna proposición es solución de "$x =$ *x no es verdadera*", a la oración del mentiroso reforzada no corresponde ninguna proposición. Sin embargo, desafortunadamente, el razonamiento informal sobre la oración del mentiroso reforzada (SL) que lleva a contradicción sigue siendo válido: si a (SL) no corresponde ninguna proposición, la oración (SL) no es verdadera, pero eso es justamente lo que afirma (SL), luego (SL) es verdadera.

Se puede alegar que este razonamiento que lleva a contradicción no puede seguirse si se usa estrictamente la versión formal (2.8), pero ello no proporciona una explicación de qué es lo que falla en la versión informal; en todo caso mostraría que la versión formal no es fiel reflejo de la informal.

2.3.7. Gupta y Belnap: teoría de la revisión de la verdad

Estos autores defienden que: "It is a fundamental intuition concerning truth that the T-biconditionals are analytic and that they *fix* the meaning of 'true'",[45] hasta el punto de que la paradoja del mentiroso debe ser resuelta sin dañar esta

[45] Gupta y Belnap (1993, p. 6). Los bicondicionales-T son las equivalencias que tienen la forma del esquema [T] de Tarski.

intuición fundamental. Añaden una precisión importante: la noción de verdad fijada por los bicondicionales-T es una noción de verdad lógica, absoluta y débil. Considero necesario destacar cuáles son para los autores las nociones de verdad débil y fuerte:

> With the weak notion, the semantic status of '"P" is true' is exactly the same as that of P. If P is neither true nor false, so is '"P" is true'. If P is both true and false, then so is '"P" is true' [...] With the strong notion(s) of truth the equivalence between P and '"P" is true' is not perfect. For example, with this reading, '"P" is true' may be false even though P is not false but only neither true nor false.[46]

A mi juicio, la noción débil de verdad no es la más acertada. Según nuestras intuiciones más básicas, si P no es verdadera ni falsa, debería poder deducirse: (1) P no es verdadera; (2) es falso que P es verdadera. Pero con la noción débil de verdad "P es verdadera", al igual que P, no será verdadera ni falsa, por lo que no es falso que P es verdadera, con lo cual contradecimos (2).[47]

Gupta y Belnap encuentran en la teoría de las definiciones la clave para resolver adecuadamente el problema de la verdad (paradojas incluidas). Tras analizar algunos ejemplos destacan que "*every kind of pathological behavior that the concept of truth exhibits can be mirrored in concepts with circular definitions*".[48]

Su teoría lógica de las definiciones se basa en que una definición fija completamente el significado del término definido (*definiendum*), *incluso aunque la definición sea circular*. Dado que, en general, una definición circular no permite determinar la extensión de su *definiendum* es preciso pensar en el significado de un modo diferente. Este nuevo modo consiste en que el significado que una definición circular asigna a un *definiendum* tiene un carácter hipotético. Solo podemos decir cuál sería la extensión del *definiendum* en función de una extensión previamente supuesta para el mismo. Esta función, $\delta_{D,M}$, dependiente de un conjunto de hechos, M, y de una definición, D, es vista como una *regla de revisión*. Aplicada a una extensión hipotética, X, nos proporciona una extensión $\delta_{D,M}(X)$, que es un

[46] Ibíd., p. 22.

[47] Además, si como en la propuesta de Martin, combinamos la noción débil de verdad con la negación electiva se produce una reducción de la capacidad expresiva del lenguaje. Porque si P no es verdadera ni falsa no hay modo de expresar la proposición cierta de que P no es verdadera, ya que tanto "'P' es verdadera" como "'P' no es verdadera" no serán ni verdaderas ni falsas.

[48] Ibíd., pp. 116-117, en cursiva en el original.

candidato mejor (o igual de bueno) para ser considerado la extensión del concepto definido. Repetidas aplicaciones de la función $\delta_{D,M}$ generan una secuencia de extensiones que cada vez son mejores (o igualmente buenos) candidatos para la extensión del concepto definido.

El comportamiento de las secuencias de extensiones que se generan a partir de cada una de las posibles extensiones iniciales será el que determine de modo categórico si el *definiendum* es ordinario o patológico. Consideremos, a modo de ejemplo, la siguiente definición:[49]

$$x\ es\ G =_{def} x\ es\ H \vee \neg(x\ es\ G) \tag{2.9}$$

y supongamos que el dominio es $D = \{a, b\}$ y la extensión de H es $I(H) = \{a\}$. Las secuencias de extensiones de G que se obtienen a partir de cada una de las posibles extensiones iniciales son:

$$\begin{array}{l}(\emptyset, D, \{a\}, D, \{a\}, D, \{a\}, D...)\\ (\{a\}, D, \{a\}, D, \{a\}, D...)\\ (\{b\}, \{a\}, D, \{a\}, D, \{a\}, D...)\\ (D, \{a\}, D, \{a\}, D, \{a\}, D...)\end{array} \tag{2.10}$$

Se observa que, a partir de un determinado número de aplicaciones de la regla de revisión (en nuestro caso, a partir de la primera aplicación) el elemento a pertenece invariablemente a la extensión de G. Esto nos permite afirmar de modo categórico que la oración "$a\ es\ G$" es verdadera. En cambio, el elemento b no tiene ese comportamiento: si b pertenece a la extensión de G en una etapa de revisión, no pertenece a esa extensión en la etapa siguiente y viceversa. El valor de verdad de "$b\ es\ G$" no se estabiliza, sino que, por el contrario oscila indefinidamente. Diremos, de modo categórico, que la oración "$b\ es\ G$" es *patológica*.

Naturalmente, las secuencias de extensiones asociadas a una definición (y a un conjunto de hechos) pueden ser de muy diversos tipos, lo que permite definir distintos tipos de oraciones patológicas. Por ejemplo, la definición

$$(TT)\ es\ verdadera =_{def} (TT)\ es\ verdadera \tag{2.11}$$

[49] He elaborado un ejemplo más sencillo que el de los autores.

da lugar a secuencias estables pero falla la convergencia porque las distintas secuencias no convergen a un mismo punto fijo. En cambio,

$$(L)\ es\ verdadera =_{def} (L)\ es\ falsa \tag{2.12}$$

da lugar a secuencias no estables que oscilan indefinidamente pero hay convergencia, en el sentido de que "the "patterns" of revisions we get do not depend upon the initial hypothesis".[50] Por tanto, la teoría de la revisión de la verdad permite distinguir el diferente comportamiento patológico de (TT) y (L).

Puesto que, según los autores, los bicondicionales de Tarski fijan el significado de "verdadero", es importante señalar que la lectura correcta de estos bicondicionales es la equivalencia *definicional* ($=_{def}$) no la equivalencia material ($\leftrightarrow$). En una derivación basada en la primera, hay un cambio en el índice que establece la etapa de revisión. Por ejemplo, dada la definición

$$X\ es\ verdadera =_{def} p \tag{2.13}$$

a partir de la hipótesis de que "X *es verdadera*" se verifica en la etapa i, se deduce que p se verifica en la etapa $i-1$ y viceversa. En cambio, dada la equivalencia

$$X\ es\ verdadera \leftrightarrow p \tag{2.14}$$

la premisa y la conclusión tienen siempre el mismo índice. Así consiguen que la oración del mentiroso no genere contradicción. La generaría si se interpretase como

$$(L)\ es\ verdadera \leftrightarrow (L)\ es\ falsa \tag{2.15}$$

pero no si se interpreta como (2.12).

Si hacemos explícita la etapa de revisión, una definición de la forma

$$x\ es\ G =_{def} A(x, G) \tag{2.16}$$

podría reducirse a

$$\forall i([x\ es\ G\ en\ la\ etapa\ i] \leftrightarrow [A(x, G)\ en\ la\ etapa\ i-1]) \tag{2.17}$$

[50] Ibíd., p. 136.

Aplicado a la oración del mentiroso, tendríamos

$$\forall i([(L)\ es\ verdadera\ en\ la\ etapa\ i] \leftrightarrow [(L)\ es\ falsa\ en\ la\ etapa\ i-1]) \quad (2.18)$$

donde se pone de manifiesto una importante similitud con otras teorías jerárquicas, con las que, a mi juicio, comparte el inconveniente de ofrecer una interpretación demasiado artificiosa de la oración del lenguaje natural "esta oración es falsa".

Una interpretación más sencilla, y al mismo tiempo más general, de las definiciones circulares consiste en interpretarlas como ecuaciones. En realidad, la interpretación de Gupta y Belnap se puede ver como una interpretación de ese tipo pero con la *innecesaria restricción* de buscar sus soluciones exclusivamente de un modo iterativo: se da un valor a la incógnita en la expresión definidora y se calcula un nuevo valor para aquella; el proceso se repite y se observa su evolución y convergencia. Es evidente que hay una forma más sencilla de buscar la solución: dar cada valor posible a la incógnita en todos los sitios de la definición/ecuación donde aparezca y comprobar en qué casos se verifica la ecuación. Tomemos por ejemplo,

$$(TT)\ es\ verdadera =_{def} (TT)\ es\ verdadera \quad (2.11)$$

en vez de usar el método de Gupta y Belnap podemos dar a (TT) todos sus posibles valores y comprobaremos que con todos se verifica la ecuación. Al no haber una única solución podemos afirmar que (TT) es patológica. Si hacemos lo mismo con

$$(L)\ es\ verdadera =_{def} (L)\ es\ falsa \quad (2.12)$$

comprobamos que no hay ninguna solución. También podemos afirmar que (L) es patológica y, a la vez, concluir que su carácter patológico es distinto del de (TT).

Así pues, no veo más ventajas ni menos inconvenientes en la solución de Gupta y Belnap que las que pueda tener una solución basada en ecuaciones. Como en este tipo de soluciones la paradoja del mentiroso reforzada[51] no es resuelta satisfactoriamente. Los autores responden que es patológica pero que, sin embargo, no se puede afirmar, dentro del lenguaje formal, que no es verdadera. En el metalenguaje, ellos mismos lo afirman de todas las oraciones patológicas:

> Revision theory results in a tripartite division of sentences: sentences that are categorically assertible, those that are categorically deniable,

[51] Denominada "Exclusion Liar" por Gupta y Belnap. V. p. 141.

> and those that are pathological in one way or another. This is similar to what one finds in the three-valued approach; here some sentences are true, some others are false, and the remaining ones are neither true nor false.[52]

Es claro que si presuponemos que las oraciones pueden tener n valores de verdad diferentes siempre hay un valor no incluido entre ellos correspondiente al caso de que la definición resulte patológica. Por ejemplo si clasificamos las oraciones en verdaderas, falsas y patológicas, la oración correspondiente a la definición

$$(SL)\ es\ verdadera =_{def} (SL)\ es\ falsa\ o\ patológica \tag{2.19}$$

resulta patológica en el metalenguaje, es decir, es patológico decir que es patológica. Así pues la clasificación original no resulta exhaustiva. Inevitablemente la teoría acaba siendo jerárquica:

> 'the Strengthened Liar[53] is categorical' and 'the Strengthened Liar is not categorical' are pathological [...] Here the concept of categoricalness that is appropriate for describing the behavior of the Ordinary Liar is not appropriate for the Strengthened Liar. To correctly describe the behavior of the latter we need to appeal to a higher-level notion of categoricalness. This higher-level notion would itself manifest paradoxical behavior in the presence of vicious reference. And we would account for it in the same way. The higher-level paradoxes would demand a still higher-level notion for their description.[54]

Una última crítica respecto a su concepto de las definiciones. Creo que su idea de mantener que, incluso en las definiciones circulares, la definición fija el significado del término definido no está justificada. Opino que una definición de la forma "$x\ es\ G =_{def} x\ no\ es\ G$" debería considerarse ilegítima. Gupta y Belnap la consideran correcta y con ello están tergiversando el concepto natural de definición.

De hecho hay una definición oculta detrás de cada definición circular, una definición convencional en el metalenguaje detrás de cada definición circular en

[52] Ibíd., p. 139.
[53] Se refieren a "esta oración no es categórica o no es verdadera".
[54] Ibíd., p. 256.

el lenguaje. Dada una definición de la forma "$x\ es\ G =_{def} A(x, G)$", realmente la definición de "'x es G' es verdadera" es del tipo de

$$\exists p \quad \forall i \geq p \quad \forall X \subseteq D \quad x \in \delta^i(X) \tag{2.20}$$

donde D es el dominio y δ^i es la aplicación de δ i veces.[55]

2.3.8. Propuesta de Skyrms[56]

Skyrms nos presenta la oración del mentiroso reforzada como 'a no es verdadera' ('$\sim Ta$'), siendo 'a' el nombre de esa oración. Da por supuesto, explícitamente, que la identidad

$$a = \text{`} \sim Ta\text{'} \tag{2.21}$$

es cierta ya que considera sostener lo contrario poco verosímil y carente de interés como solución. Sin embargo no lo descarta como último recurso al que acudir si se probase que no es posible una teoría satisfactoria de la autorreferencia semántica.

Para bloquear la paradoja, Skyrms afirma que la oración del mentiroso reforzada carece de significado y, por tanto, no es verdadera. Pero de aquí no se sigue que lo que dice es correcto porque, si carece de significado, no dice nada.

En mi opinión, el problema de este planteamiento es que no podemos afirmar que la oración del mentiroso reforzada no es verdadera, porque, al intentarlo, construiríamos la propia oración que, según Skyrms, carece de significado. Como resultado, la capacidad expresiva del lenguaje queda limitada. Veamos todo esto desde el punto de vista formal.

El autor da por cierto que

$$\sim T\text{`} \sim Ta\text{'} \tag{2.22}$$

Si, teniendo en cuenta (2.21), aplicamos el principio de sustitución de los idénticos obtenemos

$$\sim Ta \tag{2.23}$$

[55] Esta no es exactamente la definición de los autores que requeriría extendernos en más tecnicismos. Simplemente he tratado de expresar la idea de que "x es G" es verdadera si a partir de un cierto número de iteraciones de la función δ, x pertenece a la extensión obtenida ($\delta^i(X)$), cualquiera que sea la extensión de partida (X).

[56] Skyrms (1970).

Finalmente, si lo anterior es verdadero, podemos afirmar

$$T \text{`} \sim Ta\text{'} \tag{2.24}$$

lo que contradice a (2.22).

Si se da por cierto (2.22), es claro que no se puede dar por cierto (2.23), que se ha deducido por aplicación del principio de sustitución de los idénticos, luego, concluye Skyrms, este principio no es correcto. Solo es correcto un principio débil de sustitución de los idénticos por el que de una premisa verdadera no se obtiene una conclusión falsa (se puede obtener una conclusión sin significado cuando ésta es una oración autorreferencial, como ocurre en nuestro ejemplo).

El problema de la falta de capacidad expresiva del lenguaje que plantea la solución de Skyrms podemos expresarlo ahora afirmando que no hay manera de decir en el lenguaje formal que a no es verdadera, pues si escribimos ‘ $\sim Ta$’ formamos una oración carente de significado.

De otro lado, podemos reprochar a la solución de Skyrms que es una solución *ad hoc* al rechazar el principio de sustitución de los idénticos solo porque prescindiendo de él se evita la paradoja. Para que no fuese *ad hoc* habría que justificar por qué (2.21) y (2.22) son verdaderos.

Ahora bien, considero una virtud del planteamiento de Skyrms el mostrar que, una vez que aceptamos (2.22) nos vemos obligados a elegir entre rechazar (2.21) y rechazar el principio de sustitución de los idénticos. Él considera que rechazar (2.21) es una posición desesperada. Yo no lo veo tanto; al fin y al cabo es bastante natural negar que afirmar a sea lo mismo que afirmar “a no es verdadera”. En cambio, rechazar el principio de sustitución de los idénticos sí parece una solución desesperada (recordemos que aquí no estamos en los clásicos contextos referencialmente opacos).

Una crítica, que considero acertada, al rechazo del principio de sustitución de los idénticos de Skyrms, es la realizada por Fitch (1970, p. 75). Este no acepta que a sea idéntico a b y a la vez una oración pueda ser cierta de a y solo neutra, ni verdadera ni falsa, de b. Ello indicaría una diferencia entre a y b.

3

Requisitos para una solución satisfactoria

Tras haber analizado un amplio abanico de intentos de solución de la paradoja del mentiroso, en el que creo se encuentran los más influyentes, es el momento de extraer conclusiones con la mirada puesta en establecer los requisitos que debe satisfacer una solución que merezca tal nombre.

Como indiqué en la introducción, me centraré en la paradoja del mentiroso por ser un buen prototipo de las paradojas semánticas, pero, no cabe duda, de que una solución a esta paradoja sería realmente valiosa si fuese aplicable a las demás paradojas semánticas.

De un modo muy general, el problema de las paradojas consiste en determinar qué principios, hasta el momento considerados como indiscutibles, deben ser revisados de modo que las paradojas desaparezcan. Es preferible que la solución no sea *ad hoc*, es decir, que la modificación de esos principios no esté únicamente justificada porque se resuelven las paradojas. Pero, desde mi punto de vista, lo realmente importante es que la nueva colección de principios no esté aquejada de defectos antes no existentes ni cree problemas mayores o de similar envergadura al que resuelven. A modo de ejemplo, la *solución* de Skyrms genera el problema de cómo puede ser cierto '$a = b$' y a la vez '$\sim Tb$' ser verdadera y '$\sim Ta$' ser neutra (ni verdadera ni falsa). En cambio, el hecho de que una solución sea *ad hoc* no necesariamente la invalida: el razonamiento que indica que la oración del mentiroso, (L), no es verdadera ni falsa porque en ambos casos se deduce una contradicción, podría, en principio, ser válido.

Será muy útil para extraer conclusiones hacer un resumen de los errores que más frecuentemente aparecen en los distintos intentos de solución.

- Se basan en propuestas demasiado artificiosas. Vimos que así ocurre en las propuestas de Russell y Tarski, al jerarquizar el predicado *verdadero* o en la de Skyrms con su forzada restricción del principio de sustitución de los idénticos, o en la de Gupta y Belnap al admitir como legítimas definiciones del tipo "x *es* $G =_{def}$ x *no es* G". Lo mismo puede decirse de distintos aspectos de las propuestas de Kripke, Martin, Barwise y Etchemendy, McGee...

- Son demasiado restrictivos. Las propuestas de Russell restringen la autorreferencia en casos inocuos. Las propuestas jerarquizadas, en general, impiden la expresión de proposiciones acerca de totalidades de proposiciones de la jerarquía: así en la propuesta de Tarski no se puede expresar "toda proposición es verdadera o no". La propuesta de Kripke impide afirmar en el lenguaje objeto que (L) no es verdadera o que una proposición es fundamentada o que es paradójica. Se trata de propuestas que acaban desplazando el problema de las paradojas al lenguaje natural, pero no lo resuelven. Problemas del mismo tipo presentan las de van Fraassen, Gupta y Belnap, Martin...

  No es difícil entender por qué los intentos de solución suelen ser demasiado restrictivos. Para evitar las paradojas se establecen restricciones a los principios normalmente aceptados puesto que conducen a contradicción. El problema es que no hay acuerdo sobre cuáles deben ser estas restricciones y, de hecho, todas parecen restringir demasiado de una u otra forma. Quizás las únicas propuestas estudiadas en este trabajo que no siguen el camino de restringir principios son las *inconsistentes* pero sabemos que tampoco solucionan la paradoja del mentiroso reforzada.

- Tienen otras consecuencias inadmisibles. Como vimos, Gupta y Belnap muestran que las propuestas *inconsistentes* tienen la consecuencia de que todas las oraciones, no solo las paradójicas, resultan verdaderas y falsas a la vez.

- No superan la paradoja del mentiroso reforzada (SL). Esta es una característica común a todas las propuestas. Las que dicen resolverlo no lo hacen de modo satisfactorio. Unas, como la de Kripke, van Fraassen, etc. impiden expresar la oración (SL) trasladando el problema al metalenguaje. Otras, que dicen permitir la expresión de (SL), no permiten, sin embargo, decir uno de

estos enunciados ciertos: (SL) no es verdadera, (SL) es patológica (paradójica, sin valor de verdad), (SL) es falsa o patológica. Como las primeras, están restringiendo la capacidad expresiva del lenguaje con la única motivación de evitar las paradojas. Pero estas reaparecen en el metalenguaje.

Por consiguiente, la paradoja del mentiroso reforzada es una importante piedra de toque de cualquier intento de solucionar las paradojas semánticas.

- Kripke (1975) señala también como defecto de diversas propuestas el hecho de que son meras sugerencias, no verdaderas teorías:

> Almost never is there any precise semantical formulation of a language, at least rich enough to speak of its own elementary syntax (either directly or via arithmetization) and containing its own truth predicate. Only if such a language were set up with formal precision could it be said that a theory of the semantic paradoxes has been presented.[1]

Sin embargo, creo, con Martin, que el problema de las paradojas no es sustancialmente un problema relativo a los lenguajes formalizados:

> I see the Liar as raising questions concerning the concepts of sentence (or statement or proposition), truth, negation, reference, etc.; in short, as a problem in the philosophy of language —our language— not primarily as a problem having to do with formalized languages.[2]

Estamos pues ante un problema de filosofía del lenguaje en el que habrá que explicar qué es lo que falla en las convenciones del lenguaje natural para que aparezcan las paradojas, o desde un punto de vista más positivo, qué convenciones hay que aceptar para que las paradojas desaparezcan. Desde luego, la forma más precisa de mostrar que las paradojas desparecen es presentar la solución mediante un lenguaje formalizado, pero nunca debemos perder de vista que esa solución formal debe estar conectada con el lenguaje natural en el sentido de que debe proporcionar una explicación a las paradojas en dicho lenguaje.

Una vez detectados los principales errores en que no debe caer una solución a las paradojas semánticas, pasaré a indicar otros requisitos que debe cumplir.

[1] Kripke (1975, p. 62).
[2] Martin (1970a, p. 91). La palabra subrayada aparece así en el original.

- Debe tomar como concepto de verdad el concepto de verdad *fuerte* ya que, como hemos visto (p. 52), el concepto de verdad *débil* conduce a conclusiones indeseables.

- Debe establecer un criterio justificado que permita determinar, suponiendo conocidos los hechos empíricos, si las oraciones autorreferenciales como las del mentiroso, del mentiroso reforzada, etc. son verdaderas, falsas o ni verdaderas ni falsas. Es interesante observar que a la misma oración se le adjudica distinto valor de verdad en distintas teorías. Por ejemplo, según Barwise y Etchemendy, la oración del mentiroso es falsa en toda *situación posible*, en tanto que para la mayoría de las teorías no es verdadera ni falsa. La oración "esta oración no es verdadera ni falsa" sería considerada falsa en varias teorías,[3] sin embargo, Mackie considera que no es verdadera ni falsa.[4] Un buen criterio debe dejar claro cuál de las respuestas es correcta.

- Debe evitar introducir conceptos que permitan generar nuevas paradojas.

Resumiendo, una lista de los principales requisitos para una solución satisfactoria es la siguiente:

1. Determinar qué principio o principios normalmente aceptados deben ser revisados para que las paradojas objeto de estudio desaparezcan. Los nuevos principios deben tener una justificación independiente de las paradojas o, al menos, no deben resultar demasiado artificiosos ni acarrear nuevas consecuencias inadmisibles.

2. Solucionar claramente la paradoja del mentiroso reforzada sin limitar la capacidad expresiva del lenguaje.

3. Los principios revisados no deben ser demasiado restrictivos, es decir, no deben evitar situaciones que no eran problemáticas.

4. Debe haber un criterio claro para determinar, suponiendo conocidos los hechos empíricos, si las oraciones autorreferenciales como las del mentiroso, del mentiroso reforzada, etc. son verdaderas, falsas o ni verdaderas ni falsas.

[3] Si no es verdadera ni falsa es verdadera. Si es verdadera, es falsa. En cambio, si es falsa no hay contradicción.

[4] Mackie (1973, p. 291). Él califica de indeterminadas a las oraciones que no son verdaderas ni falsas y su comentario se refiere exactamente a la oración "This sentence, standardly construed, is indeterminate".

5. Debe evitar introducir conceptos poco acertados, como el concepto de verdad *débil*, o conceptos que permitan generar nuevas paradojas.

Finalmente, la solución será mejor cuanto mayor variedad de paradojas resuelva y será más completa si va acompañada de una formulación precisa mediante el uso de lenguajes formales.

4

Caracterización del problema de la paradoja del mentiroso (y afines)

4.1. El problema desde la perspectiva de los lenguajes formales interpretados

4.1.1. Introducción

Entenderemos por paradoja un razonamiento en el que, a partir de supuestos aparentemente verdaderos, y principios aparentemente correctos, se deduce una conclusión aparentemente inaceptable.

En el caso particular de la paradoja del mentiroso (reforzada), una interesante caracterización nos la ofrece Martin (1984, pp. 1 y 2), dando buenas razones para aceptar como verdaderas las dos siguientes afirmaciones:

(S) Hay una oración que dice de sí misma únicamente que no es verdadera

(T) Una oración es verdadera ssi las cosas son como la oración dice que son

Pero —continúa su argumentación— de (S) y (T) se deduce una conclusión inaceptable. Si (S) es verdadera, hay una oración, s, que dice de sí misma únicamente que no es verdadera. La oración s no puede ser verdadera: en efecto, si s fuera verdadera, según (T), y según lo que s dice de sí misma, s no sería verdadera. Tras concluir que s no es verdadera, y puesto que eso es lo único que s dice de sí misma, se deduce, al aplicar (T), que s es verdadera. En resumen, se tiene la

inaceptable conclusión de que *s* no puede ser verdadera y, al mismo tiempo, tiene que ser verdadera.

Entre las razones para aceptar (S), Martin señala la simple existencia de las oraciones "esta oración no es verdadera" o "lo que estoy diciendo ahora no es verdadero". Nosotros podríamos añadir otros ejemplos como:

(R) la oración que hay dentro del recuadro no es verdadera

o, la interesante versión de Quine de la paradoja del mentiroso reforzada:

(Q) "añadida a su propia cita es una oración no verdadera" añadida a su propia cita es una oración no verdadera

Si en las oraciones anteriores cambiamos el no ser verdadero por ser falso, parece inevitable tener que admitir (T) y (S_o):

(S_o) Hay una oración que dice de sí misma únicamente que es falsa

De (T) y (S_o) también se puede llegar a una conclusión inaceptable. En este caso estaríamos ante la paradoja del mentiroso (simple).

El enfoque de Martin tiene la virtud de mostrar con claridad que no debemos asociar la paradoja del mentiroso exclusivamente a una oración concreta. Puede que en algún momento sea adecuado elegir una para facilitar el estudio de la paradoja, pero cualquiera de las oraciones que parece llevarnos inevitablemente a aceptar (S_o) o (S) puede tomarse como oración asociada a la paradoja del mentiroso o del mentiroso reforzada, respectivamente.

Aunque en (S_o) y (S) los predicados "verdadero" y "falso" se aplican a oraciones, no es esencial, para que la paradoja del mentiroso se presente, que consideremos las oraciones como los portadores de verdad. El propio Martin nos dice que(S) puede expresarse de otras formas y pone ejemplos de ello, pero no llega a ofrecernos un sustituto de (S) independiente de cuáles sean los portadores de verdad.

En este trabajo se mejorará la caracterización de la paradoja del mentiroso de Martin encontrando un esquema de dicha paradoja más general que, entre otras virtudes, será independiente de cuáles se consideren los portadores de verdad.

También conseguiremos una caracterización de otras paradojas semánticas estrechamente emparentadas con la del mentiroso; para algunos, simples versiones de la misma. Por ejemplo, entre otras, Simmons (1993, pp. 2-7) señala como tales las paradojas asociadas a las oraciones o sistemas de oraciones siguientes:

(1) Si (1) es verdadera entonces Dios existe

(2) (3) es verdadera

(3) (2) es falsa

(4) (5) es verdadera

(5) (4) es verdadera

(6) (6) es verdadera

La oración (1) corresponde a la llamada paradoja de Curry, por la que se puede demostrar que Dios existe o cualquier afirmación que pongamos en su lugar. Las oraciones (2) y (3) no pueden ser evaluadas consistentemente con los valores de verdad verdadero o falso. Las oraciones (4) y (5) pueden evaluarse ambas como verdaderas o ambas como falsas, pero no hay más motivo para una evaluación que para la otra; el problema es muy similar al de la oración del veraz (6).

Cada una de las versiones de la paradoja tiene variantes "empíricas". Por mencionar solo un ejemplo, supongamos que Platón afirma la oración (7) y, al mismo tiempo, Aristóteles afirma la oración (8):

(7) lo que Aristóteles dice en este momento es verdadero

(8) lo que Platón dice en este momento es falso

Estaríamos ante una paradoja similar a la que corresponde a las oraciones (2) y (3), con la diferencia de que la existencia o no de paradoja depende ahora, no solo de las oraciones sino también de hechos empíricos como que Platón afirme (7) y Aristóteles afirme (8) al mismo tiempo.

Nuestro primer objetivo es pues encontrar un esquema de la paradoja del mentiroso que resulte útil para clarificar el problema, para distinguir los elementos fundamentales a los que se debe la paradoja de aquellos otros simplemente circunstanciales.

Dos enfoques pueden ayudar a conseguirlo. Uno consiste en extraer lo que tienen en común un conjunto lo más amplio posible de paradojas similares. Como hemos visto esto es lo que hizo Russell con su principio del círculo vicioso y más recientemente ha hecho Priest (1994). Otro consiste en desprenderse de las

ambigüedades y complejidades del lenguaje natural y analizar el problema en un lenguaje formal. Naturalmente, no se trata de decir que se soluciona la paradoja si de alguna manera "se soluciona" en el lenguaje formal; el uso de lenguajes formales ha de verse solo como una herramienta para comprenderla mejor.

Considero que antes de pretender comprender un conjunto amplio de paradojas es conveniente entender, lo mejor posible, alguna bien elegida por su carácter paradigmático. Por eso, al menos inicialmente, me centraré en el segundo de los enfoques referidos y, por supuesto, elegiré la paradoja del mentiroso. Después, en la medida de lo posible, será el momento de intentar extender las conclusiones o de buscar planteamientos comunes con otras paradojas.

4.1.2. Una formalización clásica para analizar la paradoja

Se trata de ver qué ocurre en un lenguaje interpretado cuando intentamos reflejar en él las propiedades del lenguaje natural relevantes para la aparición de la paradoja del mentiroso.

Llamaremos lenguaje interpretado a una terna $(\mathcal{L}, \xi, \mathfrak{M})$ formada por un lenguaje, $\mathcal{L}$, una matriz, ξ, que proporciona la interpretación de las constantes lógicas y un modelo, $\mathfrak{M}$, que proporciona la interpretación de las constantes no lógicas. También diremos que, dado un lenguaje, una matriz determina una lógica para ese lenguaje.

Nos restringiremos a lenguajes de primer orden y comenzaremos, en el siguiente apartado, con una lógica clásica, es decir, una lógica bivaluada y una interpretación clásica de las conectivas y cuantificadores. Después añadiremos capacidad de cita a estos lenguajes.

4.1.2.1. Lenguajes interpretados de primer orden clásicos

No basaremos nuestro lenguaje de primer orden, $\mathcal{L}$, en un alfabeto de símbolos minimalista, que haría más incómodo reflejar las propiedades del lenguaje natural en el formal. Por el contrario, supondremos que su alfabeto dispone de los siguientes símbolos:

- Símbolos lógicos:
  - Símbolos de operaciones lógicas o conectivas $(\neg, \wedge, \vee, \rightarrow, \leftrightarrow)$. La negación, $\neg$, es monádica, las demás conectivas son diádicas.

  - Variables (hay un conjunto numerable de variables, V).
  - Cuantificadores ($\exists, \forall$).
  - Símbolos auxiliares (paréntesis, corchetes).

- Símbolo de identidad[1] (=).[2]
- Símbolos no lógicos:
  - Símbolos de oración.
  - Símbolos de constante.
  - Para cada número entero $n \geq 1$, un conjunto (posiblemente vacío) de símbolos de relación n-ádica.[3]
  - Para cada número entero $n \geq 1$, un conjunto (posiblemente vacío) de símbolos de función n-ádica.[4]

Por supuesto, ningún símbolo del alfabeto pertenece a más de una de las clases de símbolos que se acaban de enumerar.

Llamaremos expresión bien formada, o simplemente expresión (formal), a cualquier término o fórmula del lenguaje. Con un alfabeto del tipo anterior, es habitual definir los términos y fórmulas del lenguaje del siguiente modo.

Definición 4.1 (término). *Los términos del lenguaje $\mathcal{L}$ son aquellas cadenas finitas de símbolos de su alfabeto que se obtienen al aplicar, un número de veces finito, las siguientes reglas:*

1. *Cualquier variable o símbolo de constante es un término.*
2. *Si f es un símbolo de función n-ádica y $t_1, t_2, ..., t_n$ son términos, entonces $f\, t_1, t_2, ..., t_n$ es un término.*

[1] Para unos autores es un símbolo lógico para otros no.

[2] Aunque el mismo símbolo se use en el metalenguaje, el contexto nos permitirá reconocer cada caso de uso.

[3] Los símbolos de relación monádica también se denominan símbolos de predicado.

[4] Los que, según es costumbre, llamamos símbolos de función no lo son de cualquier tipo de función sino de operadores con individuos, es decir, funciones que tienen como argumento una secuencia de individuos y, como valor, un individuo.

Como es costumbre, para referirnos genéricamente a cadenas de símbolos del lenguaje (constantes, variables, términos, fórmulas, etc.) usaremos variables sobre las respectivas cadenas de símbolos. Estas variables no pertenecen al lenguaje formal sino al metalenguaje, por lo que se denominan metavariables (por ejemplo, el símbolo "f" utilizado en la definición anterior no es un símbolo del alfabeto sino una metavariable que puede ser sustituida por un símbolo de función n-ádico).

Si lo que deseamos es referirnos a una cadena de símbolos concreta del lenguaje podemos usar un nombre del metalenguaje para hacerlo. Con estos nombres y las metavariables es útil formar esquemas de términos o fórmulas. Un esquema de término/fórmula se caracteriza por convertirse en un término/fórmula al sustituir los nombres y variables del metalenguaje por cadenas de símbolos del lenguaje adecuadas (cada nombre por la cadena a la que se refiere y cada variable por una cadena del tipo de la variable). Por ejemplo, "$f\,t_1, t_2, ..., t_n$" es un esquema de término que se convertirá en un término al sustituir el símbolo "f" por un símbolo de función n-ádico, y cada uno de los símbolos "t_1", "t_2",..., "t_n", por sendos términos.

Definición 4.2 (fórmula). *Las fórmulas (bien formadas) del lenguaje $\mathcal{L}$ son aquellas cadenas finitas de símbolos de su alfabeto que se obtienen al aplicar, un número de veces finito, las siguientes reglas:*

1. *Cualquier símbolo de oración es una fórmula.*
2. *Si t_1 y t_2 son términos, $t_1 = t_2$ es una fórmula.*
3. *Si $t_1, t_2, ..., t_n$son términos y R es un símbolo de relación n-ádico, entonces $R\,t_1, t_2, ..., t_n$ es una fórmula.*
4. *Si φ y ψ son fórmulas, entonces $\neg\varphi$, $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, $(\varphi \rightarrow \psi)$ y $(\varphi \leftrightarrow \psi)$ son fórmulas.*
5. *Si φ es una fórmula y x una variable, entonces $\exists x\,\varphi$ y $\forall x\,\varphi$ son fórmulas.*

La expresión obtenida al sustituir simultáneamente en la expresión formal ε las variables libres $x_1, x_2, ..., x_n$ (todas distintas) por los términos $t_1, t_2, ..., t_n$, respectivamente, la designaremos mediante $\varepsilon\frac{t_1,t_2,...,t_n}{x_1,x_2,...,x_n}$. Para señalar que las variables libres de una expresión formal ε son $x_1, x_2, ..., x_n$, podremos designar dicha expresión mediante $\varepsilon(x_1, x_2, ..., x_n)$. Por ser más intuitivo, preferiremos escribir $\varepsilon(t_1, t_2, ..., t_n)$ en lugar de $\varepsilon(x_1, x_2, ..., x_n)\frac{t_1,t_2,...,t_n}{x_1,x_2,...,x_n}$.

No nos extenderemos aquí en explicar otros conceptos sintácticos de uso corriente como variables libres y ligadas, fórmulas abiertas y cerradas, fórmulas atómicas, subfórmula de una fórmula, etc. Solamente señalaremos que llamaremos oraciones a las fórmulas cerradas. También llamaremos términos abiertos a los que contienen variables y términos cerrados o nombres a los que carecen de ellas.

A continuación, introducimos la matriz, ξ, propia de una lógica clásica:

Definición 4.3 (matriz para la lógica clásica). *La matriz, ξ, que determina una lógica clásica en el lenguaje de primer orden $\mathcal{L}$ está formada por:*

1. *Un conjunto de dos valores de verdad que designaremos $\mathfrak{f}$ (falso) y $\mathfrak{v}$ (verdadero): $W = \{\mathfrak{f}, \mathfrak{v}\}$. Además, estableceremos en este conjunto una relación de orden de modo que $\mathfrak{f}$ sea anterior a $\mathfrak{v}$, es decir,* $\text{mín}(\{\mathfrak{f}, \mathfrak{v}\}) = \mathfrak{f}$.

2. *Una interpretación clásica, C, de las constantes lógicas (conectivas y cuantificadores), es decir, una función total,*[5] *C_k, por cada constante lógica, k, definida del siguiente modo:*

   - *Negación:* $C_\neg : W \to W$; $C_\neg(\mathfrak{f}) =_{def} \mathfrak{v}$; $C_\neg(\mathfrak{v}) =_{def} \mathfrak{f}$.
   - *Conjunción:* $C_\wedge : W^2 \to W$; $C_\wedge(x, y) =_{def} \text{mín}(\{x, y\})$.
   - *Disyunción:* $C_\vee : W^2 \to W$; $C_\vee(x, y) =_{def} \text{máx}(\{x, y\})$.
   - *Condicional:* $C_\to : W^2 \to W$; $C_\to(x, y) =_{def} C_\vee(C_\neg(x), y)$.
   - *Bicondicional:* $C_\leftrightarrow : W^2 \to W$; $C_\leftrightarrow(x, y) =_{def} C_\wedge(C_\to(x, y), C_\to(y, x))$.
   - *Cuantificador universal:* $C_\forall : (\mathscr{P}(W) - \{\varnothing\}) \to W$; $C_\forall(A) =_{def} \text{mín}(A)$.
   - *Cuantificador existencial:* $C_\exists : (\mathscr{P}(W) - \{\varnothing\}) \to W$; $C_\exists(A) =_{def} \text{máx}(A)$.

Para completar nuestro lenguaje interpretado, solo queda añadir el concepto de modelo $\mathfrak{M}$ para $(\mathcal{L}, \xi)$.

[5] Una función total es una función cuyo dominio coincide con el conjunto inicial. Entendemos por dominio de una función a su campo de existencia, es decir, al conjunto de elementos para los que la función está definida. Mediante $dom(f)$ denotaremos el dominio de una función f. Más detalles en apéndice A.

Definición 4.4 (modelo). *Un modelo $\mathfrak{M}$ para $(\mathcal{L}, \xi)$ está formado por:*

1. *Un conjunto no vacío $|\mathfrak{M}|$ llamado dominio o universo de $\mathfrak{M}$.*[6]

2. *Una función de interpretación de los símbolos no lógicos, $\mathfrak{I}_{\mathfrak{M}}$, que asocia:*

   - A cada símbolo de oración, p, un único valor de verdad, $\mathfrak{I}_{\mathfrak{M}}(p)$ (elemento de W).

   - A cada símbolo de constante, c, un único elemento del universo, $\mathfrak{I}_{\mathfrak{M}}(c)$, ($\mathfrak{I}_{\mathfrak{M}}(c) \in |\mathfrak{M}|$).

   - A cada símbolo de relación n-ádica, R, una función total, $\mathfrak{I}_{\mathfrak{M}}(R)$, definida entre $|\mathfrak{M}|^n$ y W.

   - A cada símbolo de función n-ádica, f, una función total, $\mathfrak{I}_{\mathfrak{M}}(f)$, definida entre $|\mathfrak{M}|^n$ y $|\mathfrak{M}|$.

El concepto de asignación de variables juega un importante papel en la interpretación de los términos y fórmulas de un lenguaje.

Definición 4.5 (asignación de variables). *Una asignación de variables para un lenguaje interpretado, $(\mathcal{L}, \xi, \mathfrak{M})$, es una función total $s : V \rightarrow |\mathfrak{M}|$ (V es el conjunto de variables del lenguaje).*

La semántica de la lógica clásica es una semántica referencial, es decir, la interpretación de los términos y fórmulas, su significado, es su referencia.

La interpretación o referencia de una expresión formal, en un lenguaje interpretado de primer orden clásico, $(\mathcal{L}, \xi, \mathfrak{M})_{cls}$,[7] con respecto a una asignación de variables, s, es el valor que la función de valuación, $\mathcal{V}_{\mathfrak{M},s}$, asocie a esa expresión. $\mathcal{V}_{\mathfrak{M},s}$ asocia a cada término un elemento de $|\mathfrak{M}|$ y a cada fórmula un valor de verdad (un elemento de W), de acuerdo con la siguiente definición:[8]

Definición 4.6 (función de valuación). *En un lenguaje interpretado de primer orden clásico, $(\mathcal{L}, \xi, \mathfrak{M})_{cls}$, sean: s, una asignación de variables; c, un símbolo de*

[6] No debe confundirse el dominio de un modelo con el dominio de una función.

[7] El subíndice *cls* lo usaré para recalcar que $\mathcal{L}$, ξ y $\mathfrak{M}$ son del tipo que hemos descrito anteriormente.

[8] Cuando no sea necesario, no especificaremos explícitamente el conjunto inicial y el conjunto final de una función para definirla (aunque, en rigor, de acuerdo con el apéndice A, habría que hacerlo). Si lo hiciésemos para la función $\mathcal{V}_{\mathfrak{M},s}$ diríamos que su conjunto inicial es el conjunto de todas las expresiones formales y, su conjunto final, el conjunto $|\mathfrak{M}| \cup W$.

constante; x, una variable; f, un símbolo de función n-ádica; $t_1, t_2, ..., t_n$, términos; p, un símbolo de oración; R, un símbolo de relación n-ádica; φ y ψ fórmulas. La función de valuación, $\mathcal{V}_{\mathfrak{M},s}$, viene definida por:

1. $\mathcal{V}_{\mathfrak{M},s}(c) =_{def} \mathfrak{I}_{\mathfrak{M}}(c)$.
2. $\mathcal{V}_{\mathfrak{M},s}(x) =_{def} s(x)$.
3. $\mathcal{V}_{\mathfrak{M},s}(f\,t_1, t_2, ..., t_n) =_{def} \mathfrak{I}_{\mathfrak{M}}(f)(\mathcal{V}_{\mathfrak{M},s}(t_1), \mathcal{V}_{\mathfrak{M},s}(t_2), ..., \mathcal{V}_{\mathfrak{M},s}(t_n))$.
4. $\mathcal{V}_{\mathfrak{M},s}(p) =_{def} \mathfrak{I}_{\mathfrak{M}}(p)$.
5. $\mathcal{V}_{\mathfrak{M},s}(t_1 = t_2) =_{def} \begin{cases} \mathfrak{v} & si\ \mathcal{V}_{\mathfrak{M},s}(t_1) = \mathcal{V}_{\mathfrak{M},s}(t_2) \\ \mathfrak{f} & si\ \mathcal{V}_{\mathfrak{M},s}(t_1) \neq \mathcal{V}_{\mathfrak{M},s}(t_2) \end{cases}$.
6. $\mathcal{V}_{\mathfrak{M},s}(R\,t_1, t_2, ..., t_n) =_{def} \mathfrak{I}_{\mathfrak{M}}(R)(\mathcal{V}_{\mathfrak{M},s}(t_1), \mathcal{V}_{\mathfrak{M},s}(t_2), ..., \mathcal{V}_{\mathfrak{M},s}(t_n))$.
7. $\mathcal{V}_{\mathfrak{M},s}(\neg\varphi) =_{def} C_{\neg}(\mathcal{V}_{\mathfrak{M},s}(\varphi))$.
8. $\mathcal{V}_{\mathfrak{M},s}((\varphi \wedge \psi)) =_{def} C_{\wedge}(\mathcal{V}_{\mathfrak{M},s}(\varphi), \mathcal{V}_{\mathfrak{M},s}(\psi))$.
9. $\mathcal{V}_{\mathfrak{M},s}((\varphi \vee \psi)) =_{def} C_{\vee}(\mathcal{V}_{\mathfrak{M},s}(\varphi), \mathcal{V}_{\mathfrak{M},s}(\psi))$.
10. $\mathcal{V}_{\mathfrak{M},s}((\varphi \rightarrow \psi)) =_{def} C_{\rightarrow}(\mathcal{V}_{\mathfrak{M},s}(\varphi), \mathcal{V}_{\mathfrak{M},s}(\psi))$.
11. $\mathcal{V}_{\mathfrak{M},s}((\varphi \leftrightarrow \psi)) =_{def} C_{\leftrightarrow}(\mathcal{V}_{\mathfrak{M},s}(\varphi), \mathcal{V}_{\mathfrak{M},s}(\psi))$.
12. $\mathcal{V}_{\mathfrak{M},s}(\forall x\,\varphi) =_{def} C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\varphi)/e \in| \mathfrak{M} |\})$ *donde* $s[e/x]$ *es igual que* s *salvo que asocia* e *a la variable* x*, es decir,*

$$s[e/x](y) =_{def} \begin{cases} s(y) & si\ x\ e\ y\ no\ son\ la\ misma\ variable \\ e & si\ x\ e\ y\ son\ la\ misma\ variable \end{cases}$$

13. $\mathcal{V}_{\mathfrak{M},s}(\exists x\,\varphi) =_{def} C_{\exists}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\varphi)/e \in| \mathfrak{M} |\})$.

Supuestos un lenguaje $\mathcal{L}$ y una matriz ξ, cuando $\mathcal{V}_{\mathfrak{M},s}(\varphi) = \mathfrak{v}$, se dice que φ es verdadera en $\mathfrak{M}$ con respecto a la asignación s, o que $\mathfrak{M}$ satisface φ con respecto a la asignación s. Esto puede expresarse del siguiente modo:

$$\mathfrak{M}, s \models \varphi$$

El valor que la función $\mathcal{V}_{\mathfrak{M},s}$ asocia a algunas fórmulas y términos es independiente de la asignación de variables s que se tome. Esto nos permite definir la función de valuación, $\mathcal{V}_{\mathfrak{M}}$, de esos términos y fórmulas:

Definición 4.7. *Sea ε una expresión formal (término o fórmula). La expresión ε pertenece al dominio de la función $\mathcal{V}_{\mathfrak{M}}$ ssi para cualesquiera asignaciones de variable, r y s, se cumple $\mathcal{V}_{\mathfrak{M},r}(\varepsilon) = \mathcal{V}_{\mathfrak{M},s}(\varepsilon)$. En tal caso, $\mathcal{V}_{\mathfrak{M}}(\varepsilon) =_{def} \mathcal{V}_{\mathfrak{M},s}(\varepsilon)$.*

Se dice que la fórmula φ es verdadera en $\mathfrak{M}$, o que $\mathfrak{M}$ satisface φ, cuando $\mathcal{V}_{\mathfrak{M}}(\varphi) = \mathfrak{v}$. Esto suele expresarse del siguiente modo:

$$\mathfrak{M} \models \varphi$$

Es interesante señalar que si dos asignaciones de variable, r y s, asocian los mismos valores a cada una de las variables libres de una expresión formal, ε, entonces $\mathcal{V}_{\mathfrak{M},r}(\varepsilon) = \mathcal{V}_{\mathfrak{M},s}(\varepsilon)$. Si la expresión formal no tiene variables libres, el valor que la función $\mathcal{V}_{\mathfrak{M},s}$ le asocia será independiente de s. Consecuentemente, la función $\mathcal{V}_{\mathfrak{M}}$ está definida para todo término cerrado y para toda oración.

Fijados un lenguaje $\mathcal{L}$ y una matriz ξ, diremos que la fórmula ψ es consecuencia lógica de la fórmula φ cuando para cualquier modelo, $\mathfrak{M}$, y cualquier asignación de variables, s, si $\mathcal{V}_{\mathfrak{M},s}(\varphi) = \mathfrak{v}$ entonces $\mathcal{V}_{\mathfrak{M},s}(\psi) = \mathfrak{v}$. Esto se expresa habitualmente mediante:

$$\varphi \models \psi$$

También diremos que las fórmulas φ y ψ son lógicamente equivalentes cuando una es consecuencia lógica de la otra y viceversa. Por tanto, las fórmulas φ y ψ son lógicamente equivalentes si y solo si para cualquier modelo, $\mathfrak{M}$, y cualquier asignación de variables, s, $\mathcal{V}_{\mathfrak{M},s}(\varphi) = \mathcal{V}_{\mathfrak{M},s}(\psi)$.

Finalmente, debemos destacar algunas características relevantes de la semántica de la lógica clásica.

En primer lugar, como ya se ha señalado, es una semántica referencial.

En segundo lugar, la semántica clásica sigue el principio de composicionalidad: la interpretación de una expresión (término o fórmula) estructurada se obtiene a partir de la estructura sintáctica de la misma y de la interpretación de sus partes. Esto es manifiesto si miramos la definición 4.6, excepto en el caso de fórmulas cuantificadas. Sin embargo, como señala Janssen (1997, p. 423), la composicionalidad del significado de todas las fórmulas compuestas se restablece si tomamos como su significado el conjunto de asignaciones de variable para los que la fórmula es verdadera.[9]

A continuación establecemos una definición más general de referencia de una expresión —por ser aplicable tanto a fórmulas como a términos— que restablece la composicionalidad:

[9] De esta forma el concepto de interpretación de una fórmula es relativo a un modelo pero no hay que relativizarlo a una asignación de variable.

Definición 4.8 (referencia de una expresión). *La referencia de una expresión formal ε, en un lenguaje interpretado con un modelo $\mathfrak{M}$, es la función $\mathcal{R}_{\mathfrak{M},\varepsilon}$ definida (entre el conjunto de asignaciones de variable y $W \cup \mid \mathfrak{M} \mid$) del siguiente modo:*

$$\mathcal{R}_{\mathfrak{M},\varepsilon}(s) =_{def} \mathcal{V}_{\mathfrak{M},s}(\varepsilon) \tag{4.1}$$

Obsérvese que cuando ε es una fórmula, $\mathcal{R}_{\mathfrak{M},\varepsilon}$ asocia a cada asignación de variables, s, un valor de verdad por lo que determina el conjunto de asignaciones de variable para los que la fórmula es verdadera.

A veces, para indicar que dos expresiones formales ε_1 y ε_2 tienen la misma referencia escribiremos la igualdad de funciones $\mathcal{R}_{\mathfrak{M},\varepsilon_1} = \mathcal{R}_{\mathfrak{M},\varepsilon_2}$, la cual equivale a decir que para toda asignación de variables, s, $\mathcal{V}_{\mathfrak{M},s}(\varepsilon_1) = \mathcal{V}_{\mathfrak{M},s}(\varepsilon_2)$.

Por otra parte, en el caso de que ε sea una expresión cerrada, tendremos $\mathcal{R}_{\mathfrak{M},\varepsilon}(s) = \mathcal{V}_{\mathfrak{M},s}(\varepsilon) = \mathcal{V}_{\mathfrak{M}}(\varepsilon)$. Aunque en rigor, según la definición anterior, la referencia de la expresión cerrada ε sería una función que asocia a toda asignación de variables el valor $\mathcal{V}_{\mathfrak{M}}(\varepsilon)$, acostumbraremos a decir simplemente que la referencia de ε es $\mathcal{V}_{\mathfrak{M}}(\varepsilon)$. Además, sin perder ningún rigor, podemos decir que dos expresiones cerradas ε_1 y ε_2 tienen la misma referencia si y solo si $\mathcal{V}_{\mathfrak{M}}(\varepsilon_1) = \mathcal{V}_{\mathfrak{M}}(\varepsilon_2)$.

Por estar dotado de una semántica referencial y composicional, un lenguaje interpretado de primer orden clásico es extensional lo que significa que dentro de una fórmula o de un término se puede sustituir un elemento (término, fórmula, símbolo de oración, de relación o de función) por otro con la misma referencia o extensión sin que cambie la referencia de la fórmula o del término.

4.1.2.2. Capacidad de cita

Ahora queremos añadir capacidad de cita a los lenguajes interpretados de primer orden clásicos. Esto viene motivado por nuestra intención de reflejar en el lenguaje formal las propiedades del lenguaje natural relevantes para la aparición de la paradoja del mentiroso.

Como hemos visto, siguiendo la caracterización de Martin, la paradoja del mentiroso (reforzada) surge cuando aceptamos la verdad de las dos siguientes afirmaciones:

(S) Hay una oración que dice de sí misma únicamente que no es verdadera

(T) Una oración es verdadera ssi las cosas son como la oración dice que son

Es claro que, para reflejar cualquiera de ellas en el lenguaje formal, éste necesita tener capacidad de referirse a sus propias oraciones. La forma más explícita de referirse a una oración en el lenguaje natural es entrecomillarla. Como se indica a continuación, este modo de referencia explícita puede añadirse fácilmente a un lenguaje interpretado de primer orden clásico, $(\mathcal{L}_0, \xi, \mathfrak{M}_0)_{cls}$, y así obtener un lenguaje interpretado, $(\mathcal{L}_1, \xi, \mathfrak{M}_1)_{cit}$, con capacidad de cita.

Formamos el albafeto de $\mathcal{L}_1$ añadiendo al alfabeto de $\mathcal{L}_0$ los símbolos de cita:[10] $\ulcorner$ y $\urcorner$. El propósito de estos símbolos es poder citar una cadena cualquiera de símbolos del alfabeto, κ, mediante un término de cita, $\ulcorner\kappa\urcorner$, es decir, un nombre obtenido encerrando la cadena entre los símbolos de cita. Si llamamos $\mathcal{A}_1$ al alfabeto de símbolos de $\mathcal{L}_1$ y $\mathcal{A}_1^*$ al conjunto de cadenas finitas de símbolos de $\mathcal{A}_1$, lo que necesitamos es: a) que una regla sintáctica de $\mathcal{L}_1$ sea "si $\kappa \in \mathcal{A}_1^*$, entonces $\ulcorner\kappa\urcorner$ es un término" (a los términos de este tipo los denominaremos términos de cita); b) que las cadenas finitas de símbolos de $\mathcal{L}_1$ pertenezcan al universo de $\mathfrak{M}_1$. En consecuencia, este universo lo estableceremos como:

$$\mid \mathfrak{M}_1 \mid = \mid \mathfrak{M}_0 \mid \cup \{\kappa / \kappa \in \mathcal{A}_1^*\} \tag{4.2}$$

Desde el punto de vista semántico, escogeremos un modelo $\mathfrak{M}_1$ de $\mathcal{L}_1$ que sea igual que $\mathfrak{M}_0$ pero con el añadido de que la interpretación de los términos de cita ha de ser la cadena de símbolos en ellos citada. Es decir,

$$para\, todo\, \kappa \in \mathcal{A}_1^* : \mathfrak{I}_{\mathfrak{M}_1}(\ulcorner\kappa\urcorner) = \kappa \tag{4.3}$$

En cuanto a $\mathcal{V}_{\mathfrak{M}_1}(\ulcorner\kappa\urcorner)$, es claro que, puesto que $\mathfrak{I}_{\mathfrak{M}_1}(\ulcorner\kappa\urcorner)$ es un *valor* (un elemento del universo $\mid \mathfrak{M}_1 \mid$), debemos establecer $\mathcal{V}_{\mathfrak{M}_1}(\ulcorner\kappa\urcorner) = \mathfrak{I}_{\mathfrak{M}_1}(\ulcorner\kappa\urcorner)$. Con más precisión y dado que $\mathfrak{I}_{\mathfrak{M}_1}(\ulcorner\kappa\urcorner) = \kappa$, estableceremos:

$$para\, todo\, \kappa \in \mathcal{A}_1^* : \mathcal{V}_{\mathfrak{M}_1}(\ulcorner\kappa\urcorner) = \kappa \tag{4.4}$$

El lenguaje interpretado, $(\mathcal{L}_1, \xi, \mathfrak{M}_1)_{cit}$, que acabamos de obtener precisa corregir un problema con el entrecomillado tal como lo hemos introducido: dado que en $\mathcal{L}_1$, la propia cadena de símbolos κ puede contener símbolos de cita, la expresión

[10] Se supone que el alfabeto de $\mathcal{L}_0$ no contiene los símbolos de cita.

$\ulcorner\kappa\urcorner$ puede resultar ambigua.[11] El problema se resuelve usando métodos de cita más elaborados como los que aparecen en Boolos (1998, capítulo 28) o en Little (2003). Nosotros, en cambio, lo solucionaremos restringiendo el tipo de cadenas de símbolos que se pueden citar a términos y fórmulas,[12] es decir, sustituiremos, en $\mathcal{L}_1$, la regla sintáctica "si $\kappa \in \mathcal{A}_1^*$, entonces $\ulcorner\kappa\urcorner$ es un término" por esta otra: "si ε es un término o fórmula de $\mathcal{L}_1$, entonces $\ulcorner\varepsilon\urcorner$ es un término".[13]

4.1.2.3. Predicado de verdad y predicados de desentrecomillado

Una vez que disponemos de nuestro lenguaje interpretado con capacidad de cita, estamos en condiciones de establecer qué requisitos tendría que cumplir dicho lenguaje para reflejar los supuestos, aparentemente ciertos, (S) y (T) que, según Martin, conducen a la paradoja del mentiroso. Comencemos por el segundo:

(T) Una oración es verdadera ssi las cosas son como la oración dice que son

Esto queda plasmado en nuestro lenguaje, si posee un predicado de verdad de acuerdo con la siguiente caracterización: en un lenguaje interpretado de primer orden con capacidad de cita $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$, el símbolo de predicado T representa un predicado de verdad ssi para toda oración, φ, se verifica

$$\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\varphi) \tag{4.5}$$

es decir,

$$\mathfrak{M} \models T^{\ulcorner}\varphi^{\urcorner} \leftrightarrow \varphi \tag{4.6}$$

—lo que indica que $\mathfrak{M}$ satisface toda oración con la forma del esquema T de Tarski—.

(4.5) pone de manifiesto que el predicado T realiza una tarea de desentrecomillado. También realizan una tarea de desentrecomillado otros posibles predicados

[11] Por ejemplo, en un determinado lenguaje, cabría entender la expresión $\ulcorner a\urcorner R\ulcorner b\urcorner$ como un término que cita la cadena de símbolos $a\urcorner R\ulcorner b$ o como una fórmula que establece la relación R entre los términos de cita $\ulcorner a\urcorner$ y $\ulcorner b\urcorner$.

[12] No necesitamos más para analizar las oraciones autorreferenciales implicadas en las paradojas semánticas.

[13] Con esta regla todo símbolo $\ulcorner$ tiene su correspondiente símbolo $\urcorner$ de modo no ambiguo, con el mismo mecanismo de emparejamiento habitual de los paréntesis.

semánticos. Por ejemplo, el predicado "falso" podría formalizarse mediante el símbolo F y debería cumplirse, para cualquier oración, φ:

$$\mathcal{V}_{\mathfrak{M}}(F^{\ulcorner}\varphi^{\urcorner}) = C_F(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.7}$$

donde C_F será una función total, definida entre W y W, mediante $C_F(\mathfrak{f}) = \mathfrak{v}$, $C_F(\mathfrak{v}) = \mathfrak{f}$.

Podemos, en general, decir que el símbolo de predicado D representa un predicado de desentrecomillado ssi para toda oración, φ, se verifica

$$\mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\varphi^{\urcorner}) = C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.8}$$

donde C_D es una función total, definida entre W y W.[14]

La caracterización anterior se puede generalizar considerando que φ puede ser cualquier fórmula en el dominio de la función $\mathcal{V}_{\mathfrak{M}}$ (no solamente cualquier oración).

Si tenemos en cuenta cómo hemos definido los lenguajes interpretados de primer orden con capacidad de cita y tomamos un término, ν, que tenga como referencia la oración φ ($\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$), es fácil comprobar que $\mathcal{V}_{\mathfrak{M}}(D\nu) = \mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\varphi^{\urcorner})$.[15]

Juntando las dos generalizaciones anteriores, podemos establecer la siguiente definición:

Definición 4.9 (predicado de desentrecomillado). *Un símbolo de predicado D representa un predicado de desentrecomillado ssi para todo término, ν, que tenga como referencia una fórmula φ perteneciente al dominio de $\mathcal{V}_{\mathfrak{M}}$, se verifica*

$$\mathcal{V}_{\mathfrak{M}}(D\nu) = C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.9}$$

donde C_D es una función total, definida entre W y W.

De modo similar a como hemos definido los predicados de desentrecomillado se pueden definir funciones de desentrecomillado. Por ejemplo, una función de denotación, h, se caracterizaría porque dado un término, τ, con $\tau \in dom(\mathcal{V}_{\mathfrak{M}})$, se verificará $\mathcal{V}_{\mathfrak{M}}(h^{\ulcorner}\tau^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\tau)$. En general, una función de desentrecomillado, g, se caracterizará porque dado un término ν cuya referencia sea un término, τ, con $\tau \in dom(\mathcal{V}_{\mathfrak{M}})$, se verificará $\mathcal{V}_{\mathfrak{M}}(g\nu) = C_g(\mathcal{V}_{\mathfrak{M}}(\tau))$, siendo C_g una función total definida entre $\mid \mathfrak{M} \mid$ y $\mid \mathfrak{M} \mid$.

[14] En el caso del predicado T anterior la función total C_T era simplemente la función identidad.

[15] $\mathcal{V}_{\mathfrak{M}}(D\nu) = \mathfrak{I}_{\mathfrak{M}}(D)(\mathcal{V}_{\mathfrak{M}}(\nu)) = \mathfrak{I}_{\mathfrak{M}}(D)(\varphi) = \mathfrak{I}_{\mathfrak{M}}(D)(\mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\varphi^{\urcorner})) = \mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\varphi^{\urcorner})$.

Aunque cabe pensar en relaciones y funciones con varios argumentos que *desentrecomillen* algunos de ellos, no los necesitamos para el estudio formal de la paradoja del mentiroso, por lo cual en los lenguajes interpretados que usemos en este trabajo solo admitiremos relaciones y funciones de desentrecomillado monádicas.

4.1.2.4. Extensionalidad de los lenguajes interpretados de primer orden con capacidad de cita

Si, dentro de una fórmula o término, sustituimos una expresión formal que aparece de forma citada por otra con la misma referencia no hay garantías de que no cambie la referencia de la fórmula o del término, pues esta dependerá con frecuencia de la propia expresión citada y no de su referencia. Es decir, como es bien sabido, generalmente el entrecomillado destruye la referencialidad de las fórmulas y términos delimitados por las comillas (los símbolos de cita). Sin embargo, no siempre es así. Tomemos una oración φ y el término $\ulcorner\varphi\urcorner$ que la cita. Si este término va precedido de un predicado de desentrecomillado, D, formamos una oración, $D\ulcorner\varphi\urcorner$, cuya referencia —su valor de verdad— depende de la referencia de φ. Por tanto, si sustituimos en $D\ulcorner\varphi\urcorner$ la oración φ por otra oración ψ con el mismo valor de verdad la referencia de la oración resultante no cambia. Más formalmente: si $\mathcal{V}_{\mathfrak{M}}(\varphi) = \mathcal{V}_{\mathfrak{M}}(\psi)$ entonces $\mathcal{V}_{\mathfrak{M}}(D\ulcorner\varphi\urcorner) = \mathcal{V}_{\mathfrak{M}}(D\ulcorner\psi\urcorner)$, porque $\mathcal{V}_{\mathfrak{M}}(D\ulcorner\varphi\urcorner) = C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)) = C_D(\mathcal{V}_{\mathfrak{M}}(\psi)) = \mathcal{V}_{\mathfrak{M}}(D\ulcorner\psi\urcorner)$.

Definición 4.10 (aparición referencial de una expresión). *Sean ε y δ dos expresiones formales. Una aparición de ε es* referencial respecto a δ *cuando: a) la aparición de ε se produce en δ y b) si ε' es una expresión con la misma referencia que ε, entonces la expresión δ', resultante de sustituir en δ la aparición de ε por ε', tiene la misma referencia que δ (abreviadamente: si $\mathcal{R}_{\mathfrak{M},\varepsilon} = \mathcal{R}_{\mathfrak{M},\varepsilon'}$, entonces $\mathcal{R}_{\mathfrak{M},\delta} = \mathcal{R}_{\mathfrak{M},\delta'}$).*[16]

Por tanto, cuando una aparición de ε es *referencial respecto a* δ la aportación de esa aparición de ε a la referencia de δ es únicamente su referencia.

Un método recursivo para determinar si una aparición de ε es referencial respecto a δ es el siguiente. ε es referencial respecto a δ ssi se cumple uno de los tres requisitos que se indican a continuación:

[16] Ver definición 4.8 en p. 75 y el comentario posterior sobre la igualdad de funciones de referencia $\mathcal{R}_{\mathfrak{M},\varepsilon_1} = \mathcal{R}_{\mathfrak{M},\varepsilon_2}$.

1. La aparición de ε no forma parte de ningún término de cita contenido en δ.

2. D es un predicado de desentrecomillado, ζ es una fórmula, $\zeta \in dom(\mathcal{V}_{\mathfrak{M}})$, la aparición de ε es referencial respecto a ζ y se da en una aparición de la oración $D^{\ulcorner}\zeta^{\urcorner}$ en δ que es referencial respecto a δ.

3. g es una función de desentrecomillado, ζ es un término, $\zeta \in dom(\mathcal{V}_{\mathfrak{M}})$, la aparición de ε es referencial respecto a ζ y se da en una aparición del término $g^{\ulcorner}\zeta^{\urcorner}$ en δ que es referencial respecto a δ.

El punto primero del procedimiento anterior está justificado por la referencialidad del lenguaje cuando no intervienen términos de cita.

El punto segundo está justificado porque al sustituir la aparición de ε por otra expresión formal, ε', con la misma referencia: a) la fórmula ζ se convertirá en otra fórmula, ζ', con la misma referencia, dado que la aparición de ε es referencial respecto a ζ; b) la fórmula $D^{\ulcorner}\zeta'^{\urcorner}$ tiene la misma referencia que $D^{\ulcorner}\zeta^{\urcorner}$ dado que $\mathcal{V}_{\mathfrak{M}}(\zeta) = \mathcal{V}_{\mathfrak{M}}(\zeta')$ y D es un predicado de desentrecomillado ($\mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\zeta^{\urcorner}) = C_D(\mathcal{V}_{\mathfrak{M}}(\zeta)) = C_D(\mathcal{V}_{\mathfrak{M}}(\zeta')) = \mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\zeta'^{\urcorner})$); c) la expresión δ' resultante de sustituir en δ la aparición de $D^{\ulcorner}\zeta^{\urcorner}$ por $D^{\ulcorner}\zeta'^{\urcorner}$, tiene la misma referencia que δ porque la aparición de $D^{\ulcorner}\zeta^{\urcorner}$ es referencial respecto a δ y la referencia de $D^{\ulcorner}\zeta^{\urcorner}$ es la misma que la de $D^{\ulcorner}\zeta'^{\urcorner}$.

Obsérvese que si la aparición de ε no fuese referencial respecto a ζ o la de $D^{\ulcorner}\zeta^{\urcorner}$ no fuese referencial respecto a δ, o el predicado D no fuese un predicado de desentrecomillado, la aparición de ε no sería referencial respecto a δ.

Los comentarios respecto al punto tercero son similares cambiando el predicado D por la función g y la fórmula $D^{\ulcorner}\zeta^{\urcorner}$ por el término $g^{\ulcorner}\zeta^{\urcorner}$.

Los siguientes ejemplos en castellano son ilustrativos de las diferentes situaciones de una expresión cerrada dentro de otra expresión. Respecto a

"Clarín" tiene seis letras

el nombre "Clarín" no es referencial (porque el predicado "tener seis letras" no es un predicado de desentrecomillado). Respecto a

"Clarín fue un escritor" es verdadero

la oración "Clarín fue un escritor" sí es referencial (el predicado "verdadero" es un predicado de desentrecomillado) y el nombre "Clarín" también es referencial

(porque es referencial respecto a "Clarín fue un escritor" y esta oración es a su vez referencial respecto a ""Clarín fue un escritor" es verdadero"). Respecto a

""Clarín fue un escritor" es verdadero" tiene treinta letras

el nombre "Clarín" no es referencial, porque aunque lo es respecto a ""Clarín fue un escritor" es verdadero", esta oración no lo es respecto a la oración completa. Respecto a

""Clarín fue un escritor" tiene seis letras" es verdadero

el nombre "Clarín" no es referencial, porque no lo es respecto a ""Clarín fue un escritor" tiene seis letras". Respecto a

"la persona llamada "Clarín" fue un escritor" es verdadero

el nombre "Clarín" es referencial, porque el nombre "Clarín" es referencial respecto a "la persona llamada "Clarín" fue un escritor" —dado que la función "la persona llamada _" es una función de desentrecomillado— y el predicado "verdadero" es un predicado de desentrecomillado. Finalmente respecto a

"Clarín" tiene seis letras y "Clarín" es el nombre de un escritor

la primera aparición de "Clarín" no es referencial pero la segunda sí.

Es fácil comprobar en los ejemplos anteriores que al sustituir una aparición de "Clarín" en una oración por otro nombre con la misma referencia como es "Leopoldo Alas", el valor de verdad de la oración resultante no cambia cuando la aparición de "Clarín" sustituida es referencial respecto a la oración considerada.

Definición 4.11 (parte propia). *La expresión, ε, constituye una parte propia de la expresión δ si y solo si una aparición de ε en δ es referencial respecto a δ y $\varepsilon \neq \delta$.*

Los símbolos de relación y símbolos de función que aparezcan en δ son partes propias de δ cuando no aparecen dentro de un término de cita contenido en δ o cuando, siendo así, forman parte propia de la aparición de una expresión, ε, referencial respecto a δ.

De este modo la extensionalidad de los lenguajes interpretados de primer orden con capacidad de cita se conserva si precisamos que para que se mantenga la referencia de una expresión formal solo se puede sustituir un elemento por otro con la misma referencia cuando el elemento sustituido era una *parte propia* de la expresión.

También se conservan la composicionalidad y la referencialidad si cambiamos *parte* por *parte propia*. Es decir, en un lenguaje interpretado de primer orden con capacidad de cita: a) la interpretación de una expresión formal estructurada se obtiene a partir de la estructura sintáctica de la misma y de la interpretación de sus partes propias; b) la interpretación de una expresión es su referencia.

4.1.2.5. Autorreferencia

Para reflejar en un lenguaje interpretado el supuesto (S) de Martin (hay una oración que dice de sí misma únicamente que no es verdadera) es necesario que el lenguaje tenga cierta capacidad de autorreferencia.

Conviene precisar la noción de autorreferencia antes de ver los modos de conseguirla. En general, una entidad es autorreferencial si alguna parte propia de ella se refiere a la totalidad de la entidad. De esta idea aplicada a oraciones surge la siguiente definición.

Definición 4.12 (oración autorreferencial). *Una oración es autorreferencial (en un lenguaje interpretado) ssi un término propio de ella tiene como referencia la oración.*

Si llamamos $\mathfrak{M}$ al modelo, τ al término y $\sigma(\tau)$ a la oración autorreferencial, tendremos

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \sigma(\tau) \tag{4.10}$$

En un lenguaje interpretado bivaluado con capacidad de cita, $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$, dado que $\mathcal{V}_{\mathfrak{M}}(\ulcorner\sigma(\tau)\urcorner) = \sigma(\tau)$, también podremos escribir

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\sigma(\tau)\urcorner) \tag{4.11}$$

de donde, si el lenguaje tiene símbolo de identidad, resulta

$$\mathfrak{M} \models \tau = \ulcorner\sigma(\tau)\urcorner \tag{4.12}$$

Definición 4.13 (lenguaje fuertemente autorreferencial). *Un lenguaje interpretado es fuertemente autorreferencial ssi para toda fórmula con una única variable libre, $\sigma(x)$, existe un término, τ, cuya referencia es la oración $\sigma(\tau)$.*

De esta definición se deduce que $\sigma(\tau)$ es una oración autorreferencial y que, si llamamos $\mathfrak{M}$ al modelo, se cumplirá (4.10). Si además el lenguaje tiene capacidad de cita y símbolo de identidad, también se cumplirán, (4.11) y (4.12).

Otra consecuencia de la definición es:

Proposición 4.1. *Si un lenguaje interpretado extensional y con capacidad de cita, $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$, es fuertemente autorreferencial, entonces para toda fórmula con una única variable libre, $\sigma(x)$, existe un término, τ, tal que:*

$$\mathcal{V}_{\mathfrak{M}}(\sigma(\tau)) = \mathcal{V}_{\mathfrak{M}}(\sigma(\ulcorner\sigma(\tau)\urcorner)) \tag{4.13}$$

Demostración. Teniendo en cuenta que el lenguaje es fuertemente autorreferencial, para toda fórmula, $\sigma(x)$, con una única variable libre, existirá un término, τ, tal que $\mathcal{V}_{\mathfrak{M}}(\tau) = \sigma(\tau)$, (4.10). Como el lenguaje tiene capacidad de cita, (4.10) equivale a $\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\sigma(\tau)\urcorner)$, (4.11); y dado que su semántica es extensional, si se cumple (4.11) se cumple (4.13). ■

Hay varios modos de conseguir que nuestro lenguaje contenga oraciones autorreferenciales. Aparentemente el más sencillo es utilizar un término *this* reflejo en el lenguaje formal de la expresión denotativa "esta oración" del lenguaje natural.[17] Un primer intento de expresar la oración del mentiroso reforzada nos daría $\neg T\,this$, pero aquí quedaría ambigua la referencia de *this*: podría ser $T\,this$ o $\neg T\,this$. La ambigüedad es resuelta por Barwise y Etchemendy (1987) añadiendo al alfabeto un símbolo de ámbito, $\downarrow$, y a las reglas sintácticas la regla de que si φ es una fórmula también lo es $\downarrow \varphi$ (op. cit., p. 32). Esto permite distinguir la oración del mentiroso reforzada, $\downarrow \neg T\,this$, de la negación de la oración del veraz, $\neg \downarrow T\,this$.

La adecuada introducción del término *this* en nuestro lenguaje permite encontrar, para toda fórmula con una única variable libre, $\sigma(x)$, que no contenga el símbolo $\downarrow$, la oración autorreferencial $\downarrow \sigma(this)$. Esto no cumple literalmente nuestra definición de lenguaje fuertemente autorreferencial aunque encaja en el

[17] Variantes del uso de "this" en lenguajes formales aquí presentado pueden verse en Barwise y Etchemendy (1987) y en Barwise y Moss (1996, capítulo 13).

espíritu de la misma. Otro detalle incómodo es que el término *this*, a diferencia de los demás términos, tiene una referencia sensible al contexto (es un demostrativo). No se puede preguntar simplemente cuál es la referencia de *this* sino cuál es la referencia de *this* cuando aparece en una oración $\downarrow \sigma$. Pero como la respuesta es la propia oración $\downarrow \sigma$, se puede decir que nos encontramos con la curiosa situación de que para elaborar la pregunta tenemos que usar la respuesta.

Probablemente, el principal motivo que puede alegarse para considerar poco adecuado el uso de *this* en el análisis de la paradoja del mentiroso es que la referencia de ese término demostrativo se establece: ha de ser la fórmula, dentro del ámbito del símbolo $\downarrow$, a la que pertenece. En cierto sentido, la autorreferencia se impone con el uso de *this* sin someterse a discusión. Pero la paradoja del mentiroso podría indicar que el término *this* no siempre puede tener la referencia que se le supone. Con otras palabras, una escapatoria a la paradoja podría consistir en afirmar que realmente el término *this* en su aparición en la oración $\downarrow \neg T\, this$ carece de referencia. Sin embargo, Quine parece cerrarnos esa salida con su versión de la paradoja del mentiroso reforzada:

(Q) "añadida a su propia cita, es una oración no verdadera" añadida a su propia cita, es una oración no verdadera

en la que se consigue la autorreferencia sin usar la expresión "esta oración". En la misma línea se manifiesta Hofstadter (1998, p. 554) para quien, en la oración de Quine, la autorreferencialidad es concretada de modo más directo, más explícito, que en la oración del mentiroso *convencional*.

La oración de Quine se ha formado, en lenguaje natural, mediante la operación de hacer preceder una secuencia de caracteres por su propia cita. En un lenguaje formal se puede conseguir autorreferencia haciendo seguir a una una secuencia de caracteres de su propia cita:[18] es lo que Smullyan (1996, p. 4) denomina *norma* de una expresión. Otras formas diversas de conseguir autorreferencia pueden verse en Smullyan (1996). Elegiré la diagonalización, seguramente la forma más conocida de conseguir autorreferencia en un lenguaje formal.

Definición 4.14 (función diagonalización). *El símbolo de función d representa una función de diagonalización —en un lenguaje interpretado extensional y con*

[18] La razón es que en un lenguaje formal se acostumbra a escribir el predicado antes que el sujeto mientras que en los lenguajes naturales suele ocurrir al revés.

capacidad de cita $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$ *— ssi para toda fórmula,* $\sigma(x)$*, con una única variable libre,* x*, se cumple*

$$\mathcal{V}_{\mathfrak{M}}(d\ulcorner\sigma(x)\urcorner) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\sigma(\ulcorner\sigma(x)\urcorner)\urcorner) \tag{4.14}$$

Nótese que, (4.14) equivale a

$$\mathcal{V}_{\mathfrak{M}}(d\ulcorner\sigma(x)\urcorner) = \sigma(\ulcorner\sigma(x)\urcorner) \tag{4.15}$$

y también a:

$$\mathfrak{M} \models d\ulcorner\sigma(x)\urcorner = \ulcorner\sigma(\ulcorner\sigma(x)\urcorner)\urcorner \tag{4.16}$$

Debido a la extensionalidad de $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$, dado cualquier término, ν, que tenga como referencia la fórmula, $\sigma(x)$, se cumplirá:

$$\mathcal{V}_{\mathfrak{M}}(d\nu) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\sigma(\ulcorner\sigma(x)\urcorner)\urcorner) \tag{4.17}$$

Además debe observarse que, cuando la referencia de un término cerrado,t, no sea una fórmula con una única variable libre, la función de valuación puede asignar adt cualquier valor del universo $\mid \mathfrak{M} \mid$ del modelo; es decir, estrictamente hablando, no hay una única función de diagonalización, sino que cualquier función que cumpla (4.14) es una función de diagonalización.[19]

Si, en el uso del término *this* cabía cuestionar si ese término tenía la referencia pretendida, resulta mucho más difícil negar que la función de diagonalización pueda aplicarse a cualquier fórmula con una única variable libre, dado que se trata de una función puramente sintáctica que, como queda de manifiesto en (4.16), transforma una fórmula en otra (resultante de sustituir la variable libre por la cita de la fórmula). Así pues, el uso de la diagonalización para conseguir autorreferencia, es en cierto modo, más impecable que el uso de demostrativos.

Proposición 4.2. *Si* d *representa una función de diagonalización en un lenguaje interpretado con capacidad de cita* $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$*, el lenguaje es fuertemente autorreferencial.*

Demostración. Dada una fórmula cualquiera con una única variable libre, $\sigma(y)$, sea τ el término $d\ulcorner\sigma(dx)\urcorner$, entonces:

[19] Algo similar puede afirmarse del predicado de verdad, T, y el esquema de Tarski.

(1) $\mathcal{V}_{\mathfrak{M}}(d^{\ulcorner}\sigma(dx)^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\sigma(d^{\ulcorner}\sigma(dx)^{\urcorner})^{\urcorner})$
; (4.14) (definición diagonalización) aplicada a la fórmula $\sigma(dx)$

(2) $\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\sigma(\tau)^{\urcorner})$; de 1, dado que τ es el término $d^{\ulcorner}\sigma(dx)^{\urcorner}$

(3) $\mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\sigma(\tau)^{\urcorner}) = \sigma(\tau)$; para toda expresión formal, ε, $\mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\varepsilon^{\urcorner}) = \varepsilon$

(4) $\mathcal{V}_{\mathfrak{M}}(\tau) = \sigma(\tau)$; de 2 y 3

luego la referencia de τ es la oración $\sigma(\tau)$.

Así pues, de acuerdo con la definición 4.13 (p. 83), se verifica que el lenguaje es fuertemente autorreferencial. ■

4.1.2.6. Imposibilidad de representar la paradoja del mentiroso

En este punto sabemos cómo conseguir cada uno de los requisitos que permitirían representar en un lenguaje interpretado de primer orden clásico con capacidad de cita, $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$, una oración del mentiroso reforzada. Pero concretemos más estos requisitos. Se trata de encontrar una oración, λ, que *diga de sí misma únicamente que no es verdadera* por lo que la oración deberá cumplir

$$\mathcal{V}_{\mathfrak{M}}(\lambda) = \mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\lambda^{\urcorner}) \tag{4.18}$$

y T debería representar un predicado de verdad.

Para conseguir (4.18) basta con que $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$ posea una función de diagonalización y un símbolo de predicado, T. En efecto, al disponer de una función de diagonalización, $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$ será fuertemente autorreferencial, y podremos aplicarle la proposición 4.1 (p. 83) por la cual para cualquier fórmula con una única variable libre, $\sigma(x)$, existe un término, τ, tal que:

$$\mathcal{V}_{\mathfrak{M}}(\sigma(\tau)) = \mathcal{V}_{\mathfrak{M}}(\sigma(^{\ulcorner}\sigma(\tau)^{\urcorner})) \tag{4.13}$$

Por tanto, para el caso particular de la fórmula $\neg Tx$, existirá un término, τ, tal que:

$$\mathcal{V}_{\mathfrak{M}}(\neg T\tau) = \mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\neg T\tau^{\urcorner}) \tag{4.19}$$

Esto satisface el esquema (4.18).

Cuál sea el término τ depende del modo en que se consiga la autorreferencia. En nuestro caso, en que usamos diagonalización, la demostración de la proposi-

ción 4.2 (p. 85), tomando la fórmula $\neg Ty$, nos lleva a encontrar que el término τ es $d^{\ulcorner}\neg Tdx^{\urcorner}$. Nuestra oración candidata a oración del mentiroso reforzada, λ, es pues:

$$\neg Td^{\ulcorner}\neg Tdx^{\urcorner} \tag{4.20}$$

Sin embargo, no es tan importante la forma concreta de la oración λ como el hecho de que cumple la condición

$$\mathcal{V}_{\mathfrak{M}}(\lambda) = \mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\lambda^{\urcorner}) \tag{4.18}$$

Para que λ sea realmente una oración del mentiroso reforzada falta un segundo requisito: que el símbolo de predicado T represente un predicado de verdad, es decir que cumpla la condición de Tarski: $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\varphi)$, para cualquier oración, φ. Si aplicamos la condición de Tarski a la oración λ, obtenemos:

$$\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\lambda) \tag{4.21}$$

Pero (4.18) y (4.21) son incompatibles pues de ambas se deduce:

$$\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\lambda^{\urcorner}) \tag{4.22}$$

y, según hemos definido nuestro lenguaje $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$, se tiene $\mathcal{V}_{\mathfrak{M}}(\psi) \neq \mathcal{V}_{\mathfrak{M}}(\neg\psi)$ cualquiera que sea la oración, ψ.

4.1.2.7. Primeras conclusiones

En definitiva, lo que hemos probado al intentar formalizar la paradoja del mentiroso es que si añadimos capacidad de cita a un lenguaje interpretado de primer orden clásico, el lenguaje resultante no puede tener un predicado de verdad y, al mismo tiempo, ser fuertemente autorreferencial.[20] Así pues, en el lenguaje formal no hay paradoja, no se puede representar una auténtica oración del mentiroso. Mas es muy importante señalar que esto no ocurre porque el lenguaje carezca de capacidad expresiva: un lenguaje obtenido al añadir capacidad de cita a un lenguaje interpretado de primer orden clásico, puede ser fuertemente autorreferencial

[20] Esta conclusión puede verse como una versión, para nuestro tipo de lenguajes, del teorema de Tarski sobre la indefinibilidad de la verdad en el lenguaje formal de la aritmética.

y puede contener un predicado de verdad, aunque no ambas cosas a la vez. Esto es prácticamente lo mismo que, respecto al lenguaje natural, dice Martin sobre la incompatibilidad de (S) y (T), así que el lenguaje formal parece reflejar suficientemente aquellos aspectos del lenguaje natural relevantes para la aparición de la paradoja del mentiroso. Sin embargo, mientras que en el lenguaje formal no hay paradoja sino la constatación de que un predicado de verdad es incompatible con la autorreferencialidad fuerte; en el lenguaje natural hay paradoja porque parece incuestionable que posee el predicado de verdad y, a la vez, oraciones que dicen de sí mismas que no son verdaderas.

Surge entonces la pregunta: ¿cómo es que el lenguaje natural posee unas características que conducen a la paradoja del mentiroso y, por tanto, a contradicción? La situación no debe extrañarnos si tenemos en cuenta que, como es sabido, y destacado especialmente por los defensores de lógicas paraconsistentes, las creencias de las personas, las leyes civiles e incluso las teorías científicas, son a menudo inconsistentes sin que por eso estemos dispuestos a creer cualquier cosa o a renunciar fácilmente a esas creencias o a esas teorías científicas, o que las leyes y normativas de una sociedad dejen de cumplir una misión. Análogamente, los lenguajes naturales pueden conllevar paradojas sin que ello impida su utilidad en la comunicación humana.

Ahora bien, que en la práctica aceptemos sistemas de creencias, normativas, etc. inconsistentes no significa que tengamos que aceptar contradicciones evidentes; es más, a mi juicio, no podemos aceptarlas. Cuando descubrimos la inconsistencia de un sistema, sabemos que no lo podemos aceptar en su totalidad. Lo podemos seguir aceptando en su generalidad por diversos motivos: que no encontremos manera de mejorarlo, que las situaciones en que se pone de manifiesto la inconsistencia se den rara vez, o que carezcan de importancia práctica, etc. Pero no podemos negar que, al menos desde un punto de vista teórico, hay un problema que resolver.

En el caso que nos ocupa, el análisis que hemos realizado nos indica que bajo los supuestos semánticos de un lenguaje interpretado clásico al que se añade capacidad de cita, la existencia de un predicado de verdad[21] es incompatible con que el lenguaje sea fuertemente autorreferencial, ya que podríamos crear una oración del mentiroso y deducir una contradicción. Así pues o cuestionamos que algunos de los supuestos semánticos de nuestro tipo de lenguajes interpretados sean fiel reflejo de

[21] Entendido como un predicado que satisface el esquema de Tarski.

una parte de la semántica del lenguaje natural, o cuestionamos que en el lenguaje natural sean ciertas todas las equivalencias con la forma del esquema de Tarski al tiempo que posee capacidad autorreferencial suficiente para enunciar una oración que diga de sí misma únicamente que no es verdadera. Ambos cuestionamientos se han realizado. Los supuestos semánticos de nuestro tipo de lenguajes interpretados no son aceptados por las lógicas inconsistentes que admiten oraciones verdaderas y falsas a la vez, ni por las propuestas de Kripke o de Martin para solventar la paradoja del mentiroso, que admiten oraciones sin valor de verdad o con un tercer valor de verdad, ni por la propuesta de Skyrms modificando el clásico principio de sustitución de los idénticos, etc. Las propuestas de Russell y Tarski suponen a la vez una jerarquización del concepto de verdad y una limitación de la capacidad autorreferencial en lenguajes formales, aunque resultan demasiado artificiosas para explicar el problema en el lenguaje natural. Las de McGee, o de Gupta y Belnap, son ejemplos de propuestas que modifican el concepto clásico de verdad.

Como ya hemos visto, hay motivos suficientes para considerar todas estas propuestas de solución de la paradoja del mentiroso como insatisfactorias. Esto nos lleva a pensar que no se soluciona la paradoja del mentiroso simplemente evitando la inconsistencia en un sistema formal a base de buscar esta o aquella ingeniosa variante o restricción del esquema de Tarski o artificiosos nuevos principios lógicos. Porque, evitar inconsistencias en el sistema formal solo sirve si se hace aplicando unos principios semánticos que sigan siendo correctos en, al menos, la parte del lenguaje natural donde puede expresarse la paradoja del mentiroso.

4.1.3. La paradoja del mentiroso en una lógica trivaluada

Con el objetivo de distinguir los elementos fundamentales a los que se debe la paradoja de aquellos puramente circunstanciales, podemos continuar nuestro estudio probando algunas alternativas al tipo de lenguajes interpretados que hemos usado hasta aquí.

Quizá la intuición que más consenso suscita en cuanto al estatus veritativo de la oración del mentiroso es que no es verdadera ni falsa. Por tanto, una alternativa, que no podemos pasar por alto, es usar lenguajes interpretados que admitan oraciones no verdaderas ni falsas. Aunque se trata de una alternativa bastante estudiada,[22] la perspectiva que aquí adoptaremos será diferente a la usual.

[22] Véanse por ejemplo Kripke (1975), Feferman (1984) o Barwise y Moss (1996, cap. 13).

El planteamiento típico lo expresa muy bien Feferman (1984)[23] para quien el primer objetivo es producir un sistema consistente que tenga "más o menos" las propiedades siguientes: 1/ (i) para cada oración, ϕ, hay un término cerrado que la nombra, $\ulcorner\phi\urcorner$; (ii) para cada fórmula, $\psi(x)$, hay una oración, ϕ, que equivale a $\psi(\ulcorner\phi\urcorner)$; 2/ se aceptan los axiomas y reglas del cálculo proposicional ordinario; 3/ el predicado $T(x)$ satisface el esquema de Tarski. Puesto que, como el propio Feferman demuestra,[24] estas propiedades conducen a contradicción, es necesario modificar al menos una de ellas para que el sistema resultante sea consistente. Así planteado, este primer objetivo es puramente técnico y puede conseguirse de formas diversas (por ejemplo, las de Kripke (1975), Feferman (1984) e incluso Tarski (1983a)). Después es preciso algún criterio para dar o no por válidas cada una de las soluciones al primer objetivo. El criterio más exigente, señala Feferman (1984),[25] sería encajar la solución en una semántica global para el lenguaje natural.

Hasta ahora, el resultado de este planteamiento es que se consiguen sistemas formales consistentes pero, las soluciones formales no son trasladables al lenguaje natural, donde persiste la paradoja. En mi opinión, se trata de un resultado previsible, porque este planteamiento fomenta el separar la solución técnica (objetivo inicial) de la filosófica (validación de las soluciones técnicas).

Nuestro planteamiento, por el contrario, no buscará forzar la consistencia del sistema formal sino, como hemos hecho hasta ahora, pero basándonos en lenguajes interpretados que admitan oraciones no verdaderas ni falsas, reflejar en el lenguaje formal las propiedades del lenguaje natural que conducen a la paradoja del mentiroso. Pretendemos así clarificar el problema de la paradoja, aislar los elementos verdaderamente responsables de la misma.

4.1.3.1. Lenguaje interpretado trivaluado

Como se ha señalado, necesitamos admitir en el lenguaje formal oraciones no verdaderas ni falsas, oraciones que, siguiendo a Kripke (1975), denominaremos*indefinidas*. Destacaremos dos motivos para considerar que una oración de la forma "a es P" es indefinida. Uno es que el término "a" carezca de referencia. En relación al estudio de la paradoja del mentiroso, este es, por ejemplo, un planteamiento que utiliza van Fraassen (1970), siguiendo la idea de presuposición fregeana.

[23] Martin (1984, p. 245).
[24] Ibíd. p. 243.
[25] Ibíd. p. 245.

Otro es el planteamiento de Martin (1970a) basado en los rangos de aplicabilidad de los predicados: aunque "a" tenga referencia, si esta no pertenece al rango de aplicabilidad del predicado P, "a es P" no será verdadera ni falsa.

Podemos abarcar ambos puntos de vista, permitiendo que haya términos sin referencia y estableciendo rangos de aplicabilidad para los predicados y, en general, para las relaciones. Esto último puede hacerse, en un lenguaje interpretado $(\mathcal{L}, \xi, \mathfrak{M})$, de dos formas:

1. Caracterizando una relación n-ádica mediante una función parcial definida entre $\mid \mathfrak{M} \mid^n$ y $\{\mathfrak{f}, \mathfrak{v}\}$. El conjunto de elementos de $\mid \mathfrak{M} \mid^n$ a los que esta función asocia el valor $\mathfrak{v}$ es la extensión de la relación; el conjunto de elementos de $\mid \mathfrak{M} \mid^n$ a los que la función asocia el valor $\mathfrak{f}$ es la antiextensión de la relación; el conjunto unión de la extensión y la antiextensión es el rango de aplicabilidad de la relación. A aquellos elementos de $\mid \mathfrak{M} \mid^n$ que no estén en el rango de aplicabilidad de la relación no se les asocia ningún valor de verdad.

2. Añadiendo un nuevo "valor de verdad", $\mathfrak{i}$, con la idea de asociarlo a aquellos elementos de $\mid \mathfrak{M} \mid^n$ que no están en el rango de aplicabilidad de la relación. Entonces tendríamos un lenguaje interpretado trivaluado y caracterizaríamos una relación n-ádica mediante una función total entre $\mid \mathfrak{M} \mid^n$ y $\{\mathfrak{f}, \mathfrak{i}, \mathfrak{v}\}$. La extensión, la antiextensión y el rango de aplicabilidad de la relación las definiríamos de igual modo.

Ambos planteamientos son similares, dado que las oraciones atómicas que no tienen valor de verdad en el caso bivaluado, toman el valor $\mathfrak{i}$ en el caso trivaluado y viceversa. Sin embargo, una aplicación estricta del principio de composicionalidad puede marcar una diferencia en la valuación de las oraciones compuestas. Porque, de acuerdo con Janssen (1997, p. 427, punto 5) uno de los supuestos de la composicionalidad es que todas las partes propias de una expresión tienen interpretación, por lo que si alguna de sus partes no la tiene tampoco la tendrá la expresión completa. Por tanto, en el caso de que φ sea una oración verdadera y ψ una oración sin valor de verdad, la oración $(\varphi \vee \psi)$ no tendría valor de verdad, a pesar de que es razonable asignarle el valor verdadero. La introducción de un tercer valor de verdad, permite solucionar el problema manteniendo la composicionalidad, porque, siguiendo con el ejemplo, φ pasaría a tener ese tercer valor de

verdad y el valor de verdad de $(\varphi \vee \psi)$ sería una función de los valores de verdad de sus componentes. Además, es fácil diseñar un lenguaje interpretado trivaluado en que las oraciones sin valor de verdad en el lenguaje interpretado bivaluado, y solo ellas, tengan el valor de verdad $\mathfrak{i}$: bastaría con que las operaciones asociadas a las conectivas y cuantificadores asocien el valor $\mathfrak{i}$ cuando alguno de sus operandos sea $\mathfrak{i}$ (en el caso de un cuantificador, cuando alguno de los elementos del conjunto de valores de verdad al que se aplica la operación asociada al cuantificador, sea $\mathfrak{i}$). En este sentido, el lenguaje bivaluado corresponde a un caso particular del trivaluado.

Por estos motivos, y porque, más adelante, utilizaremos lenguajes interpretados donde se contempla la posibilidad alternativa (solo dos valores de verdad y posibilidad de oraciones sin valor de verdad por la causa explicada), elegiré la opción de una semántica trivaluada.

Sea $\mathcal{L}$ nuestro lenguaje de primer orden ampliado con capacidad de cita. Estableceremos en él una lógica trivaluada mediante la siguiente matriz:

Definición 4.15 (matriz para la lógica trivaluada). *La matriz, ξ, que determina una lógica trivaluada en el lenguaje $\mathcal{L}$ está formada por:*

1. *Un conjunto de tres valores de verdad que designaremos $\mathfrak{f}$ (falso), $\mathfrak{i}$ (impropio), $\mathfrak{v}$ (verdadero): $W = \{\mathfrak{f}, \mathfrak{i}, \mathfrak{v}\}$.*

2. *Una interpretación, C, de las constantes lógicas (conectivas y cuantificadores), es decir: a) por cada conectiva n-ádica, k, una función total n-ádica, C_k, definida entre W^n y W, de modo compatible con la interpretación clásica; b) por cada cuantificador, una función total definida entre $(\mathscr{P}(W) - \{\varnothing\})$ y W de modo compatible con la interpretación clásica.*[26]

Ahora solo queda añadir a $(\mathcal{L}, \xi)$ el concepto de modelo parcial para los términos $\mathfrak{M}$. El resultado será un lenguaje interpretado trivaluado $(\mathcal{L}, \xi, \mathfrak{M})$. Para recalcar el tipo de lenguaje, matriz y modelo que lo forman lo designaremos mediante $(\mathcal{L}, \xi, \mathfrak{M})_{tri}$.

Definición 4.16 (modelo parcial (para los términos)). *Un modelo parcial para los términos, $\mathfrak{M}$, para $(\mathcal{L}, \xi)$ está formado por:*

[26] Es obvio que las constantes lógicas admiten diversas interpretaciones compatibles con la clásica lo cual daría lugar a distintas lógicas trivaluadas.

1. *Un conjunto no vacío* $|\,\mathfrak{M}\,|$ *llamado dominio o universo de* $\mathfrak{M}$. *Puesto que* $\mathcal{L}$ *tiene capacidad de cita, el conjunto de términos y fórmulas de* $\mathcal{L}$ *debe ser un subconjunto de* $|\,\mathfrak{M}\,|$.

2. *Una función de interpretación,* $\mathfrak{I}_{\mathfrak{M}}$, *que asocia:*

   - *A cada símbolo de oración,* p, *un único valor de verdad,* $\mathfrak{I}_{\mathfrak{M}}(p)$ *(elemento de* W*).*

   - *A cada símbolo de constante,* c, *para el que* $\mathfrak{I}_{\mathfrak{M}}$ *esté definido, un elemento de* $|\,\mathfrak{M}\,|$, $\mathfrak{I}_{\mathfrak{M}}(c)$.

   - *A cada término de cita, la expresión citada:* $\mathfrak{I}_{\mathfrak{M}}(\ulcorner\varepsilon\urcorner) = \varepsilon$.

   - *A cada símbolo de relación n-ádica,* R, *una función total,* $\mathfrak{I}_{\mathfrak{M}}(R)$, *definida entre* $|\,\mathfrak{M}\,|^n$ *y* W.

   - *A cada símbolo de función n-ádica,* f, *una función parcial,* $\mathfrak{I}_{\mathfrak{M}}(f)$, *definida entre* $|\,\mathfrak{M}\,|^n$ *y* $|\,\mathfrak{M}\,|$.

El concepto de asignación de variables es el mismo, salvo que, ahora, la función entre el conjunto de variables y el universo del modelo es una función parcial.

Dado un lenguaje interpretado, $(\mathcal{L}, \xi, \mathfrak{M})_{tri}$, y una asignación de variables, s, para el mismo, la función de valuación con respecto a la asignación de variables$s$, es la función $\mathcal{V}_{\mathfrak{M},s}$, que está parcialmente definida, entre el conjunto de términos y $|\,\mathfrak{M}\,|$, y totalmente definida entre el conjunto de fórmulas yW, de acuerdo con la siguiente definición.

Definición 4.17. *En un lenguaje* $(\mathcal{L}, \xi, \mathfrak{M})_{tri}$, *sean:* s, *una asignación de variables;* ε, *una expresión;* c, *un símbolo de constante;* x, *una variable;* f, *un símbolo de función n-ádica;* $t_1, t_2, ..., t_n$, *términos;* p, *un símbolo de oración;* R, *un símbolo de relación n-ádica;* φ *y* ψ *fórmulas. La función de valuación,* $\mathcal{V}_{\mathfrak{M},s}$, *viene definida por:*

1. $\mathcal{V}_{\mathfrak{M},s}(\ulcorner\varepsilon\urcorner) = \varepsilon$.
2. $\mathcal{V}_{\mathfrak{M},s}(c) =_{def} \mathfrak{I}_{\mathfrak{M}}(c)$.[27]
3. $\mathcal{V}_{\mathfrak{M},s}(x) =_{def} s(x)$.

[27] Si $\mathfrak{I}_{\mathfrak{M}}$ no está definida sobre c, es decir, si c no pertenece al dominio de la función $\mathfrak{I}_{\mathfrak{M}}$ —abreviadamente, $c \notin dom(\mathfrak{I}_{\mathfrak{M}})$—, entonces, $c \notin dom(\mathcal{V}_{\mathfrak{M},s})$. En general, cuando escribamos algo de la forma $A =_{def} B$, se entenderá que si B no está definido, A tampoco lo estará.

4. $\mathcal{V}_{\mathfrak{M},s}(f\,t_1,t_2,...,t_n) =_{def} \mathfrak{I}_{\mathfrak{M}}(f)(\mathcal{V}_{\mathfrak{M},s}(t_1),\mathcal{V}_{\mathfrak{M},s}(t_2),...,\mathcal{V}_{\mathfrak{M},s}(t_n))$.[28]

5. $\mathcal{V}_{\mathfrak{M},s}(p) =_{def} \mathfrak{I}_{\mathfrak{M}}(p)$.

6. $\mathcal{V}_{\mathfrak{M},s}(t_1 = t_2) =_{def} \begin{cases} \mathfrak{i} & si \quad t_1 \notin dom(\mathcal{V}_{\mathfrak{M},s}) \quad o \quad t_2 \notin dom(\mathcal{V}_{\mathfrak{M},s}), \\ & en\,otro\,caso: \\ \mathfrak{v} & si \quad \mathcal{V}_{\mathfrak{M},s}(t_1) = \mathcal{V}_{\mathfrak{M},s}(t_2) \\ \mathfrak{f} & si \quad \mathcal{V}_{\mathfrak{M},s}(t_1) \neq \mathcal{V}_{\mathfrak{M},s}(t_2). \end{cases}$

7. $\mathcal{V}_{\mathfrak{M},s}(R\,t_1,t_2,...,t_n) =_{def}$

$=_{def} \begin{cases} \mathfrak{i} & si\ existe\ t_i\ (1 \leq i \leq n)\ tal\ que\ t_i \notin dom(\mathcal{V}_{\mathfrak{M},s}) \\ \mathfrak{I}_{\mathfrak{M}}(R)(\mathcal{V}_{\mathfrak{M},s}(t_1),\mathcal{V}_{\mathfrak{M},s}(t_2),...,\mathcal{V}_{\mathfrak{M},s}(t_n)) & en\,otro\,caso. \end{cases}$

8. $\mathcal{V}_{\mathfrak{M},s}(\neg\varphi) =_{def} C_{\neg}(\mathcal{V}_{\mathfrak{M},s}(\varphi))$.

9. $\mathcal{V}_{\mathfrak{M},s}((\varphi \wedge \psi)) =_{def} C_{\wedge}(\mathcal{V}_{\mathfrak{M},s}(\varphi),\mathcal{V}_{\mathfrak{M},s}(\psi))$.

10. $\mathcal{V}_{\mathfrak{M},s}((\varphi \vee \psi)) =_{def} C_{\vee}(\mathcal{V}_{\mathfrak{M},s}(\varphi),\mathcal{V}_{\mathfrak{M},s}(\psi))$.

11. $\mathcal{V}_{\mathfrak{M},s}((\varphi \rightarrow \psi)) =_{def} C_{\rightarrow}(\mathcal{V}_{\mathfrak{M},s}(\varphi),\mathcal{V}_{\mathfrak{M},s}(\psi))$.

12. $\mathcal{V}_{\mathfrak{M},s}((\varphi \leftrightarrow \psi)) =_{def} C_{\leftrightarrow}(\mathcal{V}_{\mathfrak{M},s}(\varphi),\mathcal{V}_{\mathfrak{M},s}(\psi))$.

13. $\mathcal{V}_{\mathfrak{M},s}(\forall x\,\varphi) =_{def} C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\varphi)/e \in| \mathfrak{M} |\})$.

14. $\mathcal{V}_{\mathfrak{M},s}(\exists x\,\varphi) =_{def} C_{\exists}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\varphi)/e \in| \mathfrak{M} |\})$.

Supuesta nuestra lógica trivaluada para $(\mathcal{L},\xi)$, diremos que φ es verdadera en $\mathfrak{M}$ con respecto a la asignación s, o que $\mathfrak{M}$ satisface φ con respecto a la asignación s, cuando $\mathcal{V}_{\mathfrak{M},s}(\varphi) = \mathfrak{v}$ (y solo en este caso). Entonces escribiremos: $\mathfrak{M}, s \models \varphi$.

El valor que la función $\mathcal{V}_{\mathfrak{M},s}$ asocia a algunas fórmulas y términos es independiente de la asignación de variables s que se tome. Esto nos permite definir la función de valuación, $\mathcal{V}_{\mathfrak{M}}$, de esos términos y fórmulas de igual modo que hicimos en un lenguaje interpretado clásico.

Así mismo, diremos que la fórmula φ es verdadera en $\mathfrak{M}$, o que $\mathfrak{M}$ satisface φ, únicamente cuando $\mathcal{V}_{\mathfrak{M}}(\varphi) = \mathfrak{v}$. Entonces escribiremos: $\mathfrak{M} \models \varphi$.

Es interesante señalar que la función $\mathcal{V}_{\mathfrak{M}}$ está definida para toda oración pero, en general, no para todo término cerrado.

Finalmente, es fácil constatar que este tipo de lenguajes interpretados es una generalización del anterior. Además conservan las propiedades de referencialidad, extensionalidad y composicionalidad.

[28] Téngase en cuenta que, para que $\mathcal{V}_{\mathfrak{M},s}$ esté definida sobre $(f\,t_1,t_2,...,t_n)$, es necesario y suficiente que esté definida para cada uno de los términos $t_1,t_2,...,t_n$ y que $(\mathcal{V}_{\mathfrak{M},s}(t_1),\mathcal{V}_{\mathfrak{M},s}(t_2),...,\mathcal{V}_{\mathfrak{M},s}(t_n)) \in dom(\mathfrak{I}_{\mathfrak{M}}(f))$.

4.1.3.2. Predicado de verdad, autorreferencia y paradoja del mentiroso.

¿Que ocurrirá ahora si intentamos representar la oración del mentiroso reforzada en un lenguaje interpretado $(\mathcal{L}, \xi, \mathfrak{M})_{tri}$ como el que acabamos de especificar? Cambiando $(\mathcal{L}, \xi, \mathfrak{M})_{cit}$ por $(\mathcal{L}, \xi, \mathfrak{M})_{tri}$, podemos mantener la misma definición de función de diagonalización y seguirán siendo verdaderas las proposiciones 4.1 (p. 83) y 4.2 (p. 85). Por la proposición 4.2, si $(\mathcal{L}, \xi, \mathfrak{M})_{tri}$ posee una función de diagonalización, es fuertemente autorreferencial y, si esto ocurre, la proposición 4.1 nos permite encontrar una oración λ que cumple

$$\mathcal{V}_{\mathfrak{M}}(\lambda) = \mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\lambda^{\urcorner}) \tag{4.18}$$

Si mantenemos como definición de predicado de verdad, T,[29] que para toda oración, φ, se verifique $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\varphi)$, podremos deducir, como antes:

$$\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\lambda) \tag{4.21}$$

Y por último, a partir (4.18) y (4.21) volvemos a obtener la conclusión:

$$\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\lambda^{\urcorner}) \tag{4.22}$$

Sin embargo, esta conclusión no es ahora necesariamente contradictoria. Todo depende de la interpretación de la conectiva $\neg$ (negación). Es razonable pensar que si una oración no es verdadera ni falsa su negación tampoco lo es. Esto significaría que la función, $C_{\neg}$, que la matriz asigna a la negación, cumpliría $C_{\neg}(\mathfrak{i}) = \mathfrak{i}$. Por consiguiente, la condición (4.22) sería verdadera si $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}) = \mathfrak{i}$. Como $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\lambda)$, tendríamos que $\mathcal{V}_{\mathfrak{M}}(\lambda) = \mathfrak{i}$, lo cual refleja la intuición de que la oración del mentiroso no es verdadera ni falsa. Puesto que $^{\ulcorner}\lambda^{\urcorner}$ sí tiene referencia en el lenguaje interpretado ($\mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\lambda^{\urcorner}) = \lambda$), la explicación que podemos dar, para justificar que $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}) = \mathfrak{i}$, consiste en afirmar que $^{\ulcorner}\lambda^{\urcorner}$ no está en el rango de aplicabilidad de T. Precisamente esta es una conclusión fundamental en las propuestas con las que Martin (1970a) y Kripke (1975) pretenden resolver la paradoja del mentiroso.

[29] Hablando con propiedad T solo es un símbolo de predicado. Pero, mientras no se genere confusión, usaremos T para referirnos tanto al símbolo de predicado como al predicado. Esto mismo podremos hacer con otros símbolos de predicado o de función.

¿Por qué esta explicación no es satisfactoria? Porque a nuestro lenguaje interpretado no le hemos dotado de capacidad para decir con verdad que una oración que no es verdadera ni falsa, no es verdadera. Si dotamos al lenguaje de esa capacidad, la oración del mentiroso nos llevará de nuevo a contradicción. Para que en nuestro lenguaje pueda decirse con verdad que una oración indefinida no es verdadera basta con definir el predicado de verdad T de modo que para toda oración, φ, se verifique:

$$\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = \begin{cases} \mathcal{V}_{\mathfrak{M}}(\varphi) & si\, \mathcal{V}_{\mathfrak{M}}(\varphi) \neq \mathfrak{i} \\ \mathfrak{f} & si\, \mathcal{V}_{\mathfrak{M}}(\varphi) = \mathfrak{i} \end{cases} \tag{4.23}$$

Esta caracterización del predicado de verdad es, en mi opinión, más correcta para un lenguaje trivaluado como el nuestro, que la anterior ($\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\varphi)$). Porque cuando φ es una oración no verdadera ni falsa, afirmar "la oración φ es verdadera" es una falsedad. Si pretendemos que el reflejo en el lenguaje formal de "la oración φ es verdadera" sea $T^{\ulcorner}\varphi^{\urcorner}$, entonces, es claro que su valor de verdad, $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner})$ tiene que ser $\mathfrak{f}$. Para expresar (4.23) de un modo más compacto, introducimos una función total $C_{\triangle}$ definida entre W y W mediante:

$$C_{\triangle}(\mathfrak{f}) = C_{\triangle}(\mathfrak{i}) = \mathfrak{f}; \quad C_{\triangle}(\mathfrak{v}) = \mathfrak{v} \tag{4.24}$$

y podremos escribir, en lugar de (4.23),

$$\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = C_{\triangle}(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.25}$$

La siguiente proposición nos muestra que, un lenguaje en el que hay un predicado, T, que para toda oración, φ, satisface $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = C_{\triangle}(\mathcal{V}_{\mathfrak{M}}(\varphi))$, no podrá ser fuertemente autorreferencial —y, por supuesto, no importa que llamemos o no predicado de verdad a T, aunque nosotros sí lo haremos—.

Proposición 4.3. *Sea $(\mathcal{L}, \xi, \mathfrak{M})_{tri}$ un lenguaje interpretado trivaluado con capacidad de cita. $(\mathcal{L}, \xi, \mathfrak{M})_{tri}$ no puede ser fuertemente autorreferencial y a la vez tener un predicado de verdad, T, entendido éste como un predicado que, para toda oración φ, satisfaga $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = C_{\triangle}(\mathcal{V}_{\mathfrak{M}}(\varphi))$.*

Demostración. Si fuese fuertemente autorreferencial y tuviese un predicado de verdad T, habría una oración, λ, que cumpliría:

(1) $\mathcal{V}_{\mathfrak{M}}(\lambda) = \mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\lambda^{\urcorner})$
; proposición 4.1 (p. 83) aplicada a la fórmula $\neg Tx$
(2) $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}) = C_{\triangle}(\mathcal{V}_{\mathfrak{M}}(\lambda))$
; por ser T un predicado de verdad
(3) $\mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\lambda^{\urcorner}) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}))$
; composicionalidad ($C_{\neg}$ interpreta la negación)
(4) $\mathcal{V}_{\mathfrak{M}}(\lambda) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\lambda^{\urcorner}))$; de 1 y 3
(5) $\mathcal{V}_{\mathfrak{M}}(\lambda) = C_{\neg}(C_{\triangle}(\mathcal{V}_{\mathfrak{M}}(\lambda)))$; de 2 y 4

Pero (5) es imposible puesto que si $\mathcal{V}_{\mathfrak{M}}(\lambda) = \mathfrak{v}$, entonces $C_{\neg}(C_{\triangle}(\mathcal{V}_{\mathfrak{M}}(\lambda))) = \mathfrak{f}$ y si $\mathcal{V}_{\mathfrak{M}}(\lambda) \neq \mathfrak{v}$, entonces $C_{\neg}(C_{\triangle}(\mathcal{V}_{\mathfrak{M}}(\lambda))) = \mathfrak{v}$ —basta tener en cuenta que $C_{\triangle}(\mathfrak{f}) = \mathfrak{f}$, $C_{\triangle}(\mathfrak{i}) = \mathfrak{f}$, $C_{\triangle}(\mathfrak{v}) = \mathfrak{v}$ y que $C_{\neg}(\mathfrak{f}) = \mathfrak{v}$, $C_{\neg}(\mathfrak{v}) = \mathfrak{f}$—.[30] ■

4.1.3.3. Nuevas conclusiones

Podemos resumir los principales resultados de nuestro estudio de la paradoja en nuestros lenguajes trivaluados, en los siguientes puntos:

- La existencia de una función de diagonalización es suficiente para que el lenguaje sea fuertemente autorreferencial.
- Si dotamos al lenguaje de suficiente capacidad expresiva para poder decir en él, de una oración *indefinida*, que no es verdadera, el lenguaje no puede ser fuertemente autorreferencial.

Por consiguiente, el haber refinado nuestra semántica mediante el uso de lenguajes interpretados en los que algunas oraciones pueden no ser verdaderas ni falsas, no ha solucionado el problema de la paradoja del mentiroso reforzada. Es más tenemos la sensación de que no hemos avanzado nada, pues volvemos a llegar a la conclusión de que la autorreferencialidad fuerte es incompatible con la existencia de un predicado de verdad.

Mas, como dijimos, nuestro objetivo no era encontrar un sistema formal consistente donde pueda representarse la oración del mentiroso, sino clarificar el problema de la paradoja. Y, en este sentido sí podemos hacer avances observando qué es lo que tienen en común los razonamientos que hemos hecho en los dos tipos de

[30] Obsérvese que el valor que demos a $C_{\neg}(\mathfrak{i})$ no es relevante.

lenguajes interpretados estudiados hasta ahora y que nos llevan a contradicción si intentamos representar la oración del mentiroso reforzada.

En ambos casos tenemos, por un lado, que, mediante diagonalización u otro método de conseguir autorreferencia, podemos encontrar, para cualquier fórmula con una única variable libre, $\sigma(x)$, un término, τ, y una oración, ψ (que es simplemente $\sigma(\tau)$), tales que:

$$\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\sigma(\ulcorner\psi\urcorner)) \tag{4.26}$$

Por otra parte, tenemos un predicado, T, que podemos caracterizar como un predicado que, para toda oración, φ, verifica

$$\mathcal{V}_{\mathfrak{M}}(T\ulcorner\varphi\urcorner) = C_T(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.27}$$

Aquí C_T es una función total definida de W en W. En el caso bivaluado, era la función identidad y en el trivaluado, la función $C_\triangle$.

De (4.26) y (4.27) se puede obtener, respectivamente, $\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\neg T\ulcorner\psi\urcorner)$ y $\mathcal{V}_{\mathfrak{M}}(T\ulcorner\psi\urcorner) = C_T(\mathcal{V}_{\mathfrak{M}}(\psi))$. Pero como $\mathcal{V}_{\mathfrak{M}}(\neg T(\ulcorner\psi\urcorner)) = C_\neg(\mathcal{V}_{\mathfrak{M}}(T\ulcorner\psi\urcorner))$, podemos concluir que

$$\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\neg T\ulcorner\psi\urcorner) = C_\neg(\mathcal{V}_{\mathfrak{M}}(T\ulcorner\psi\urcorner)) = C_\neg(C_T(\mathcal{V}_{\mathfrak{M}}(\psi))) \tag{4.28}$$

y, en definitiva:

$$\mathcal{V}_{\mathfrak{M}}(\psi) = C_\neg(C_T(\mathcal{V}_{\mathfrak{M}}(\psi))) \tag{4.29}$$

Usando la notación habitual para funciones compuestas,[31] la condición (4.29) no será posible a menos que exista un $a \in W$ que cumpla $a = (C_\neg \circ C_T)(a)$, es decir, a menos que la función $(C_\neg \circ C_T)$ tenga un punto fijo.[32] Así, si como sucede en los casos estudiados, la función $(C_\neg \circ C_T)$ no tiene ningún punto fijo, la condición (4.29) no puede cumplirse.

El aspecto clave de (4.29) es que relaciona $\mathcal{V}_{\mathfrak{M}}(\psi)$ con una función de sí mismo. Pero la función será distinta a $(C_\neg \circ C_T)$ si en vez de usar la oración del mentiroso reforzada usamos otra oración autorreferencial. El razonamiento anterior que nos

[31] Si f es una función definida entre A y B y g es una función definida entre B y C, se llama función compuesta de f y g a la función $(g \circ f)$ definida entre A y C de manera que $(g \circ f)(x) =_{def} g(f(x))$. Para una definición más rigurosa véase el apéndice A.

[32] Se dice que una función, h, tiene un punto fijo, a, cuando se cumple $h(a) = a$.

ha llevado a (4.29) es un caso particular del que se indica después. Pero, antes de adentrarnos en él, introduciré unas notaciones que nos serán útiles en numerosas partes de este trabajo.

Para señalar que una fórmula ρ tiene una única variable libre, x, que en todas sus apariciones va precedida del símbolo de predicado P, podremos designar dicha fórmula mediante $\rho(Px)$ o decir que ρ es de la forma

$$\rho(Px) \tag{4.30}$$

A la oración resultante de sustituir x, en todas sus apariciones libres dentro de $\rho(Px)$, por un término cerrado, τ, la podemos designar mediante

$$\rho(P\tau) \tag{4.31}$$

La composicionalidad nos permite decir que existe una función, C_ρ, que obtiene $\mathcal{V}_{\mathfrak{M}}(\rho(P\tau))$ a partir de $\mathcal{V}_{\mathfrak{M}}(P\tau)$ —cualesquiera que sean el símbolo de predicado P y el término cerrado τ—:

$$\mathcal{V}_{\mathfrak{M}}(\rho(P\tau)) = C_\rho(\mathcal{V}_{\mathfrak{M}}(P\tau)) \tag{4.32}$$

Por ejemplo, sea ρ la fórmula $(Px \wedge q)$, donde P es un símbolo de predicado y q, un símbolo de oración. Podemos decir que ρ es de la forma $\rho(Px)$. Si τ es un término cerrado, $\rho(P\tau)$ es la oración $(P\tau \wedge q)$. Por definición de $\mathcal{V}_{\mathfrak{M}}$, tenemos

$$\mathcal{V}_{\mathfrak{M}}(P\tau \wedge q) = C_\wedge(\mathcal{V}_{\mathfrak{M}}(P\tau), \mathcal{V}_{\mathfrak{M}}(q)) \tag{4.33}$$

Por tanto, en este caso particular,

$$\mathcal{V}_{\mathfrak{M}}(\rho(P\tau)) = C_\wedge(\mathcal{V}_{\mathfrak{M}}(P\tau), \mathcal{V}_{\mathfrak{M}}(q)) \tag{4.34}$$

y, teniendo en cuenta (4.32), encontramos que C_ρ es una función que cumple:

$$C_\rho(\mathcal{V}_{\mathfrak{M}}(P\tau)) = C_\wedge(\mathcal{V}_{\mathfrak{M}}(P\tau), \mathcal{V}_{\mathfrak{M}}(q)) \tag{4.35}$$

Si llamamos z al argumento de C_ρ, es decir a $\mathcal{V}_{\mathfrak{M}}(P\tau)$, podemos escribir:

$$C_\rho(z) = C_\wedge(z, \mathcal{V}_{\mathfrak{M}}(q)) \tag{4.36}$$

Como $\mathcal{V}_{\mathfrak{M}}(P\tau)$ puede ser cualquier elemento del conjunto de valores de verdad, W, podemos decir que C_ρ es la función, que asocia a un valor de verdad z el valor de verdad $C_\wedge(z, \mathcal{V}_{\mathfrak{M}}(q))$.

Hagamos ahora uso de estas notaciones y consideraciones. Tomemos una fórmula no cuantificada de la forma $\rho(Tx)$. Si el lenguaje es fuertemente autorreferencial, será aplicable la proposición 4.1 (p. 83) a la fórmula $\rho(Tx)$, por lo que habrá una oración, ψ, tal que

$$\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\rho(T^{\ulcorner}\psi^{\urcorner})) \tag{4.37}$$

El predicado, T, se caracteriza como un predicado que, para toda oración, φ, verifica

$$\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = C_T(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.27}$$

Por la composicionalidad de $\mathcal{V}_{\mathfrak{M}}$, existirá una función C_ρ definida entre W y W, tal que $\mathcal{V}_{\mathfrak{M}}(\rho(T^{\ulcorner}\psi^{\urcorner})) = C_\rho(\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\psi^{\urcorner}))$. Como, además, $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\psi^{\urcorner}) = C_T(\mathcal{V}_{\mathfrak{M}}(\psi))$, tenemos:

$$\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\rho(T^{\ulcorner}\psi^{\urcorner})) = C_\rho(\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\psi^{\urcorner})) = C_\rho(C_T(\mathcal{V}_{\mathfrak{M}}(\psi))) \tag{4.38}$$

y, en definitiva:

$$\mathcal{V}_{\mathfrak{M}}(\psi) = C_\rho(C_T(\mathcal{V}_{\mathfrak{M}}(\psi))) \tag{4.39}$$

de la que (4.29) no es más que un caso particular.

Este resultado nos indica que aunque definiésemos T usando la condición $\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\varphi^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\varphi)$ —es decir, tomando para C_T la función identidad— y definiésemos $C_\neg(\mathfrak{i}) = \mathfrak{i}$, no estaríamos libres de contradicción. Bastaría introducir una conectiva, $\sim$, con $C_\sim$ definida, entre W y W, de modo que no tenga punto fijo.[33] En aplicación de (4.39), llegaríamos a la conclusión contradictoria de que existiría una oración ψ que cumpliría $\mathcal{V}_{\mathfrak{M}}(\psi) = C_\sim(\mathcal{V}_{\mathfrak{M}}(\psi))$.

Otra observación importante es que la deducción de (4.39) sería idéntica en un sistema multivaluado con más de tres valores de verdad. Así ocurre, por ejemplo, en una lógica paraconsistente, con un cuarto valor de verdad, $\mathfrak{p}$, para las oraciones que son verdaderas y falsas a la vez e incluso en una lógica difusa que pretenda representar la vaguedad de los predicados usando un conjunto continuo de valores de verdad —típicamente el intervalo de números reales [0,1]—. En todos estos casos, si dotamos de suficiente capacidad expresiva al lenguaje formal, no pueden

[33] Hay 8 funciones de W en W sin punto fijo. Una de ellas, que correspondería a lo que suele denominarse negación exclusiva, se definiría mediante $C_\sim(\mathfrak{f}) = C_\sim(\mathfrak{i}) = \mathfrak{v}$; $C_\sim(\mathfrak{v}) = \mathfrak{f}$.

coexistir el predicado T y la autorreferencialidad fuerte, porque entonces podríamos encontrar una fórmula de la forma $\rho(Tx)$ tal que la función $(C_\rho \circ C_T)$ no tuviese ningún punto fijo.[34] Como consecuencia, estamos también refutando que ciertas explicaciones de la paradoja basadas en lógicas difusas o inconsistentes sean satisfactorias.

Quizá la conclusión principal que sugiere lo estudiado hasta este momento, es que la paradoja del mentiroso es independiente de numerosas cuestiones semánticas, como si las oraciones tienen dos, tres o más valores de verdad; si el predicado de verdad es vago o si la generalización a una lógica multivaluada de la caracterización tarskiana de ese predicado se debe realizar de una u otra forma. Por el contrario, se refuerza la idea de que, si $\mathcal{V}_\mathfrak{M}$ es composicional para fórmulas no cuantificadas, la causa de la paradoja está en el conflicto entre la existencia de un predicado, T, que, para toda oración, φ, verifique $\mathcal{V}_\mathfrak{M}(T^\ulcorner\varphi^\urcorner) = C_T(\mathcal{V}_\mathfrak{M}(\varphi))$ y la autorreferencia, en grado suficiente para conseguir $\mathcal{V}_\mathfrak{M}(\psi) = \mathcal{V}_\mathfrak{M}(\rho(T^\ulcorner\psi^\urcorner))$, de modo que la función $(C_\rho \circ C_T)$ no tenga ningún punto fijo.

4.1.4. Contenidos enunciativos como portadores de verdad y paradoja del mentiroso

En los lenguajes interpretados extensionales, como los estudiados hasta el momento: 1/ la interpretación de un término cerrado es, si la tiene, su referencia, es decir un elemento del universo del modelo; 2/ la interpretación de una oración es su valor de verdad (su referencia fregeana).

Este planteamiento se ajusta a la idea de que los portadores de verdad son las oraciones (enunciativas). Una alternativa, cuanto menos razonable, consiste en que no es la oración lo que es verdadero o falso sino, si lo tiene, el contenido enunciativo de la oración cuando se profiere en un contexto determinado.

Hay muchas opiniones acerca de qué características e incluso qué nombre se debe dar a ese contenido enunciativo (proposición, enunciado, aserción...), pero no se trata, aquí, de discutirlas o añadir una más. Lo que interesa investigar es la idea, nada nueva, de que la paradoja del mentiroso podría desaparecer al considerar que los portadores de verdad son los contenidos enunciativos.

[34] Si fuera necesario, añadiríamos una conectiva nueva, π, definida mediante C_π, elegido de forma que $(C_\pi \circ C_T)$ no tenga ningún punto fijo.

En primer lugar, si hablamos estrictamente, las oraciones ya no son verdaderas ni falsas por lo que los contenidos enunciativos de las oraciones "esta oración es falsa" y "esta oración no es verdadera" serán respectivamente falso y verdadero y las oraciones dejarán de ser paradójicas. Pero, claro está, si los portadores de verdad no son las oraciones sino los contenidos enunciativos, las oraciones del mentiroso y del mentiroso reforzada habrá que reescribirlas como "el contenido enunciativo de esta oración es falso" y "el contenido enunciativo de esta oración no es verdadero".

La idea que supuestamente hace desaparecer la paradoja es que la oración del mentiroso reforzada (y lo mismo podría decirse de la del mentiroso simple) carece de contenido enunciativo. En tal caso, el sujeto de esa oración carece de referencia, lo que, a su vez, consolida la idea de que la oración carece de contenido enunciativo.

La réplica a esta idea consiste en construir una nueva oración del mentiroso más *reforzada* que la anterior: "esta oración carece de contenido enunciativo o tiene un contenido enunciativo falso". Si afirmamos que la oración carece de contenido enunciativo, lo que dice es verdad y, por tanto, no carece de contenido enunciativo. Si admitimos que la oración tiene un contenido enunciativo, ese contenido enunciativo tiene que ser, como todos, verdadero o falso. Pero, es fácil comprobar que cada uno de estos dos supuestos tiene como consecuencia el contrario.

Aunque esta réplica sugiere que el cambio de las oraciones por los contenidos enunciativos como portadores de verdad no supondrá una solución definitiva a la paradoja, ello no impide que intentemos una formalización de la misma con este nuevo planteamiento, dado que, las formalizaciones sirven a nuestro objetivo de extraer un esquema más depurado de la causa real de la paradoja.

4.1.4.1. Formalización

Deseamos construir un tipo de lenguajes interpretados en los que tengan cabida los contenidos enunciativos de las oraciones como portadores de verdad. La interpretación de las oraciones debe ser parcial, en el sentido de que se debe permitir la existencia de oraciones sin contenido enunciativo, para poder reflejar en el lenguaje formal la explicación informal de que algunas oraciones como "el contenido enunciativo de esta oración no es verdadero" dejan de ser paradójicas al considerar que carecen de contenido enunciativo. Por consiguiente, denominaremos lenguajes interpretados enunciativos parciales al tipo de lenguajes sugerido.

Una función de interpretación asociará a cada oración un contenido enunciativo (o ninguno). A su vez, puesto que todo contenido enunciativo es verdadero o es falso, necesitamos una función que asocie, a cada uno de ellos su valor de verdad. Denominaremos a esta función, función de extensionalización, dado que el contenido enunciativo suele considerarse una entidad intensional mientras que el valor de verdad es una entidad extensional.

Aunque el contenido enunciativo depende, en general, no solo de la oración enunciativa sino también del contexto en que se emite —tiempo, lugar, persona que afirma la oración...—, haremos abstracción de éste, igual que hemos hecho en los lenguajes interpretados anteriormente estudiados, pues, según hemos visto en esos lenguajes, no es preciso considerarlo para reflejar en el lenguaje formal aquellos aspectos del lenguaje natural relevantes en la paradoja del mentiroso. Consiguientemente, no buscamos un lenguaje interpretado en el que se asocie a cada oración enunciativa una función de contextos en contenidos enunciativos, ni tampoco, una función de contextos en valores de verdad; bastará con que se asocie, a cada oración, un contenido enunciativo (o ninguno).

El tipo de modelos que buscamos se parece más a la estructura intensional esbozada en Bealer (1998)[35] que a los modelos basados en la idea de que una proposición es la intensión de una oración y esta, a su vez, es una función que asigna a cada contexto el valor de verdad de la oración en ese contexto. Para Bealer, este tipo de modelos responde a una visión reduccionista de las proposiciones —reducidas a funciones extensionales— que él rechaza. Propone desarrollar una visión no reduccionista basándose en lo que denomina estructuras intensionales, donde las proposiciones, propiedades, relaciones binarias, ternarias, etc. son entidades primitivas. En una estructura intensional de Bealer hay un conjunto de funciones de extensionalización que asignan una extensión a cada entidad intensional y, en particular, un valor de verdad a cada proposición. Una de las funciones de extensionalización se distingue como la función de extensionalización real ("actual extensionalization function") —se entiende que las otras corresponden a otros mundos posibles—.

Ahora bien, no parece que la paradoja del mentiroso tenga una estrecha relación con conceptos modales o mundos posibles; de hecho, como hemos visto con el uso de lenguajes interpretados extensionales, no es necesario introducir dichos conceptos para estudiar la paradoja en un lenguaje formal. Consiguientemente, en

[35] Ver Jacquette (2002, p. 125, y nota al pie 12 en pp. 135-6).

nuestros lenguajes interpretados enunciativos parciales nos bastará con disponer de una única función de extensionalización.

No necesitamos en la formalización de la paradoja ninguna construcción referencialmente opaca salvo el entrecomillado. Por eso, aunque ahora queremos dar cabida a los contenidos enunciativos —las proposiciones de Bealer— no pretendemos profundizar en los aspectos intensionales del lenguaje natural. Queremos dejar abierta la forma en que, en el lenguaje interpretado, se asocien contenidos enunciativos a las oraciones ya que ello dependerá del modo de entender qué es exactamente un contenido enunciativo, cuándo dos oraciones distintas expresan el mismo contenido, etc. Lo que sí mantendremos es el principio de composicionalidad en la interpretación de las expresiones: se trata de un principio básico en los lenguajes formales interpretados que no tenemos motivo para abandonar.[36] Es interesante el estudio que Janssen realiza de dicho principio así como su principal conclusión que nos reafirma en nuestra decisión: la composicionalidad guía la investigación en la dirección correcta.[37]

Así pues, la interpretación de las oraciones dependerá de su estructura sintáctica y de la interpretación de sus partes significativas (términos, símbolos de predicado, etc.). Por supuesto, tampoco es objeto de este trabajo pretender estudiar cuál es la interpretación correcta de esas partes. Para poder trabajar en el estudio de la paradoja nos basta con que la función de extensionalización sea también composicional. A modo de ejemplo, si $\Pi\alpha, \beta$ representa el contenido enunciativo de la oración Pa, b (donde Π representa la interpretación de P, α la de a y β la de b), al aplicar la función de extensionalización E a $\Pi\alpha, \beta$ obtendremos un valor de verdad por aplicación de una función, $E(\Pi)$ a los valores $E(\alpha)$ y $E(\beta)$, es decir, $E(\Pi\alpha, \beta) = E(\Pi)(E(\alpha), E(\beta))$.

Si queremos introducir el predicado *verdadero* en el lenguaje formal, su argumento ha de ser aquello que se interpreta como un contenido enunciativo, es decir, una oración y no, como hasta ahora, un nombre de oración. Pero entonces debemos considerar una sintaxis más flexible en la que las oraciones son también términos.

Una alternativa es trasladar al lenguaje formal el concepto de *verdadero* no como predicado sino como operador que aplicado a un contenido enunciativo da

[36] Únicamente, hay que mantener la precaución de considerar los términos de cita como elementos sin partes —a efectos de composicionalidad.

[37] Janssen (1997, p. 461).

como resultado un contenido enunciativo, lo que significa que si en el lenguaje formal llamamos T a ese operador, T debe ir acompañado de una oración (aquello que se interpreta como un contenido enunciativo) que será su operando. Lo más directo es que, si la oración σ es el operando al que se aplica T, escribamos $T\sigma$ para indicarlo. Naturalmente, $T\sigma$ será, a su vez, una oración ya que debe interpretarse, en su caso, como un contenido enunciativo. Así pues, T tiene el comportamiento de una conectiva monádica.

En castellano, la diferencia entre usar *verdadero/verdad* como un predicado o como un operador es bastante clara: la oración "el contenido enunciativo de la oración "Carlos es alto" es verdadero" ejemplifica el uso de *verdadero* como predicado de contenidos enunciativos; la oración "es verdad que Carlos es alto" ejemplifica el uso de *verdad* como operador. Pero, si formalizamos "Carlos es alto" mediante Ac, y *verdad/verdadero* mediante T, la oración TAc serviría para formalizar ambas oraciones, solo que en el primer caso T se consideraría un predicado y Ac un término (además de ser una oración), mientras en el segundo, T se consideraría un operador y Ac una oración. La opción de considerar que las oraciones son también términos tiene como consecuencia que las oraciones puedan ser argumentos de cualesquiera predicados o cambiar las reglas sintácticas para que solo algunos términos puedan ser argumentos de ciertos predicados como T, lo cual es, en cualquier caso, una complejidad añadida sin interés para nosotros. Por todo ello, en el lenguaje formal trataremos T como un operador, sin que esto suponga un rechazo conceptual a considerar *verdadero* como un predicado, sino simplemente una opción que facilita nuestro trabajo. Gracias a esta elección, la sintaxis del lenguaje de predicados al que se añade capacidad de cita podemos reutilizarla con solo añadir T como una conectiva monádica.

Otra ventaja de tratar T como un operador es una simplificación del lenguaje interpretado que necesitamos. Como hemos visto, la explicación informal que evita la paradoja en la oración "el contenido enunciativo de esta oración no es verdadero" es que el sujeto de la oración carece de referencia (ya que la oración no tiene contenido enunciativo). Para formalizar esta explicación, que considera "verdadero" como un predicado, tendríamos que admitir términos cerrados fuera del dominio de la función de interpretación y oraciones sin contenido enunciativo a causa de contener alguno de esos términos. Mas, al tratar T como un operador, todo esto no es necesario, dado que la oración del mentiroso reforzada será de la forma $\neg T\varphi$ y la explicación que evite la paradoja se basará en que φ no

tiene contenido enunciativo y, por composicionalidad, tampoco lo tendrán$T\varphi$ ni $\neg T\varphi$.[38]

Ahora bien, puesto que queremos dar cabida a oraciones sin contenido enunciativo, necesitamos que los predicados y relaciones no sean totales, es decir, que no con cualesquiera argumentos cerrados produzcan un contenido enunciativo.

4.1.4.2. Lenguajes interpretados enunciativos parciales

Teniendo en cuenta las consideraciones anteriores, establecemos que un lenguaje interpretado enunciativo parcial, $(\mathcal{L},\xi,\mathfrak{M},E)_{enp}$, está constituido por un lenguaje formal, $\mathcal{L}$, una matriz, ξ, un modelo, $\mathfrak{M}$, y una función de extensionalización, E. El lenguaje $\mathcal{L}$ es el resultante de añadir una conectiva u operador monádico, T, al lenguaje de primer orden con capacidad de cita que hemos venido usando hasta el momento. La matriz, el modelo y la función de extensionalización deben ajustarse a las siguientes definiciones.

Definición 4.18 (matriz para lenguajes enunciativos bivaluados). *La matriz, ξ, para una interpretación enunciativa bivaluada del lenguaje$\mathcal{L}$, está formada por:*

1. *Un conjunto de dos valores de verdad: $W = \{\mathfrak{f},\mathfrak{v}\}$. Estableceremos en este conjunto una relación de orden de modo que $\mathfrak{f}$ sea anterior a $\mathfrak{v}$, es decir,* $\text{mín}(\{\mathfrak{f},\mathfrak{v}\}) = \mathfrak{f}$.

2. *Una interpretación extensional, C, de las constantes lógicas (conectivas y cuantificadores), es decir, una función total, C_k, por cada constante lógica. Las funciones $C_\neg$, $C_\wedge$, $C_\vee$, $C_\rightarrow$, $C_\leftrightarrow$, $C_\forall$ y $C_\exists$ son las de la lógica clásica (definición 4.3, p. 71). La única novedad la aporta el operador monádico T ("es verdad que"). Es claro que afirmar que un contenido enunciativo es verdadero debe tener el mismo valor de verdad que el propio contenido enunciativo. Por tanto, para cualquier$x \in W$ se ha de tener:*

$$C_T(x) =_{def} x \tag{4.40}$$

[38] Uno de los supuestos de la composicionalidad es que todas las partes propias de una expresión tienen interpretación (Janssen, 1997, p. 427, punto 5). Así para que$T\varphi$ tenga interpretación es preciso que φ la tenga, es decir, que tenga un contenido enunciativo. Al no ser así, la oración $T\varphi$ no tendría contenido enunciativo y, por ser una parte propia de$\neg T\varphi$, esta última oración, tampoco lo tendría.

3. *Un conjunto no vacío de contenidos enunciativos,* $\mathfrak{E}$*, llamado dominio enunciativo o universo enunciativo.*

4. *Una interpretación,* $\mathfrak{C}$*, de las constantes lógicas (conectivas y cuantificadores) que cumpla los siguientes requisitos:*

   a) *Asociar a cada conectiva n-ádica, k, una función total n-ádica,* $\mathfrak{C}_k$*, definida entre* $\mathfrak{E}^n$ *y* $\mathfrak{E}$*; y a cada cuantificador, una función total definida entre* $(\mathscr{P}(\mathfrak{E}) - \{\varnothing\})$ *y* $\mathfrak{E}$*.*

   b) *Ser compatible con la interpretación extensional de las constantes lógicas. Esto significa que dada cualquier función* F *que asocie a los elementos de* $\mathfrak{E}$ *elementos de* W*, debe cumplirse:*

      1) *Si* k *es una conectiva n-ádica y* $(\Phi_1...\Phi_n) \in \mathfrak{E}^n$*, entonces*

$$F(\mathfrak{C}_k(\Phi_1...\Phi_n)) =_{def} C_k(F(\Phi_1), ..., F(\Phi_n))$$

      2) *Si* Q *es un cuantificador,* B *es un elemento de* $(\mathscr{P}(\mathfrak{E}) - \{\varnothing\})$ *y escribimos* $F(B)$ *como forma abreviada de* $\{F(x)/x \in B\}$*, entonces*

$$F(\mathfrak{C}_Q(B)) =_{def} C_Q(F(B)).$$

Definición 4.19 (modelo para un lenguaje interpretado enunciativo parcial). *El modelo,* $\mathfrak{M}$*, para una interpretación enunciativa parcial de* $(\mathcal{L}, \xi)$ *está formado por:*

1. *Un conjunto no vacío,* $|\ \mathfrak{M}_x\ |$*, llamado dominio extensional o universo extensional de* $\mathfrak{M}$*. Puesto que* $\mathcal{L}$ *tiene capacidad de cita, el conjunto de expresiones de* $\mathcal{L}$ *debe ser un subconjunto de* $|\ \mathfrak{M}_x\ |$*.*

2. *Un conjunto no vacío,* $|\ \mathfrak{M}_y\ |$*.*[39]

3. *Una función de interpretación,* $\mathfrak{I}_{\mathfrak{M}}$*, definida parcialmente sobre el conjunto de símbolos de oración y, totalmente, sobre los símbolos de constante, los*

[39] La naturaleza de los elementos de este conjunto dependerá de cómo se considere la aportación de los términos cerrados al contenido enunciativo de las oraciones en que aparecen. Por ejemplo, si supusiéramos que su aportación es únicamente su referencia, tendríamos $|\ \mathfrak{M}_y\ | = |\ \mathfrak{M}_x\ |$. Si suponemos que su aportación son conceptos que permiten fijar la referencia, los elementos de $|\ \mathfrak{M}_y\ |$ serían conceptos, etc. Pero aquí no pretendemos decantarnos por ninguna opción al respecto. Únicamente decimos que si un término tiene interpretación en el modelo, esa interpretación es un elemento de $|\ \mathfrak{M}_y\ |$.

términos de cita, los símbolos de relación y los símbolos de función.[40] $\mathfrak{I}_{\mathfrak{M}}$ *asocia:*

- A cada símbolo de oración de su dominio, p, un contenido enunciativo, $\mathfrak{I}_{\mathfrak{M}}(p) \in \mathfrak{E}$.

- A cada símbolo de constante, c, un elemento de $\mid \mathfrak{M}_y \mid$, $\mathfrak{I}_{\mathfrak{M}}(c)$.

- A cada término de cita, $\ulcorner\varepsilon\urcorner$, *un elemento de* $\mid \mathfrak{M}_y \mid$, $\mathfrak{I}_{\mathfrak{M}}(\ulcorner\varepsilon\urcorner)$.

- A cada símbolo de relación n-ádica, R, una función parcial, $\mathfrak{I}_{\mathfrak{M}}(R)$, *definida entre* $\mid \mathfrak{M}_y \mid^n$ *y* $\mathfrak{E}$.

- A cada símbolo de función n-ádica, f, una función total, $\mathfrak{I}_{\mathfrak{M}}(f)$, *definida entre* $\mid \mathfrak{M}_y \mid^n$ *y* $\mid \mathfrak{M}_y \mid$.[41]

Definición 4.20 (función de extensionalización). *La función de extensionalización, E, de un lenguaje interpretado enunciativo parcial,* $(\mathcal{L}, \xi, \mathfrak{M}, E)_{enp}$ *se caracteriza por asociar a cada elemento de* $\mid \mathfrak{M}_y \mid \cup \mathfrak{E}$ *y a cada elemento del rango de las funciones de interpretación,* $\mathfrak{C}$ *e* $\mathfrak{I}_{\mathfrak{M}}$, *su correspondiente extensionalización:*

1. *E asocia a cada contenido enunciativo (elemento de* $\mathfrak{E}$*), un valor de verdad (elemento de* W*).*

2. *Para cada conectiva, k, se tiene* $E(\mathfrak{C}_k) = C_k$, *y para cada cuantificador, Q,* $E(\mathfrak{C}_Q) = C_Q$.

3. *E asocia a cada elemento de* $\mid \mathfrak{M}_y \mid$, *un elemento de* $\mid \mathfrak{M}_x \mid$ *y, en particular,* $E(\mathfrak{I}_{\mathfrak{M}}(\ulcorner\varepsilon\urcorner)) = \varepsilon$, *para cualquier expresión bien formada del lenguaje,* ε.

4. *Para todo elemento, o, de* $\mid \mathfrak{M}_x \mid$ *existe al menos un elemento, c, de* $\mid \mathfrak{M}_y \mid$ *tal que* $E(c) = o$.

5. *Dado un símbolo de relación n-ádico, R, la función E asocia a cada función parcial,* $\mathfrak{I}_{\mathfrak{M}}(R)$, *definida entre* $\mid \mathfrak{M}_y \mid^n$ *y* $\mathfrak{E}$, *una función parcial,* $E(\mathfrak{I}_{\mathfrak{M}}(R))$,

[40] Como señalamos, dejamos abierta la naturaleza de los valores de la función $\mathfrak{I}_{\mathfrak{M}}$, la cual dependerá de cómo se considere la aportación de los respectivos argumentos al contenido enunciativo de las oraciones en que aparecen.

[41] Recuérdese que, al tratar T como un operador, no es necesario tener términos sin denotación y, por tanto, podemos evitar la complejidad añadida que supondría que $\mathfrak{I}_{\mathfrak{M}}(f)$ pudiese ser una función parcial.

definida entre $|\ \mathfrak{M}_x\ |^n$ *y* W. *Si llamamos* D_R *al dominio de* $\mathfrak{I}_{\mathfrak{M}}(R)$, *el dominio de* $E(\mathfrak{I}_{\mathfrak{M}}(R))$ *será el conjunto:*

$$\{(a_1, ..., a_n) \in |\ \mathfrak{M}_x\ |^n /(b_1, ..., b_n) \in D_R,\ a_1 = E(b_1), ..., a_n = E(b_n)\}$$

6. *E asocia a cada función,* $\mathfrak{I}_{\mathfrak{M}}(f)$, *definida entre* $|\ \mathfrak{M}_y\ |^n$ *y* $|\ \mathfrak{M}_y\ |$, *una función,* $E(\mathfrak{I}_{\mathfrak{M}}(f))$, *definida entre* $|\ \mathfrak{M}_x\ |^n$ *y* $|\ \mathfrak{M}_x\ |$.

7. *Además, establecemos la composicionalidad de la función*E *para lo cual debe cumplirse:*

   a) *Si* R *es un símbolo de relación n-ádica,* D_R *es el dominio de* $\mathfrak{I}_{\mathfrak{M}}(R)$, *y* $(\tau_1...\tau_n) \in D_R$, *entonces*

   $$E(\mathfrak{I}_{\mathfrak{M}}(R)(\tau_1...\tau_n)) = E(\mathfrak{I}_{\mathfrak{M}}(R))(E(\tau_1), ..., E(\tau_n))$$

   b) *Si* $(\tau_1, \tau_2) \in |\ \mathfrak{M}_y\ |^2$ *y, en el metalenguaje, usamos el símbolo* $\equiv$ *para la identidad de elementos de* $|\ \mathfrak{M}_y\ |$ *o de* $\mathfrak{E}$ *y el símbolo* $=$ *para la identidad de elementos de* $|\ \mathfrak{M}_x\ |$ *o de* W, *entonces*

   $$E(\tau_1 \equiv \tau_2) =_{def} \begin{cases} \mathfrak{v} & si\ E(\tau_1) = E(\tau_2) \\ \mathfrak{f} & en\ otro\ caso \end{cases}$$

   c) *Si* f *es un símbolo de función n-ádica y* $(\tau_1...\tau_n) \in |\ \mathfrak{M}_y\ |^n$, *entonces*

   $$E(\mathfrak{I}_{\mathfrak{M}}(f)(\tau_1...\tau_n)) =_{def} E(\mathfrak{I}_{\mathfrak{M}}(f))(E(\tau_1), ..., E(\tau_n))$$

La asignación de variables en un lenguaje interpretado enunciativo parcial, $(\mathcal{L}, \xi, \mathfrak{M}, E)_{enp}$, es una función $s : V \rightarrow |\ \mathfrak{M}_y\ |$ (V es el conjunto de variables del lenguaje). Si $e \in |\ \mathfrak{M}_y\ |$, entenderemos que $s[e/x]$ es una asignación de variables igual que s salvo que asocia e a la variable x.

Dado un lenguaje interpretado enunciativo parcial, $(\mathcal{L}, \xi, \mathfrak{M}, E)_{enp}$, y una asignación de variables, s, para el mismo; la interpretación en el lenguaje con respecto a s, es la función $\mathcal{U}_{\mathfrak{M},s}$, establecida entre el conjunto de términos y $|\ \mathfrak{M}_y\ |$ y, parcialmente, entre el conjunto de fórmulas y $\mathfrak{E}$, de acuerdo con la siguiente definición.

Definición 4.21. *Sean:* c *un símbolo de constante o un término de cita,* x *una variable,* f *un símbolo de función n-ádica,* $t_1, t_2, ..., t_n$ *términos,* p *un símbolo de*

oración, R un símbolo de relación n-ádica, φ y ψ fórmulas. La función $\mathcal{U}_{\mathfrak{M},s}$ viene definida por:[42]

1. $\mathcal{U}_{\mathfrak{M},s}(c) \equiv_{def} \mathfrak{I}_{\mathfrak{M}}(c)$.
2. $\mathcal{U}_{\mathfrak{M},s}(x) \equiv_{def} s(x)$.
3. $\mathcal{U}_{\mathfrak{M},s}(f\,t_1, t_2, ..., t_n) \equiv_{def} \mathfrak{I}_{\mathfrak{M}}(f)(\mathcal{U}_{\mathfrak{M},s}(t_1), \mathcal{U}_{\mathfrak{M},s}(t_2), ..., \mathcal{U}_{\mathfrak{M},s}(t_n))$.
4. $\mathcal{U}_{\mathfrak{M},s}(p) \equiv_{def} \mathfrak{I}_{\mathfrak{M}}(p)$.
5. $\mathcal{U}_{\mathfrak{M},s}(t_1 = t_2) \equiv_{def} [\mathcal{U}_{\mathfrak{M},s}(t_1) \equiv \mathcal{U}_{\mathfrak{M},s}(t_2)]$.
6. $\mathcal{U}_{\mathfrak{M},s}(R\,t_1, t_2, ..., t_n) \equiv_{def} \mathfrak{I}_{\mathfrak{M}}(R)(\mathcal{U}_{\mathfrak{M},s}(t_1), \mathcal{U}_{\mathfrak{M},s}(t_2), ..., \mathcal{U}_{\mathfrak{M},s}(t_n))$
7. $\mathcal{U}_{\mathfrak{M},s}(T\varphi) \equiv_{def} \mathfrak{C}_T(\mathcal{U}_{\mathfrak{M},s}(\varphi))$.
8. $\mathcal{U}_{\mathfrak{M},s}(\neg\varphi) \equiv_{def} \mathfrak{C}_{\neg}(\mathcal{U}_{\mathfrak{M},s}(\varphi))$.
9. $\mathcal{U}_{\mathfrak{M},s}((\varphi \wedge \psi)) \equiv_{def} \mathfrak{C}_{\wedge}(\mathcal{U}_{\mathfrak{M},s}(\varphi), \mathcal{U}_{\mathfrak{M},s}(\psi))$.
10. $\mathcal{U}_{\mathfrak{M},s}((\varphi \vee \psi)) \equiv_{def} \mathfrak{C}_{\vee}(\mathcal{U}_{\mathfrak{M},s}(\varphi), \mathcal{U}_{\mathfrak{M},s}(\psi))$.
11. $\mathcal{U}_{\mathfrak{M},s}((\varphi \rightarrow \psi)) \equiv_{def} \mathfrak{C}_{\rightarrow}(\mathcal{U}_{\mathfrak{M},s}(\varphi), \mathcal{U}_{\mathfrak{M},s}(\psi))$.
12. $\mathcal{U}_{\mathfrak{M},s}((\varphi \leftrightarrow \psi)) \equiv_{def} \mathfrak{C}_{\leftrightarrow}(\mathcal{U}_{\mathfrak{M},s}(\varphi), \mathcal{U}_{\mathfrak{M},s}(\psi))$.
13. $\mathcal{U}_{\mathfrak{M},s}(\forall x\,\varphi) \equiv_{def} \mathfrak{C}_{\forall}(\{\mathcal{U}_{\mathfrak{M},s[e/x]}(\varphi)/e \in |\, \mathfrak{M}_y \,|\})$.
14. $\mathcal{U}_{\mathfrak{M},s}(\exists x\,\varphi) \equiv_{def} \mathfrak{C}_{\exists}(\{\mathcal{U}_{\mathfrak{M},s[e/x]}(\varphi)/e \in |\, \mathfrak{M}_y \,|\})$.

El valor que la función $\mathcal{U}_{\mathfrak{M},s}$ asocia a algunas fórmulas y términos es independiente de la asignación de variables s que se tome. Esto nos permite definir la función de interpretación, $\mathcal{U}_{\mathfrak{M}}$, como una restricción de $\mathcal{U}_{\mathfrak{M},s}$ a (el conjunto inicial de) esos términos y fórmulas.

Una consecuencia destacable del hecho de que E esté definido para todo elemento de $|\, \mathfrak{M}_y \,| \cup \mathfrak{E}$ es que, para todo h en el dominio de $\mathcal{U}_{\mathfrak{M},s}$, $\mathcal{U}_{\mathfrak{M},s}(h)$ está en el dominio de E. Será útil introducir las funciones de valuación $\mathcal{V}_{\mathfrak{M},s}$ y $\mathcal{V}_{\mathfrak{M}}$ mediante:

$$\mathcal{V}_{\mathfrak{M},s}(h) =_{def} E(\mathcal{U}_{\mathfrak{M},s}(h)) \tag{4.41}$$

$$\mathcal{V}_{\mathfrak{M}}(h) =_{def} E(\mathcal{U}_{\mathfrak{M}}(h)) \tag{4.42}$$

Obsérvese que el dominio de $\mathcal{V}_{\mathfrak{M},s}$ es el mismo que el de $\mathcal{U}_{\mathfrak{M},s}$ y el de $\mathcal{V}_{\mathfrak{M}}$ el mismo que el de $\mathcal{U}_{\mathfrak{M}}$. Por otra parte, $\mathcal{V}_{\mathfrak{M},s}$ es la función compuesta $(E \circ \mathcal{U}_{\mathfrak{M},s})$ y $\mathcal{V}_{\mathfrak{M}}$ es la función $(E \circ \mathcal{U}_{\mathfrak{M}})$.

También es destacable que, como en los lenguajes interpretados estudiados con anterioridad, $\mathcal{V}_{\mathfrak{M},s}(\forall x\,\varphi)$ y $\mathcal{V}_{\mathfrak{M},s}(\exists x\,\varphi)$ proporcionan una interpretación objetual

[42] En las siguientes definiciones considérese indefinido el *definiendum* cuando el *definiens* es indefinido.

de los cuantificadores. En efecto, tomemos el cuantificador universal.[43] Es sencillo ver que

$$\mathcal{V}_{\mathfrak{M},s}(\forall x\,\varphi) = E(\mathcal{U}_{\mathfrak{M},s}(\forall x\,\varphi)) = C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\varphi)/e \in |\,\mathfrak{M}_y\,|\}) \tag{4.43}$$

Y, para que la interpretación del cuantificador sea objetual, es necesario y suficiente que $E(e)$ *recorra* exactamente todos los *objetos*, es decir, que se cumpla: 1/ dado un e cualquiera perteneciente a $|\,\mathfrak{M}_y\,|$, $E(e) \in |\,\mathfrak{M}_x\,|$; 2/ dado, o, perteneciente a $|\,\mathfrak{M}_x\,|$ debe existir al menos un elemento, e, de $|\,\mathfrak{M}_y\,|$ tal que $E(e) = o$. Ahora bien, estos dos requisitos están garantizados, respectivamente, por los puntos 3 (p. 108) y 4 de la definición de E.

Diremos que la fórmula φ es verdadera en $\mathfrak{M}$, o que $\mathfrak{M}$ satisface φ, cuando $\mathcal{V}_{\mathfrak{M}}(\varphi) = \mathfrak{v}$. Entonces escribiremos: $\mathfrak{M} \models \varphi$.

Fijados un lenguaje $\mathcal{L}$ y una matriz ξ propios de un lenguaje interpretado enunciativo parcial, diremos que la fórmula ψ es consecuencia lógica de la fórmula φ cuando para cualquier función de extensionalización, cualquier modelo$\mathfrak{M}$ y cualquier asignación de variables s —propios de ese tipo de lenguajes interpretados— si $\mathcal{V}_{\mathfrak{M},s}(\varphi) = \mathfrak{v}$ entonces $\mathcal{V}_{\mathfrak{M},s}(\psi) = \mathfrak{v}$. Esto se expresa habitualmente mediante:

$$\varphi \models \psi$$

También diremos que las fórmulas φ y ψ son lógicamente equivalentes cuando una es consecuencia lógica de la otra y viceversa.

Un lenguaje interpretado bivaluado con capacidad de cita se puede considerar un caso particular de un lenguaje interpretado enunciativo parcial: el caso en que E es la función identidad (y, por ello, $|\,\mathfrak{M}_x\,| = |\,\mathfrak{M}_y\,|$ y $\mathfrak{E} = W$) y tanto la función $\mathfrak{I}_{\mathfrak{M}}$ como las funciones $\mathfrak{I}_{\mathfrak{M}}(R)$, para cualquier símbolo de relación, R, son funciones totales. Esto explica que en un lenguaje interpretado extensional las funciones de interpretación y valuación sean la misma.

Como en un lenguaje interpretado de primer orden clásico, en un lenguaje interpretado enunciativo parcial se cumple:

Proposición 4.4. *Sean t_1, t_2 dos términos cerrados, entonces:*

$$\mathfrak{M} \models t_1 = t_2 \quad ssi \quad \mathcal{V}_{\mathfrak{M}}(t_1) = \mathcal{V}_{\mathfrak{M}}(t_2)$$

[43] El mismo razonamiento será válido para el cuantificador existencial.

Demostración.

(1) $\mathcal{U}_{\mathfrak{M}}(t_1 = t_2) = [\mathcal{U}_{\mathfrak{M}}(t_1) \equiv \mathcal{U}_{\mathfrak{M}}(t_2)]$
; def. 4.21, punto 5 (p. 110), dado que t_1 y t_2 son términos cerrados

(2) $E(\mathcal{U}_{\mathfrak{M}}(t_1 = t_2)) = E(\mathcal{U}_{\mathfrak{M}}(t_1) \equiv \mathcal{U}_{\mathfrak{M}}(t_2))$; de 1

(3) $E(\mathcal{U}_{\mathfrak{M}}(t_1) \equiv \mathcal{U}_{\mathfrak{M}}(t_2)) = \mathfrak{v}$ *ssi* $E(\mathcal{U}_{\mathfrak{M}}(t_1)) = E(\mathcal{U}_{\mathfrak{M}}(t_2))$
; definición función de extensionalización (def. 4.20, punto 7*b*, p. 109)

(4) $E(\mathcal{U}_{\mathfrak{M}}(t_1 = t_2)) = \mathfrak{v}$ *ssi* $E(\mathcal{U}_{\mathfrak{M}}(t_1)) = E(\mathcal{U}_{\mathfrak{M}}(t_2))$
; de 2 y 3

(5) $\mathcal{V}_{\mathfrak{M}}(t_1 = t_2) = \mathfrak{v}$ *ssi* $\mathcal{V}_{\mathfrak{M}}(t_1) = \mathcal{V}_{\mathfrak{M}}(t_2)$
; de 4 y definición de $\mathcal{V}_{\mathfrak{M}}$

(6) $\mathfrak{M} \models t_1 = t_2$ *ssi* $\mathcal{V}_{\mathfrak{M}}(t_1) = \mathcal{V}_{\mathfrak{M}}(t_2)$
; de 5 y concepto de $\mathfrak{M} \models \varphi$

■

La extensionalidad de $\mathcal{V}_{\mathfrak{M}}$ se conserva siempre que sus argumentos estén en el dominio de la función. Por ejemplo, dada una fórmula $\varsigma(x)$, de $\mathcal{V}_{\mathfrak{M}}(t_1) = \mathcal{V}_{\mathfrak{M}}(t_2)$ se deduce $\mathcal{V}_{\mathfrak{M}}(\varsigma(t_1)) = \mathcal{V}_{\mathfrak{M}}(\varsigma(t_2))$ solo si $\varsigma(t_1)$ y $\varsigma(t_2)$ están en el dominio de $\mathcal{V}_{\mathfrak{M}}$.

Proposición 4.5. *Dados una fórmula con una única variable libre, $\varsigma(x)$, y los términos cerrados t_1 y t_2; si $\mathcal{V}_{\mathfrak{M}}(t_1) = \mathcal{V}_{\mathfrak{M}}(t_2)$ y $\varsigma(t_1)$ y $\varsigma(t_2)$ están en el dominio de $\mathcal{V}_{\mathfrak{M}}$, entonces $\mathcal{V}_{\mathfrak{M}}(\varsigma(t_1)) = \mathcal{V}_{\mathfrak{M}}(\varsigma(t_2))$.*

Demostración. Puesto que E y $\mathcal{U}_{\mathfrak{M}}$ se definen composicionalmente y $\mathcal{V}_{\mathfrak{M}}$ es la función $(E \circ \mathcal{U}_{\mathfrak{M}})$, $\mathcal{V}_{\mathfrak{M}}$ es composicional cuando se aplica a un argumento de su dominio. Por eso, si $\varsigma(t_1) \in dom(\mathcal{V}_{\mathfrak{M}})$ y $\varsigma(t_2) \in dom(\mathcal{V}_{\mathfrak{M}})$, existirá una función, C_ς, tal que $\mathcal{V}_{\mathfrak{M}}(\varsigma(t_1)) = C_\varsigma(\mathcal{V}_{\mathfrak{M}}(t_1))$ y $\mathcal{V}_{\mathfrak{M}}(\varsigma(t_2)) = C_\varsigma(\mathcal{V}_{\mathfrak{M}}(t_2))$. Por otra parte, si $\mathcal{V}_{\mathfrak{M}}(t_1) = \mathcal{V}_{\mathfrak{M}}(t_2)$, se cumple $C_\varsigma(\mathcal{V}_{\mathfrak{M}}(t_1)) = C_\varsigma(\mathcal{V}_{\mathfrak{M}}(t_2))$. Finalmente tendremos: $\mathcal{V}_{\mathfrak{M}}(\varsigma(t_1)) = C_\varsigma(\mathcal{V}_{\mathfrak{M}}(t_1)) = C_\varsigma(\mathcal{V}_{\mathfrak{M}}(t_2)) = \mathcal{V}_{\mathfrak{M}}(\varsigma(t_2))$.

■

4.1.4.3. Autorreferencia en un lenguaje interpretado enunciativo parcial.

La introducción de la función de valuación $\mathcal{V}_{\mathfrak{M}}$ en los lenguajes interpretados enunciativos parciales nos permite mantener las definiciones de oración autorreferencial, lenguaje fuertemente autorreferencial y diagonalización.

Sin embargo, la proposición que correspondería a la proposición 4.1 (p. 83), no se cumple en un lenguaje interpretado enunciativo parcial: que el lenguaje sea

fuertemente autorreferencial no garantiza que para toda fórmula con una única variable libre, $\sigma(x)$, exista un término, τ, tal que:

$$\mathcal{V}_{\mathfrak{M}}(\sigma(\tau)) = \mathcal{V}_{\mathfrak{M}}(\sigma(\ulcorner\sigma(\tau)\urcorner)) \tag{4.13}$$

es decir, tal que:

$$E(\mathcal{U}_{\mathfrak{M}}(\sigma(\tau))) = E(\mathcal{U}_{\mathfrak{M}}(\sigma(\ulcorner\sigma(\tau)\urcorner))) \tag{4.44}$$

A pesar de que, en un lenguaje fuertemente autorreferencial, está garantizado $\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\sigma(\tau)\urcorner)$, y su equivalente, $E(\mathcal{U}_{\mathfrak{M}}(\tau)) = E(\mathcal{U}_{\mathfrak{M}}(\ulcorner\sigma(\tau)\urcorner))$, de aquí no se deduce (4.44), ya que, en un lenguaje interpretado enunciativo parcial, $\mathcal{U}_{\mathfrak{M}}(\sigma(\tau))$ podría no estar definida aunque $\mathcal{U}_{\mathfrak{M}}(\tau)$ sí lo esté.

En cambio, se sigue cumpliendo la proposición 4.2 (p. 85) por la que si en el lenguaje interpretado hay una función de diagonalización entonces es fuertemente autorreferencial.

4.1.4.4. Paradoja del mentiroso

Consideremos, en primer lugar, la oración "el contenido enunciativo de esta oración no es verdadero". Teniendo en cuenta su estructura, podría formalizarse mediante $\neg TS\tau$, en un lenguaje en el que T represente un predicado de verdad, τ sea un término cuya referencia sea la oración $\neg TS\tau$ y S sea un símbolo de la función que asocia a cada oración su contenido enunciativo (si lo tiene).

Queremos, sin embargo, formalizar la oración en nuestro último lenguaje formal, donde T es un operador en vez de un predicado. Para ello podemos volver a usar la oración $\neg TS\tau$ con la diferencia de que ahora T es un operador y S representa un predicado de contenido enunciativo de acuerdo con la siguiente definición.

Definición 4.22 (predicado de contenido enunciativo). *En un lenguaje interpretado enunciativo parcial, $(\mathcal{L}, \xi, \mathfrak{M}, E)_{enp}$, S representa un predicado de contenido enunciativo, si y solo si, para toda oración φ en el dominio de $\mathcal{U}_{\mathfrak{M}}$ y todo término ν tales que $\mathcal{V}_{\mathfrak{M}}(\nu) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\varphi\urcorner)$ se cumple:*

$$\mathcal{U}_{\mathfrak{M}}(S\nu) \equiv \mathcal{U}_{\mathfrak{M}}(\varphi) \tag{4.45}$$

Dos consecuencias de (4.45) son:

$$\mathcal{U}_{\mathfrak{M}}(S^{\ulcorner}\varphi^{\urcorner}) \equiv \mathcal{U}_{\mathfrak{M}}(\varphi) \tag{4.46}$$

$$\mathcal{V}_{\mathfrak{M}}(S^{\ulcorner}\varphi^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\varphi) \tag{4.47}$$

Es destacable que, el predicado S hace una función de desentrecomillado similar a la que hacía el predicado T en los lenguajes interpretados que utilizamos con anterioridad.

Por otra parte, debido a la composicionalidad de $\mathcal{U}_{\mathfrak{M}}$ y E en $(\mathcal{L}, \xi, \mathfrak{M}, E)_{enp}$, para toda oración φ en el dominio de $\mathcal{U}_{\mathfrak{M}}$, se cumple $\mathcal{U}_{\mathfrak{M}}(T\varphi) \equiv \mathfrak{C}_T(\mathcal{U}_{\mathfrak{M}}(\varphi))$ así como $E(\mathfrak{C}_T(\mathcal{U}_{\mathfrak{M}}(\varphi))) = E(\mathfrak{C}_T)(E(\mathcal{U}_{\mathfrak{M}}(\varphi)))$. Si además tenemos en cuenta que $E_{\mathfrak{M}}(\mathfrak{C}_T)$ es C_T y que $\mathcal{V}_{\mathfrak{M}}$ es $(E \circ \mathcal{U}_{\mathfrak{M}})$, obtenemos,

$$\mathcal{V}_{\mathfrak{M}}(T\varphi) = E(\mathcal{U}_{\mathfrak{M}}(T\varphi)) = E(\mathfrak{C}_T(\mathcal{U}_{\mathfrak{M}}(\varphi))) \tag{4.48}$$

y

$$E(\mathfrak{C}_T(\mathcal{U}_{\mathfrak{M}}(\varphi))) = E(\mathfrak{C}_T)(E(\mathcal{U}_{\mathfrak{M}}(\varphi))) = C_T(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.49}$$

de donde

$$\mathcal{V}_{\mathfrak{M}}(T\varphi) = C_T(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.50}$$

Pero como además, para cualquier valor de verdad x, $C_T(x) = x$, ((4.40) p. 106), podemos concluir:

$$\mathcal{V}_{\mathfrak{M}}(T\varphi) = \mathcal{V}_{\mathfrak{M}}(\varphi) \tag{4.51}$$

En el tipo de lenguajes interpretados que estamos considerando, hay pues una importante similitud entre el predicado S y el operador T, dado que, como se deduce de (4.47) y (4.51), para toda oración φ en el dominio de $\mathcal{U}_{\mathfrak{M}}$:

$$\mathcal{V}_{\mathfrak{M}}(T\varphi) = \mathcal{V}_{\mathfrak{M}}(S^{\ulcorner}\varphi^{\urcorner}) \tag{4.52}$$

Gracias al predicado S tenemos instrumentos suficientes para formalizar la oración "el contenido enunciativo de esta oración no es verdadero" mediante una oración de la forma $\neg T\varphi$ donde la interpretación pretendida de la oración φ sea el contenido enunciativo de $\neg T\varphi$. Basta con tomar φ de la forma $S\tau$ y que la referencia del término τ sea la propia oración $\neg TS\tau$. Esto sabemos conseguirlo si en el lenguaje disponemos de una función de diagonalización, ya que entonces,

será fuertemente autorreferencial y, por ello, dada la fórmula $\neg TSx$, existe un término, τ, cuya referencia es la oración $\neg TS\tau$, es decir:

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner \neg TS\tau \urcorner) \tag{4.53}$$

Pero, la siguiente proposición nos indica que de (4.53) se deduce que la oración $\neg TS\tau$ no está en el dominio de $\mathcal{U}_{\mathfrak{M}}$.

Proposición 4.6. *Si* $\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner \neg TS\tau \urcorner)$ *entonces* $\neg TS\tau \notin dom(\mathcal{U}_{\mathfrak{M}})$.

Demostración.

1	(1)	$\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner \neg TS\tau \urcorner)$	; premisa
2	(2)	$\neg TS\tau \in dom(\mathcal{U}_{\mathfrak{M}})$	; hipótesis auxiliar
	(3)	$dom(\mathcal{U}_{\mathfrak{M}}) = dom(\mathcal{V}_{\mathfrak{M}})$	
		; propiedad de nuestros lenguajes interpretados enunciativos	
2	(4)	$\neg TS\tau \in dom(\mathcal{V}_{\mathfrak{M}})$	; 2 y 3
1,2	(5)	$\mathcal{U}_{\mathfrak{M}}(S\tau) \equiv \mathcal{U}_{\mathfrak{M}}(\neg TS\tau)$	
		; 1, 2 y definición de S (definición 4.22, p. 113)	
1,2	(6)	$S\tau \in dom(\mathcal{U}_{\mathfrak{M}}) = dom(\mathcal{V}_{\mathfrak{M}})$	; 3 y 5
1,2	(7)	$E(\mathcal{U}_{\mathfrak{M}}(S\tau)) = E(\mathcal{U}_{\mathfrak{M}}(\neg TS\tau))$	; 5
1,2	(8)	$\mathcal{V}_{\mathfrak{M}}(S\tau) = \mathcal{V}_{\mathfrak{M}}(\neg TS\tau)$	; 7, dado que $\mathcal{V}_{\mathfrak{M}}$ es $(E \circ \mathcal{U}_{\mathfrak{M}})$
2	(9)	$\mathcal{V}_{\mathfrak{M}}(\neg TS\tau) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(TS\tau))$	; 4 y composicionalidad
1,2	(10)	$\mathcal{V}_{\mathfrak{M}}(TS\tau) = \mathcal{V}_{\mathfrak{M}}(S\tau)$	; 6 y (4.51)
1,2	(11)	$\mathcal{V}_{\mathfrak{M}}(S\tau) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(S\tau))$	; 8, 9 y 10
1,2	(12)	$\mathcal{V}_{\mathfrak{M}}(S\tau) \neq C_{\neg}(\mathcal{V}_{\mathfrak{M}}(S\tau))$	; 6 y $C_{\neg}(\mathfrak{v}) = \mathfrak{f}$; $C_{\neg}(\mathfrak{f}) = \mathfrak{v}$
1,2	(13)	$\mathcal{V}_{\mathfrak{M}}(S\tau) \neq \mathcal{V}_{\mathfrak{M}}(S\tau)$	; 11 y 12
1	(14)	$\neg TS\tau \notin dom(\mathcal{U}_{\mathfrak{M}})$	
		; falsedad de 13 y eliminación de hipótesis 2	

Hemos concluido así que la oración $\neg TS\tau$ carece de contenido enunciativo ■

La proposición anterior ha servido para comprobar que, cuando un término τ tiene como referencia la oración $\neg TS\tau$, no hay una contradicción sino la conclusión de que dicha oración carece de contenido enunciativo. De este modo, hemos trasladado al lenguaje formal el razonamiento informal sobre la oración "el contenido enunciativo de esta oración no es verdadero", por el cual podemos escapar de la paradoja afirmando que la oración carece de contenido enunciativo.

Pero, dijimos, la escapatoria anterior no sirve con la oración "esta oración carece de contenido enunciativo o tiene un contenido enunciativo falso". Al querer formalizarla nos encontramos con que, en principio, no disponemos en nuestro lenguaje de un medio para expresar si una oración tiene o no contenido enunciativo. Necesitaríamos un nuevo predicado, R, en consonancia con la siguiente definición.

Definición 4.23 (predicado de existencia de contenido enunciativo). *En un lenguaje interpretado enunciativo parcial,* $(\mathcal{L}, \xi, \mathfrak{M}, E)_{enp}$*, R representa un predicado de existencia de contenido enunciativo, si y solo si, para toda oración* φ *en el dominio de* $\mathcal{U}_{\mathfrak{M}}$ *y todo término* ν *tales que* $\mathcal{V}_{\mathfrak{M}}(\nu) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\varphi\urcorner)$ *se cumple:*

$$\mathcal{U}_{\mathfrak{M}}(R\nu) \equiv (\varphi \in dom(\mathcal{U}_{\mathfrak{M}})) \tag{4.54}$$

(Suponemos que, en el metalenguaje, $\varphi \in dom(\mathcal{U}_{\mathfrak{M}})$ *representa el contenido enunciativo de "*φ *está en el dominio de* $\mathcal{U}_{\mathfrak{M}}$*").*

La consecuencia extensional de la definición anterior es que

$$\mathcal{V}_{\mathfrak{M}}(R\nu) = \begin{cases} \mathfrak{v} & si\ \varphi \in dom(\mathcal{U}_{\mathfrak{M}}) \\ \mathfrak{f} & si\ \varphi \notin dom(\mathcal{U}_{\mathfrak{M}}) \end{cases} \tag{4.55}$$

En principio, para expresar que una oración, φ, carece de contenido enunciativo o tiene un contenido enunciativo falso —es decir, no verdadero— debería servir la oración

$$\neg R\ulcorner\varphi\urcorner \vee \neg T\varphi \tag{4.56}$$

pero no es así. Cuando una oración carece de contenido enunciativo, informalmente consideramos que afirmar sobre ella que *carece de contenido enunciativo o tiene un contenido enunciativo falso* es afirmar algo verdadero —al menos es una manera bastante razonable de entender la afirmación, que será la que supongamos a continuación—. Por tanto, si φ carece de contenido enunciativo ($\varphi \notin dom(\mathcal{U}_{\mathfrak{M}})$) y la fórmula $\neg R\ulcorner\varphi\urcorner \vee \neg T\varphi$ expresara que la oración φ carece de contenido enunciativo o tiene un contenido enunciativo falso, dicha fórmula debería ser verdadera, es decir, debería cumplirse $\mathcal{V}_{\mathfrak{M}}(\neg R\ulcorner\varphi\urcorner \vee \neg T\varphi) = \mathfrak{v}$. Sin embargo, en nuestro lenguaje interpretado, dadas dos oraciones χ y ψ, si $\psi \notin dom(\mathcal{U}_{\mathfrak{M}})$ entonces $(\chi \vee \psi) \notin dom(\mathcal{U}_{\mathfrak{M}})$ incluso aunque $\mathcal{V}_{\mathfrak{M}}(\chi) = \mathfrak{v}$ porque la composicionalidad de $\mathcal{U}_{\mathfrak{M}}$ exige que todas las partes propias de una oración estén en $dom(\mathcal{U}_{\mathfrak{M}})$ para que la oración también lo esté. Aplicado a nuestro caso, si $\varphi \notin dom(\mathcal{U}_{\mathfrak{M}})$, tendremos

que $\neg T\varphi \notin dom(\mathcal{U}_{\mathfrak{M}})$ y que $(\neg R^{\ulcorner}\varphi^{\urcorner} \vee \neg T\varphi) \notin dom(\mathcal{U}_{\mathfrak{M}}) = dom(\mathcal{V}_{\mathfrak{M}})$ por lo que no podremos afirmar $\mathcal{V}_{\mathfrak{M}}(\neg R^{\ulcorner}\varphi^{\urcorner} \vee \neg T\varphi) = \mathfrak{v}$.

Esta disonancia entre nuestro lenguaje formal interpretado y el natural se puede soslayar si, en lugar de combinar el predicado R con el operador T, usamos un predicado, Q, que exprese la propiedad "carecer de contenido enunciativo o tener un contenido enunciativo falso" en concordancia a como la hemos entendido en castellano. Por tanto, Q debe satisfacer, para todo término ν que tenga como referencia la oración φ (es decir, que cumpla $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$):

$$\mathcal{U}_{\mathfrak{M}}(Q\nu) \equiv (\varphi \notin dom(\mathcal{U}_{\mathfrak{M}})\, o\, \mathcal{V}_{\mathfrak{M}}(\varphi) = \mathfrak{f}) \tag{4.57}$$

siempre que la expresión metalingüística $\varphi \notin dom(\mathcal{U}_{\mathfrak{M}})\, o\, \mathcal{V}_{\mathfrak{M}}(\varphi) = \mathfrak{f}$ se entienda como hemos dicho, es decir, de modo que cuando φ no esté en el dominio de $\mathcal{U}_{\mathfrak{M}}$, sea verdadera. En concreto, la extensionalización de (4.57) ha de ser, para todo ν tal que $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$, la siguiente:

$$\mathcal{V}_{\mathfrak{M}}(Q\nu) = \begin{cases} \mathcal{V}_{\mathfrak{M}}(\neg T\varphi) & si\ \varphi \in dom(\mathcal{U}_{\mathfrak{M}}) \\ \mathfrak{v} & si\ \varphi \notin dom(\mathcal{U}_{\mathfrak{M}}) \end{cases} \tag{4.58}$$

Dado que, si $\varphi \in dom(\mathcal{U}_{\mathfrak{M}})$, se cumple $\mathcal{V}_{\mathfrak{M}}(T\varphi) = \mathcal{V}_{\mathfrak{M}}(\varphi)$, (4.51), tendremos $C_{\neg}(\mathcal{V}_{\mathfrak{M}}(T\varphi)) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(\varphi))$ y la composicionalidad nos permitirá deducir que $\mathcal{V}_{\mathfrak{M}}(\neg T\varphi) = \mathcal{V}_{\mathfrak{M}}(\neg\varphi)$. Por consiguiente, podemos escribir de modo algo más simple la condición (4.58):

$$\mathcal{V}_{\mathfrak{M}}(Q\nu) = \begin{cases} \mathcal{V}_{\mathfrak{M}}(\neg\varphi) & si\ \varphi \in dom(\mathcal{U}_{\mathfrak{M}}) \\ \mathfrak{v} & si\ \varphi \notin dom(\mathcal{U}_{\mathfrak{M}}) \end{cases} \tag{4.59}$$

de la cual se deduce el siguiente corolario:

Corolario 4.1. *Dado un término cualquiera, ν, cuya referencia sea una oración: $Q\nu \in dom(\mathcal{U}_{\mathfrak{M}}) = dom(\mathcal{V}_{\mathfrak{M}})$.*

La formalización de "esta oración carece de contenido enunciativo o tiene un contenido enunciativo falso" es la oración $Q\tau$, donde $\mathcal{V}_{\mathfrak{M}}(\tau) = Q\tau$ —lo que confiere a $Q\tau$ su carácter de oración autorreferencial—. Pero ahora es inevitable la contradicción:

Proposición 4.7. *Si, φ es una oración y siempre que $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$, se cumple (4.59), entonces no hay un término, τ, tal que $\mathcal{V}_{\mathfrak{M}}(\tau) = Q\tau$.*

Demostración.

1	(1)	$\mathcal{V}_{\mathfrak{M}}(\tau) = Q\tau$	; hipótesis
	(2)	$\mathcal{V}_{\mathfrak{M}}(\ulcorner Q\tau \urcorner) = Q\tau$	; para toda oración φ, $\mathcal{V}_{\mathfrak{M}}(\ulcorner \varphi \urcorner) = \varphi$
1	(3)	$\mathcal{V}_{\mathfrak{M}}(\ulcorner Q\tau \urcorner) = \mathcal{V}_{\mathfrak{M}}(\tau)$	; 1 y 2
1	(4)	$Q\ulcorner Q\tau \urcorner \in dom(\mathcal{V}_{\mathfrak{M}})$, $Q\tau \in dom(\mathcal{V}_{\mathfrak{M}})$	; 1, 2 y corolario 4.1
1	(5)	$\mathcal{V}_{\mathfrak{M}}(Q\ulcorner Q\tau \urcorner) = \mathcal{V}_{\mathfrak{M}}(Q\tau)$	; 3, 4 y proposición 4.5 (p. 112)
1	(6)	$\mathcal{V}_{\mathfrak{M}}(Q\ulcorner Q\tau \urcorner) = \mathcal{V}_{\mathfrak{M}}(\neg Q\tau)$	; 4 y (4.59)
1	(7)	$\mathcal{V}_{\mathfrak{M}}(Q\tau) = \mathcal{V}_{\mathfrak{M}}(\neg Q\tau)$	; 5 y 6

Pero, en un lenguaje interpretado enunciativo parcial, (7) es imposible, luego podemos negar la hipótesis, $\mathcal{V}_{\mathfrak{M}}(\tau) = Q\tau$, de la que depende (7). ■

4.1.5. Conclusiones

Nuestro análisis de las oraciones autorreferenciales $\neg TS\tau$ y $Q\tau$ muestra que, en el primer caso no hay contradicción debido a que $\neg TS\tau \notin dom(\mathcal{U}_{\mathfrak{M}})$, lo que es posible en un lenguaje interpretado enunciativo parcial. Pero, en el segundo, no es posible $Q\tau \notin dom(\mathcal{U}_{\mathfrak{M}})$, porque la caracterización de Q conlleva que para cualquier término, τ, cuya referencia sea una oración, $Q\tau \in dom(\mathcal{U}_{\mathfrak{M}})$.

Por otra parte, antes de estudiar los lenguajes interpretados enunciativos parciales, habíamos llegado a la conclusión de que la autorreferencialidad fuerte es incompatible con la existencia de un predicado de verdad,T. Sin embargo, en nuestra última formalización de la paradoja no hemos necesitado ningún predicado —ni operador— T: la autorreferencia ha resultado ser incompatible con el predicado Q que no es estrictamente un predicado de verdad. La propiedad esencial que los predicados Q y T comparten es que realizan una función de desentrecomillado.

En resumen, una primera conclusión importante es que el intento de formalizar una oración del mentiroso, en un sistema composicional, acaba sacando a la luz una incompatibilidad entre la autorreferencialidad fuerte y la existencia de un predicado, D, que: a) para cualquier oración φ, cumpla $D\ulcorner \varphi \urcorner \in dom(\mathcal{V}_{\mathfrak{M}})$; b) realice una función de *desentrecomillado* en el sentido de que, dada una oración, φ, referenciada por un término, ν (es decir, $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$), se cumpla:

$$\mathcal{V}_{\mathfrak{M}}(D\nu) = \begin{cases} C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)) & si\ \varphi \in dom(\mathcal{V}_{\mathfrak{M}}) \\ w & si\ \varphi \notin dom(\mathcal{V}_{\mathfrak{M}}) \end{cases} \tag{4.60}$$

siendo w un elemento del conjunto W de valores de verdad y C_D una función total definida entre W y W. Puesto que $\ulcorner\varphi\urcorner$ es un término cuya referencia es la oración φ ($\mathcal{V}_{\mathfrak{M}}(\ulcorner\varphi\urcorner) = \varphi$), una consecuencia trivial es:

$$\mathcal{V}_{\mathfrak{M}}(D\ulcorner\varphi\urcorner) = \begin{cases} C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)) & si\ \varphi \in dom(\mathcal{V}_{\mathfrak{M}}) \\ w & si\ \varphi \notin dom(\mathcal{V}_{\mathfrak{M}}) \end{cases} \tag{4.61}$$

donde el desentrecomillado es más explícito.

Interesa señalar que en un lenguaje interpretado total, es decir, aquel en el que para cualquier oración, φ, se cumple $\varphi \in dom(\mathcal{V}_{\mathfrak{M}})$, la condición (4.60) se simplifica.

Es fácil ver que los predicados T y Q que hemos usado, se atienen a este esquema. Por ejemplo, en el caso del predicado Q, teniendo en cuenta (4.59) y que $\mathcal{V}_{\mathfrak{M}}(\neg\varphi) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(\varphi))$, la función C_D del esquema (4.60) sería $C_{\neg}$ —función que verifica $C_{\neg}(\mathfrak{f}) = \mathfrak{v}$, $C_{\neg}(\mathfrak{v}) = \mathfrak{f}$— y w sería $\mathfrak{v}$.

Si combinamos la existencia de un predicado del tipo de D con la autorreferencialidad fuerte podemos generalizar el razonamiento que, al intentar formalizar una oración autorreferencial, ψ, pone en relación $\mathcal{V}_{\mathfrak{M}}(\psi)$ con una función de sí misma.

Tomemos una fórmula de la forma $\rho(Dx)$. Si el lenguaje es fuertemente autorreferencial, existe un término, τ, cuya referencia es la oración $\rho(D\tau)$, es decir:

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\rho(D\tau)\urcorner) \tag{4.62}$$

Y, según la caracterización (4.60) de D:

$$\mathcal{V}_{\mathfrak{M}}(D\tau) = \mathcal{V}_{\mathfrak{M}}(D\ulcorner\rho(D\tau)\urcorner) = \begin{cases} C_D(\mathcal{V}_{\mathfrak{M}}(\rho(D\tau))) & si\ \rho(D\tau) \in dom(\mathcal{V}_{\mathfrak{M}}) \\ w & si\ \rho(D\tau) \notin dom(\mathcal{V}_{\mathfrak{M}}) \end{cases} \tag{4.63}$$

Supongamos que escogemos $\rho(Dx)$ de modo que $\rho(D\tau) \in dom(\mathcal{V}_{\mathfrak{M}})$, lo cual es siempre cierto en los lenguajes interpretados totales y puede conseguirse fácilmente en los parciales.[44] (4.63) quedará reducido a:

$$\mathcal{V}_{\mathfrak{M}}(D\tau) = \mathcal{V}_{\mathfrak{M}}(D\ulcorner\rho(D\tau)\urcorner) = C_D(\mathcal{V}_{\mathfrak{M}}(\rho(D\tau))) \tag{4.64}$$

[44] El caso más sencillo es que $\rho(D\tau)$ sea simplemente $D\tau$.

Por composicionalidad, existe una función, C_ρ, tal que

$$\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = C_\rho(\mathcal{V}_{\mathfrak{M}}(D\tau)) \tag{4.65}$$

Teniendo en cuenta (4.64) y (4.65):

$$\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = C_\rho(\mathcal{V}_{\mathfrak{M}}(D\tau)) = C_\rho(C_D(\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)))) \tag{4.66}$$

y, por tanto:

$$\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = (C_\rho \circ C_D)(\mathcal{V}_{\mathfrak{M}}(\rho(D\tau))) \tag{4.67}$$

que será una condición imposible en los casos en que la función compuesta$(C_\rho \circ C_D)$ no tenga ningún punto fijo.

(4.67) es la relación que buscábamos que iguala $\mathcal{V}_{\mathfrak{M}}(\psi)$ con una función de sí misma (ψ es la oración autorreferencial $\rho(D\tau)$). Sin embargo, a partir de (4.64) y (4.65) se obtiene un resultado más simple:

$$\mathcal{V}_{\mathfrak{M}}(D\tau) = C_D(C_\rho(\mathcal{V}_{\mathfrak{M}}(D\tau))) = (C_D \circ C_\rho)(\mathcal{V}_{\mathfrak{M}}(D\tau)) \tag{4.68}$$

que iguala $\mathcal{V}_{\mathfrak{M}}(D\tau)$ con una función de sí misma y será una condición imposible en los casos en que $(C_D \circ C_\rho)$ no tenga ningún punto fijo.

Afortunadamente, si sabemos que $(C_\rho \circ C_D)$ no tiene ningún punto fijo, podemos afirmar que $(C_D \circ C_\rho)$ tampoco lo tiene y viceversa. Es más, el número de puntos fijos de ambas funciones compuestas es forzosamente el mismo. Véase la proposición A.1 (p. 237) en el apéndice A y téngase en cuenta que tantoC_ρ como C_D son funciones totales definidas entre W y W.

4.2. El problema visto desde una perspectiva más general

4.2.1. Lenguajes interpretados enunciativos trivaluados

En primer lugar, trabajaremos con lenguajes interpretados enunciativos trivaluados de los que, los lenguajes interpretados utilizados hasta ahora son un caso particular. Un lenguaje interpretado enunciativo trivaluado, $(\mathcal{L}, \xi, \mathfrak{M}, E)_{etr}$, se obtiene fácilmente a partir de un lenguaje interpretado enunciativo parcial. La

principal diferencia es que cualquiera que sea la fórmula φ para la que la función $\mathcal{U}_{\mathfrak{M},s}$ no estaba definida, ahora sí lo estará, pero, en esos casos,$\mathcal{U}_{\mathfrak{M},s}(\varphi)$ no será un contenido enunciativo completo y ello supondrá que$\mathcal{U}_{\mathfrak{M},s}(\varphi)$ no será verdadero ni falso sino que tendrá un tercer valor de verdad, $\mathfrak{i}$, es decir, $E(\mathcal{U}_{\mathfrak{M},s}(\varphi)) = \mathfrak{i}$. Los cambios que hay que hacer en la definición de un lenguaje interpretado enunciativo parcial para definir un lenguaje interpretado enunciativo trivaluado son pequeños: a) el conjunto de valores de verdad será $W = \{\mathfrak{f}, \mathfrak{i}, \mathfrak{v}\}$ y la interpretación extensional de las constantes lógicas, C, será la de una lógica trivaluada; b) los elementos del universo enunciativo, $\mathfrak{E}$, no son únicamente contenidos enunciativos completos sino también lo que podríamos llamar contenidos enunciativos incompletos o defectuosos (los de aquellas oraciones cuyo valor de verdad es$\mathfrak{i}$); c) la función $\mathfrak{I}_{\mathfrak{M}}$ no está definida parcialmente sino totalmente sobre el conjunto de símbolos de oración; d) la función asociada al símbolo de relación R, $\mathfrak{I}_{\mathfrak{M}}(R)$, definida entre $\mid \mathfrak{M}_y \mid^n$ y $\mathfrak{E}$, no es parcial sino total (y lo mismo diremos de su extensionalización $E(\mathfrak{I}_{\mathfrak{M}}(R))$, definida entre $\mid \mathfrak{M}_x \mid^n$ y W).

Como ya vimos —al introducir los lenguajes interpretados trivaluados en el apartado 4.1.3.1—, partiendo de un lenguaje interpretado bivaluado parcial, es fácil conseguir un lenguaje interpretado trivaluado en que las oraciones que carecían de valor de verdad en el primero tomen el valor $\mathfrak{i}$ y el resto mantengan el mismo valor de verdad ($\mathfrak{v}$ o $\mathfrak{f}$). Para ello es necesario escoger una interpretación extensional de las constantes lógicas mediante operaciones en las que$\mathfrak{i}$ sea un elemento absorbente. En este sentido, los lenguajes interpretados enunciativos parciales corresponden a un caso particular de los lenguajes interpretados enunciativos trivaluados.

Una simplificación que podemos hacer es prescindir del operador$T$ que, como hemos visto, no es necesario para formalizar la paradoja.

Nuestros lenguajes interpretados extensionales trivaluados eran parciales para los términos. Si queremos que sean un caso particular de los lenguajes interpretados enunciativos trivaluados debemos permitir en estos que haya términos fuera del dominio de $\mathcal{V}_{\mathfrak{M},s}$. Para ello basta con que la función de extensionalización, E, no esté definida sobre todo elemento de $\mid \mathfrak{M}_y \mid$.[45] Como consecuencia, la función de extensionalización hay que redefinirla:

[45]Por tanto, no podremos afirmar que $dom(\mathcal{U}_{\mathfrak{M},s}) = dom(\mathcal{V}_{\mathfrak{M},s})$ ni que $dom(\mathcal{U}_{\mathfrak{M}}) = dom(\mathcal{V}_{\mathfrak{M}})$. Solo es seguro que $dom(\mathcal{V}_{\mathfrak{M},s}) \subseteq dom(\mathcal{U}_{\mathfrak{M},s})$ y $dom(\mathcal{V}_{\mathfrak{M}}) \subseteq dom(\mathcal{U}_{\mathfrak{M}})$.

Definición 4.24 (función de extensionalización). *La función de extensionalización, E, de un lenguaje interpretado enunciativo trivaluado,* $(\mathcal{L}, \xi, \mathfrak{M}, E)_{etr}$ *se caracteriza por lo siguiente:*

1. *E asocia a cada contenido enunciativo (elemento de* $\mathfrak{E}$*), un valor de verdad (elemento de* W*).*

2. *Para cada conectiva, k, se tiene* $E(\mathfrak{C}_k) = C_k$*, y para cada cuantificador, Q,* $E(\mathfrak{C}_Q) = C_Q$*.*

3. *E asocia a cada elemento de* $\mid \mathfrak{M}_y \mid$*, a lo sumo un elemento de* $\mid \mathfrak{M}_x \mid$ *y, en particular,* $E(\mathfrak{I}_{\mathfrak{M}}(\ulcorner \varepsilon \urcorner)) = \varepsilon$*, para cualquier expresión bien formada del lenguaje,* ε*.*

4. *Para todo elemento, o, de* $\mid \mathfrak{M}_x \mid$ *existe al menos un elemento, c, de* $\mid \mathfrak{M}_y \mid$ *tal que* $E(c) = o$

5. *Dado un símbolo de relación n-ádico, R, la función E asocia a cada función,* $\mathfrak{I}_{\mathfrak{M}}(R)$*, definida entre* $\mid \mathfrak{M}_y \mid^n$ *y* $\mathfrak{E}$*, una función total,* $E(\mathfrak{I}_{\mathfrak{M}}(R))$*, definida entre* $\mid \mathfrak{M}_x \mid^n$ *y* W*.*

6. *E asocia a cada función,* $\mathfrak{I}_{\mathfrak{M}}(f)$*, definida entre* $\mid \mathfrak{M}_y \mid^n$ *y* $\mid \mathfrak{M}_y \mid$*, una función,* $E(\mathfrak{I}_{\mathfrak{M}}(f))$*, definida entre* $\mid \mathfrak{M}_x \mid^n$ *y* $\mid \mathfrak{M}_x \mid$*.*

7. *Además, establecemos la composicionalidad de la función*E *para lo cual debe cumplirse:*

   a) *Si* R *es un símbolo de relación n-ádica y* $(\tau_1 ... \tau_n) \in \mid \mathfrak{M}_y \mid^n$*, entonces*

$$E(\mathfrak{I}_{\mathfrak{M}}(R)(\tau_1 ... \tau_n)) = \begin{cases} \mathfrak{i} & si\, existe\, j \,/\, 1 \leq j \leq n \ \wedge \tau_j \notin dom(E) \\ E(\mathfrak{I}_{\mathfrak{M}}(R))(E(\tau_1), ..., E(\tau_n)) & en\, otro\, caso \end{cases}$$

   b) *Si* $(\tau_1, \tau_2) \in \mid \mathfrak{M}_y \mid^2$ *y, en el metalenguaje, usamos el símbolo* $\equiv$ *para la identidad de elementos de* $\mid \mathfrak{M}_y \mid$ *o de* $\mathfrak{E}$ *y el símbolo* $=$ *para la identidad de elementos de* $\mid \mathfrak{M}_x \mid$ *o de* W*, entonces*

$$E(\tau_1 \equiv \tau_2) =_{def} \begin{cases} \mathfrak{i} & si\, \tau_1 \notin dom(E) \vee \tau_2 \notin dom(E), \\ & en\, otro\, caso: \\ \mathfrak{v} & si\, E(\tau_1) = E(\tau_2) \\ \mathfrak{f} & si\, E(\tau_1) \neq E(\tau_2) \end{cases}$$

c) *Si f es un símbolo de función n-ádica y $(\tau_1...\tau_n) \in \mid \mathfrak{M}_y \mid^n$, entonces $(\mathfrak{I}_{\mathfrak{M}}(f)(\tau_1...\tau_n)) \in dom(E)$ si y solo si $\tau_1 \in dom(E), ..., \tau_n \in dom(E)$. Cuando $(\mathfrak{I}_{\mathfrak{M}}(f)(\tau_1...\tau_n)) \in dom(E)$:*

$$E(\mathfrak{I}_{\mathfrak{M}}(f)(\tau_1...\tau_n)) =_{def} E(\mathfrak{I}_{\mathfrak{M}}(f))(E(\tau_1), ..., E(\tau_n))$$

El concepto de asignación de variables es el mismo que en los lenguajes interpretados enunciativos parciales. Las funciones $\mathcal{V}_{\mathfrak{M},s}$ y $\mathcal{V}_{\mathfrak{M}}$ se siguen definiendo respectivamente como $(E \circ \mathcal{U}_{\mathfrak{M},s})$ y $(E \circ \mathcal{U}_{\mathfrak{M}})$.

La función $\mathcal{U}_{\mathfrak{M},s}$ se define igual que en la definición 4.21 (p. 109) excepto en los casos en que su argumento es una oración cuantificada donde la definición es:

$$\begin{aligned} &\mathcal{U}_{\mathfrak{M},s}(\forall x\, \varphi) \equiv_{def} \mathfrak{C}_{\forall}(\{\mathcal{U}_{\mathfrak{M},s[e/x]}(\varphi)/e \in \mid \mathfrak{M}_y \mid \cap\, dom(E)\}) \\ &\mathcal{U}_{\mathfrak{M},s}(\exists x\, \varphi) \equiv_{def} \mathfrak{C}_{\exists}(\{\mathcal{U}_{\mathfrak{M},s[e/x]}(\varphi)/e \in \mid \mathfrak{M}_y \mid \cap\, dom(E)\}) \end{aligned} \tag{4.69}$$

La única diferencia es que se ha cambiado $\mid \mathfrak{M}_y \mid$ por $\mid \mathfrak{M}_y \mid \cap\, dom(E)$. Las nuevas definiciones también son válidas para los lenguajes interpretados enunciativos parciales, pues en ellos $\mid \mathfrak{M}_y \mid \cap\, dom(E) = \mid \mathfrak{M}_y \mid$ debido a que todo elemento de $\mid \mathfrak{M}_y \mid$ estaba en el dominio de E, es decir, $\mid \mathfrak{M}_y \mid \subset\ dom(E)$.

Gracias a (*4.69*), $\mathcal{V}_{\mathfrak{M},s}(\forall x\, \varphi)$ y $\mathcal{V}_{\mathfrak{M},s}(\exists x\, \varphi)$ proporcionan interpretaciones objetuales de los cuantificadores ya que:[46]

$$\begin{aligned} &\mathcal{V}_{\mathfrak{M},s}(\forall x\, \varphi) = E(\mathcal{U}_{\mathfrak{M},s}(\forall x\, \varphi)) = \\ &= E(\mathfrak{C}_{\forall}(\{\mathcal{U}_{\mathfrak{M},s[e/x]}(\varphi)/e \in \mid \mathfrak{M}_y \mid \cap\, dom(E)\})) \end{aligned} \tag{4.70}$$

y, teniendo en cuenta el punto $4b2$ (p. 107) de la definición 4.18:

$$\begin{aligned} &E(\mathfrak{C}_{\forall}(\{\mathcal{U}_{\mathfrak{M},s[e/x]}(\varphi)/e \in \mid \mathfrak{M}_y \mid \cap\, dom(E)\})) = \\ &= C_{\forall}(\{E(\mathcal{U}_{\mathfrak{M},s[e/x]}(\varphi))/e \in \mid \mathfrak{M}_y \mid \cap\, dom(E)\}) = \\ &= C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\varphi)/e \in \mid \mathfrak{M}_y \mid \cap\, dom(E)\}) \end{aligned} \tag{4.71}$$

Para que la interpretación del cuantificador sea objetual, es necesario y suficiente que $E(e)$ *recorra* exactamente todos los *objetos*, es decir, que se cumpla: 1/ dado un e cualquiera perteneciente a $\mid \mathfrak{M}_y \mid \cap\, dom(E)$, $e \in dom(E)$ y $E(e) \in \mid \mathfrak{M}_x \mid$;[47] 2/ dado, o, perteneciente a $\mid \mathfrak{M}_x \mid$ debe existir al menos un elemento, e, de

[46] El razonamiento para el cuantificador $\exists$ es igual que el que vemos a continuación para $\forall$.

[47] Cuando $e \notin dom(E)$, E no asocia a e ningún objeto y $\mathcal{V}_{\mathfrak{M},s[e/x]}(\varphi)$ no debe formar parte del conjunto que constituye el argumento de $C_{\forall}$ —en una interpretación objetual—. Por eso es exigible $e \in dom(E)$.

$\mid \mathfrak{M}_y \mid$ tal que $E(e) = o$. Ahora bien, estos dos requisitos están garantizados, respectivamente, por los puntos 3 (p. 122) y 4 de la definición de E. En cambio, si en lugar de $\mid \mathfrak{M}_y \mid \cap dom(E)$ hubiésemos dejado $\mid \mathfrak{M}_y \mid$ en (*4.69*), el requisito 1/ sería: dado un e cualquiera perteneciente a $\mid \mathfrak{M}_y \mid$, $e \in dom(E)$ y $E(e) \in \mid \mathfrak{M}_x \mid$. Mas, como, según el punto 3 de la definición de E, en un lenguaje enunciativo trivaluado, puede haber elementos de $\mid \mathfrak{M}_y \mid$ fuera del dominio de la función E, el requisito no se cumpliría.

Una vez que se ha definido la función $\mathcal{U}_{\mathfrak{M},s}$, el modo de definir la función $\mathcal{U}_{\mathfrak{M}}$ no sufre ningún cambio.

Una ventaja de los lenguajes interpretados enunciativos trivaluados sobre los enunciativos parciales es que se vuelve a recuperar la propiedad de que, si el lenguaje es fuertemente autorreferencial, para toda fórmula con una única variable libre, $\sigma(x)$, existe un término, τ, tal que

$$\mathcal{V}_{\mathfrak{M}}(\sigma(\tau)) = \mathcal{V}_{\mathfrak{M}}(\sigma(\ulcorner\sigma(\tau)\urcorner)) \tag{4.13}$$

En efecto, por ser el lenguaje fuertemente autorreferencial, habrá un término,τ, tal que

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\sigma(\tau)\urcorner) \tag{4.11}$$

Ahora bien, puesto que $\mathcal{V}_{\mathfrak{M}}$ es una función definida para toda oración —algo que no ocurre en un lenguaje interpretado parcial—, de (4.11) podemos deducir (4.13).

Conviene destacar que en los lenguajes interpretados enunciativos trivaluados se conserva la siguiente propiedad de los lenguajes interpretados trivaluados extensionales (ver definición 4.17, punto 7, p. 94):

Proposición 4.8. *Sean $t_1, t_2, ..., t_n$, términos y R, un símbolo de relación n-ádica. En un lenguaje interpretado enunciativo trivaluado, si alguno de los términos no está en el dominio de $\mathcal{V}_{\mathfrak{M},s}$ entonces $\mathcal{V}_{\mathfrak{M},s}(R\,t_1, t_2, ..., t_n) = \mathfrak{i}$*

Demostración. Teniendo en cuenta que, según las definiciones de $\mathcal{V}_{\mathfrak{M},s}$ y $\mathcal{U}_{\mathfrak{M},s}$:

$$\mathcal{V}_{\mathfrak{M},s}(R\,t_1, t_2, ..., t_n) = E(\mathcal{U}_{\mathfrak{M},s}(R\,t_1, t_2, ..., t_n)) \tag{4.72}$$

y

$$\mathcal{U}_{\mathfrak{M},s}(R\,t_1, t_2, ..., t_n) = \mathfrak{I}_{\mathfrak{M}}(R)(\mathcal{U}_{\mathfrak{M},s}(t_1), \mathcal{U}_{\mathfrak{M},s}(t_2), ..., \mathcal{U}_{\mathfrak{M},s}(t_n)) \tag{4.73}$$

resulta

$$\mathcal{V}_{\mathfrak{M},s}(R\,t_1,t_2,...,t_n) = E(\mathfrak{I}_{\mathfrak{M}}(R)(\mathcal{U}_{\mathfrak{M},s}(t_1),\mathcal{U}_{\mathfrak{M},s}(t_2),...,\mathcal{U}_{\mathfrak{M},s}(t_n))) \tag{4.74}$$

Por otra parte, en un lenguaje interpretado enunciativo trivaluado, por definición, todos los términos están en el dominio de $\mathcal{U}_{\mathfrak{M},s}$ pero no en el de $\mathcal{V}_{\mathfrak{M},s}$. Como $\mathcal{V}_{\mathfrak{M},s}$ es la función $E \circ \mathcal{U}_{\mathfrak{M},s}$, afirmar que $t \notin dom(\mathcal{V}_{\mathfrak{M},s})$ equivale a afirmar $\mathcal{U}_{\mathfrak{M},s}(t) \notin dom(E)$. Así pues, si algún t_j (con $1 \leq j \leq n$) no está en el dominio de $\mathcal{V}_{\mathfrak{M},s}$, tendremos $\mathcal{U}_{\mathfrak{M},s}(t_j) \notin dom(E)$ y, de acuerdo con la definición 4.24, punto *7a* (p. 122):

$$E(\mathfrak{I}_{\mathfrak{M}}(R)(\mathcal{U}_{\mathfrak{M},s}(t_1),\mathcal{U}_{\mathfrak{M},s}(t_2),...,\mathcal{U}_{\mathfrak{M},s}(t_n))) = \mathfrak{i} \tag{4.75}$$

Finalmente, de (4.74) y (4.75), se deduce:

$$\mathcal{V}_{\mathfrak{M},s}(R\,t_1,t_2,...,t_n) = \mathfrak{i} \tag{4.76}$$

■

Ya hemos visto que un lenguaje interpretado enunciativo parcial se puede considerar un caso particular de un lenguaje interpretado enunciativo trivaluado. También es claro que un lenguaje interpretado extensional trivaluado, parcial para los términos, es un caso particular de un lenguaje interpretado enunciativo trivaluado como el que hemos terminado de esbozar en el párrafo anterior. Y, por supuesto, un lenguaje interpretado clásico (extensional, bivaluado) es un caso particular de los demás.

Debido a su mayor generalidad, en lo sucesivo, mientras no se indique lo contrario, utilizaremos lenguajes interpretados enunciativos trivaluados, tal como los hemos caracterizado en este apartado, para analizar la paradoja del mentiroso y otras oraciones autorreferenciales.

En este tipo de lenguajes podemos definir las nociones de consecuencia y de equivalencia lógicas del mismo modo que lo hicimos en los lenguajes interpretados enunciativos parciales. Sin embargo, en lenguajes interpretados con más de dos valores de verdad conviene añadir la noción de fórmulas fuertemente equivalentes.

Definición 4.25 (equivalencia lógica fuerte). *Fijados un lenguaje $\mathcal{L}$ y una matriz ξ (propios de lenguajes interpretados enunciativos trivaluados), la fórmula φ es fuertemente equivalente a la fórmula ψ ssi para cualquier función de exten-*

sionalización, cualquier modelo $\mathfrak{M}$ *y cualquier asignación de variables,* s *(propios de ese tipo de lenguajes interpretados),* $\mathcal{V}_{\mathfrak{M},s}(\varphi) = \mathcal{V}_{\mathfrak{M},s}(\psi)$.

Es claro que dos fórmulas fuertemente equivalentes también serán lógicamente equivalentes pero no necesariamente a la inversa. Sin embargo, en un lenguaje interpretado con una lógica bivaluada clásica no hay distinción entre equivalencia lógica y equivalencia fuerte.

4.2.2. Autorreferencia y desentrecomillado generalizados

4.2.2.1. Términos paradójicos

Hasta ahora, una de las principales conclusiones es que el intento de formalizar una oración del mentiroso muestra una incompatibilidad entre la autorreferencialidad fuerte y la existencia de un predicado, D, que: a) para cualquier oración φ, cumpla $D^{\ulcorner}\varphi^{\urcorner} \in dom(\mathcal{V}_{\mathfrak{M}})$; b) realice una función de desentrecomillado. La condición a) está garantizada en un lenguaje interpretado enunciativo trivaluado por lo que no será necesario explicitarla. La condición b) resalta el hecho de que lo incompatible con la autorreferencialidad fuerte del lenguaje no es tanto la existencia de un predicado de verdad como de un predicado que realice una función de desentrecomillado. Esto resulta aún más patente al constatar que no es preciso que la función de desentrecomillado se aplique a nombres de oraciones para que surja la incompatibilidad, sino que esta también aparece cuando se aplica a nombres de términos. El razonamiento con términos es similar al realizado con fórmulas, por lo que podemos generalizar nuestras definiciones relativas a la autorreferencialidad a ambos tipos de expresiones.

Definición 4.26 (expresión autorreferencial). *Una expresión bien formada es autorreferencial ssi un término propio de ella tiene como referencia la expresión.*

Si llamamos $(\mathcal{L}, \xi, \mathfrak{M}, E)_{etr}$ al lenguaje interpretado enunciativo trivaluado, τ al término y $\varepsilon(\tau)$ a la expresión autorreferencial, tendremos

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \varepsilon(\tau) \tag{4.77}$$

o, dicho de otros modos:

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\varepsilon(\tau)^{\urcorner}) \tag{4.78}$$

$$\mathfrak{M} \models \tau = {}^{\ulcorner}\varepsilon(\tau)^{\urcorner} \tag{4.79}$$

Definición 4.27 (lenguaje fuertemente autorreferencial). *Un lenguaje interpretado es fuertemente autorreferencial ssi para toda expresión formal con una única variable libre, $\varepsilon(x)$, existe un término, τ, cuya referencia es la expresión formal $\varepsilon(\tau)$.*

Definición 4.28 (función diagonalización). *d representa una función de diagonalización —en el lenguaje interpretado enunciativo trivaluado,$(\mathcal{L}, \xi, \mathfrak{M}, E)_{etr}$— ssi para toda expresión formal, $\varepsilon(x)$, con una única variable libre x, se cumple*

$$\mathcal{V}_{\mathfrak{M}}(d^{\ulcorner}\varepsilon(x)^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\varepsilon(^{\ulcorner}\varepsilon(x)^{\urcorner})^{\urcorner}) \tag{4.80}$$

Nótese que, (4.80) equivale a

$$\mathcal{V}_{\mathfrak{M}}(d^{\ulcorner}\varepsilon(x)^{\urcorner}) = \varepsilon(^{\ulcorner}\varepsilon(x)^{\urcorner}) \tag{4.81}$$

y también a:

$$\mathfrak{M} \models d^{\ulcorner}\varepsilon(x)^{\urcorner} = {}^{\ulcorner}\varepsilon(^{\ulcorner}\varepsilon(x)^{\urcorner})^{\urcorner} \tag{4.82}$$

Proposición 4.9. *Si d representa una función de diagonalización en el lenguaje interpretado enunciativo trivaluado, $(\mathcal{L}, \xi, \mathfrak{M}, E)_{etr}$, el lenguaje es fuertemente autorreferencial.*

Demostración. La misma que la de la proposición 4.2 (p. 85) cambiando la fórmula σ por la expresión formal ε. ■

La tarea de desentrecomillado de un término citado, τ, consiste en que dado un término ν tal que $\mathcal{V}_{\mathfrak{M}}(\nu) = \mathcal{V}_{\mathfrak{M}}(^{\ulcorner}\tau^{\urcorner})$, se cumpla:

$$\mathcal{V}_{\mathfrak{M}}(D\nu) = \begin{cases} C_D(\mathcal{V}_{\mathfrak{M}}(\tau)) & si\, \tau \in dom(\mathcal{V}_{\mathfrak{M}}) \\ x & si\, \tau \notin dom(\mathcal{V}_{\mathfrak{M}}) \end{cases} \tag{4.83}$$

donde x es un elemento cualquiera de $\mid \mathfrak{M}_x \mid$ y C_D, una función total definida entre $\mid \mathfrak{M}_x \mid$ y $\mid \mathfrak{M}_x \mid$. Además, queda de manifiesto que D ha de ser un símbolo de función, puesto que debe cumplirse $\mathcal{V}_{\mathfrak{M}}(D\nu) \in \mid \mathfrak{M}_x \mid$, mientras que, en el desentrecomillado de fórmulas, D era un símbolo de predicado.

Hecha esta salvedad, la incompatibilidad entre la autorreferencialidad fuerte del lenguaje y la existencia de una función de desentrecomillado, D, se pone de manifiesto en un razonamiento prácticamente idéntico al que vimos, cambiando

función D por predicado D y término por fórmula. En efecto, tomemos un término de la forma $t(Dx)$.[48] Si el lenguaje es fuertemente autorreferencial, existe un término, τ, cuya referencia es $t(D\tau)$, es decir:

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner t(D\tau) \urcorner) \tag{4.84}$$

Y, según la caracterización (4.83) de D:

$$\mathcal{V}_{\mathfrak{M}}(D\tau) = \mathcal{V}_{\mathfrak{M}}(D\ulcorner t(D\tau) \urcorner) = \begin{cases} C_D(\mathcal{V}_{\mathfrak{M}}(t(D\tau))) & si\, t(D\tau) \in dom(\mathcal{V}_{\mathfrak{M}}) \\ x & si\, t(D\tau) \notin dom(\mathcal{V}_{\mathfrak{M}}) \end{cases} \tag{4.85}$$

Supongamos que escogemos $t(Dx)$ de modo que $t(D\tau) \in dom(\mathcal{V}_{\mathfrak{M}})$.[49] Entonces (4.85) quedará reducido a:

$$\mathcal{V}_{\mathfrak{M}}(D\tau) = \mathcal{V}_{\mathfrak{M}}(D\ulcorner t(D\tau) \urcorner) = C_D(\mathcal{V}_{\mathfrak{M}}(t(D\tau))) \tag{4.86}$$

Por composicionalidad, existe una función, C_t, tal que

$$\mathcal{V}_{\mathfrak{M}}(t(D\tau)) = C_t(\mathcal{V}_{\mathfrak{M}}(D\tau)) \tag{4.87}$$

Teniendo en cuenta (4.86) y (4.87):

$$\mathcal{V}_{\mathfrak{M}}(t(D\tau)) = C_t(\mathcal{V}_{\mathfrak{M}}(D\tau)) = C_t(C_D(\mathcal{V}_{\mathfrak{M}}(t(D\tau)))) \tag{4.88}$$

y, por tanto:

$$\mathcal{V}_{\mathfrak{M}}(t(D\tau)) = (C_t \circ C_D)(\mathcal{V}_{\mathfrak{M}}(t(D\tau))) \tag{4.89}$$

que será una condición imposible en los casos en que la función compuesta$(C_t \circ C_D)$ no tenga ningún punto fijo.

A partir de (4.86) y (4.87) también se obtiene:

$$\mathcal{V}_{\mathfrak{M}}(D\tau) = C_D(C_t(\mathcal{V}_{\mathfrak{M}}(D\tau))) = (C_D \circ C_t)(\mathcal{V}_{\mathfrak{M}}(D\tau)) \tag{4.90}$$

[48] Decimos que un término t es de la forma $t(Dx)$ cuando contiene una variable, x, que en todas sus apariciones va precedida del símbolo de función D. Al término cerrado resultante de sustituir x, en todas sus apariciones dentro de $t(Dx)$, por un término cerrado, τ, lo designamos mediante $t(D\tau)$.

[49] El caso más sencillo es que $t(D\tau)$ sea simplemente $D\tau$.

que iguala $\mathcal{V}_{\mathfrak{M}}(D\tau)$ con una función de sí misma y será una condición imposible en los casos en que $(C_D \circ C_t)$ no tenga ningún punto fijo.

Como resultado importante, hemos ubicado el problema de la paradoja del mentiroso en el marco de un problema más general: el de la incompatibilidad, en un sistema composicional, entre la autorreferencialidad fuerte del lenguaje y la existencia de un predicado o función de desentrecomillado. Además queda de manifiesto que el razonamiento que muestra esa incompatibilidad es independiente de si usamos un lenguaje enunciativo trivaluado o un lenguaje trivaluado extensional.

4.2.2.2. Tarskificación

Para que una oración diga de sí misma únicamente que no es verdadera no es necesario que sea de la forma $\neg T\tau$ y que el término τ tenga como referencia la oración. Un modo distinto de conseguir autorreferencia es mediante la operación que Smullyan (1996, pp. 88-89) llama *Tarskificación*. Nosotros adaptaremos la idea de Smullyan a nuestro tipo de lenguajes interpretados y estableceremos la siguiente definición.

Definición 4.29 (función tarskificación). *El símbolo de función $f_{_T}$ representa una función de* tarskificación *ssi para toda fórmula, φ, se cumple:*

$$\mathcal{V}_{\mathfrak{M}}(f_{_T}{}^{\ulcorner}\varphi^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}({}^{\ulcorner}\forall x(x = {}^{\ulcorner}\varphi^{\urcorner} \rightarrow \varphi)^{\urcorner}) \tag{4.91}$$

Nótese que, (4.91) equivale a

$$\mathfrak{M} \models f_{_T}{}^{\ulcorner}\varphi^{\urcorner} = {}^{\ulcorner}\forall x(x = {}^{\ulcorner}\varphi^{\urcorner} \rightarrow \varphi)^{\urcorner} \tag{4.92}$$

Consideremos la fórmula Jx (J es un símbolo de predicado y x una variable). Su tarskificación es la oración $\forall x(x = {}^{\ulcorner}Jx^{\urcorner} \rightarrow Jx)$. Si interpretamos $J...$ como "Juan escribe ...", la interpretación de $\forall x(x = {}^{\ulcorner}Jx^{\urcorner} \rightarrow Jx)$ será: Juan escribe (algo idéntico a) la fórmula Jx. Para ver cómo se consigue la autorreferencia consideremos la fórmula

$$Jf_{_T}x \tag{4.93}$$

Su tarskificación es

$$\forall x(x = {}^{\ulcorner}Jf_{_T}x^{\urcorner} \rightarrow Jf_{_T}x) \tag{4.94}$$

cuya interpretación será: Juan escribe la tarskificación de (algo idéntico a) la fórmula $Jf_{T}x$. Pero la tarskificación de la fórmula $Jf_{T}x$ es la propia fórmula (4.94). Así pues la interpretación de (4.94) es que Juan escribe (4.94), lo que pone de manifiesto la autorreferencialidad de dicha oración.

En general, cuando φ es una fórmula de la forma $Pf_{T}x$ (P es un símbolo de predicado y x una variable), su tarskificación es la oración $\forall x(x = \ulcorner Pf_{T}x \urcorner \rightarrow Pf_{T}x)$. En una lógica en la que $\forall x(x = \ulcorner Pf_{T}x \urcorner \rightarrow Pf_{T}x)$ es fuertemente equivalente a $Pf_{T}\ulcorner Pf_{T}x \urcorner$, llegamos a la conclusión de que la tarskificación de $Pf_{T}x$ equivale fuertemente a una oración ($Pf_{T}\ulcorner Pf_{T}x \urcorner$) que puede leerse: "la tarskificación de $Pf_{T}x$ tiene la propiedad P". Entonces, en un sentido extensional, la tarskificación de $Pf_{T}x$ es una oración que dice de sí misma que tiene la propiedad P. Es pues una oración autorreferencial aunque literalmente no se ajuste a la definición 4.12 (p. 82). Este pequeño desajuste se resuelve si sustituimos dicha definición por una nueva definición más general de oración autorreferencial.

Definición 4.30 (oración autorreferencial). *Una oración ψ, es autorreferencial ssi: a) un término propio de ella tiene como referencia una oración fuertemente equivalente a ψ; o b) ψ, o una subfórmula de ψ, es de la forma $\forall x\varphi(x)$ o $\exists x\varphi(x)$ —x es una variable y $\varphi(x)$ una fórmula con la variable libre x—.*

De momento, solo nos interesa el caso *a* de la definición. En el apartado 5.3.1 justificaremos por qué, en nuestros lenguajes formalizados, las oraciones de las que $\forall x\varphi(x)$ o $\exists x\varphi(x)$ son una sub-oración deben considerarse también autorreferenciales.

En el ejemplo que nos ocupa, ψ es la oración $Pf_{T}\ulcorner Pf_{T}x \urcorner$ que tiene un término propio, $f_{T}\ulcorner Pf_{T}x \urcorner$ cuya referencia es la oración $\forall x(x = \ulcorner Pf_{T}x \urcorner \rightarrow Pf_{T}x)$.

Es sabido que

$$\forall x(x = \ulcorner Pf_{T}x \urcorner \rightarrow Pf_{T}x)$$

es fuertemente equivalente a

$$Pf_{T}\ulcorner Pf_{T}x \urcorner$$

en un lenguaje interpretado clásico (con capacidad de cita). Sin embargo, en caso de un lenguaje interpretado trivaluado, la equivalencia anterior depende de la interpretación extensional de la conectiva $\rightarrow$ y el cuantificador $\forall$, es decir, de las funciones $C_{\rightarrow}$ y $C_{\forall}$. Demostraremos que si interpretamos esas conectivas según la lógica trivaluada fuerte de Kleene, se produce la equivalencia. En esta interpretación

$C_\forall$ es la función *mín,* suponiendo que los valores de verdad están ordenados del siguiente modo: $\mathfrak{f} \prec \mathfrak{i} \prec \mathfrak{v}$. En cuanto a $C_\rightarrow$ se caracteriza por cumplir: $C_\rightarrow(\mathfrak{i}, \mathfrak{v}) = \mathfrak{v}$, $C_\rightarrow(\mathfrak{i}, \mathfrak{f}) = \mathfrak{i}$, $C_\rightarrow(\mathfrak{i}, \mathfrak{i}) = \mathfrak{i}$ y *para todo* $x \in W : ((C_\rightarrow(\mathfrak{f}, x) = \mathfrak{v}) \wedge (C_\rightarrow(\mathfrak{v}, x) = x))$.

Proposición 4.10. *Sean:* $\varsigma(x)$ *una fórmula con una única variable libre;* τ *un término en el dominio de* $\mathcal{V}_\mathfrak{M}$. *Para un lenguaje* $\mathcal{L}$ *y una matriz* ξ *en que el cuantificador* $\forall$ *y la conectiva* $\rightarrow$ *se interpretan extensionalmente según la lógica trivaluada fuerte de Kleene, la oración* $\forall x(x = \tau \rightarrow \varsigma(x))$ *es fuertemente equivalente a* $\varsigma(\tau)$.

Demostración. Fijado $(\mathcal{L}, \xi)$, tomemos una función de extensionalización, E, un modelo $\mathfrak{M}$ y una asignación de variables, s, cualesquiera.

Considerando las propiedades definitorias de los lenguajes interpretados enunciativos, incluidas las de función $\mathcal{U}_{\mathfrak{M},s}$ y función, $\mathcal{V}_{\mathfrak{M},s}$, tendremos:[50]

$$\begin{aligned}
&\mathcal{V}_{\mathfrak{M},s}(\forall x(x = \tau \rightarrow \varsigma(x))) = E(\mathcal{U}_{\mathfrak{M},s}(\forall x(x = \tau \rightarrow \varsigma(x)))) = \\
&= E(\mathfrak{C}_\forall(\{\mathcal{U}_{\mathfrak{M},s[e/x]}(x = \tau \rightarrow \varsigma(x))/e \in |\, \mathfrak{M}_y \,| \cap dom(E)\})) = \\
&= C_\forall(\{E(\mathcal{U}_{\mathfrak{M},s[e/x]}(x = \tau \rightarrow \varsigma(x)))/e \in |\, \mathfrak{M}_y \,| \cap dom(E)\}) = \\
&= C_\forall(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(x = \tau \rightarrow \varsigma(x))/e \in |\, \mathfrak{M}_y \,| \cap dom(E)\}\})
\end{aligned} \tag{4.95}$$

de donde:

$$\begin{aligned}
&\mathcal{V}_{\mathfrak{M},s}(\forall x(x = \tau \rightarrow \varsigma(x))) = \\
&= C_\forall(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(x = \tau \rightarrow \varsigma(x))/e \in |\, \mathfrak{M}_y \,| \cap dom(E)\})
\end{aligned} \tag{4.96}$$

Ahora bien:

$$\begin{aligned}
&\mathcal{V}_{\mathfrak{M},s[e/x]}(x = \tau \rightarrow \varsigma(x)) = \\
&= C_\rightarrow(\mathcal{V}_{\mathfrak{M},s[e/x]}(x = \tau), \mathcal{V}_{\mathfrak{M},s[e/x]}(\varsigma(x)))
\end{aligned} \tag{4.97}$$

Cuando $E(e) = \mathcal{V}_\mathfrak{M}(\tau)$:

$$\mathcal{V}_{\mathfrak{M},s[e/x]}(x = \tau) = \mathfrak{v} \tag{4.98}$$

y

$$\mathcal{V}_{\mathfrak{M},s[e/x]}(\varsigma(x)) = \mathcal{V}_{\mathfrak{M},s[e/x]}(\varsigma(\tau)) = \mathcal{V}_{\mathfrak{M},s}(\varsigma(\tau)) \tag{4.99}$$

[50] Por brevedad, no especificaré detalladamente cada una de las propiedades utilizadas en cada paso de la presente demostración.

Si además tenemos en cuenta que $C_{\rightarrow}$ corresponde a la lógica trivaluada fuerte de Kleene:

$$\begin{aligned} &C_{\rightarrow}(\mathcal{V}_{\mathfrak{M},s[e/x]}(x=\tau),\mathcal{V}_{\mathfrak{M},s[e/x]}(\varsigma(x))) = \\ &= C_{\rightarrow}(\mathfrak{v},\mathcal{V}_{\mathfrak{M},s[e/x]}(\varsigma(x))) = \mathcal{V}_{\mathfrak{M},s[e/x]}(\varsigma(x)) = \mathcal{V}_{\mathfrak{M},s}(\varsigma(\tau)) \end{aligned} \tag{4.100}$$

Cuando $E(e) \neq \mathcal{V}_{\mathfrak{M}}(\tau)$: $\mathcal{V}_{\mathfrak{M},s[e/x]}(x=\tau) = \mathfrak{f}$. Por tanto:

$$\begin{aligned} &C_{\rightarrow}(\mathcal{V}_{\mathfrak{M},s[e/x]}(x=\tau),\mathcal{V}_{\mathfrak{M},s[e/x]}(\varsigma(x))) = \\ &= C_{\rightarrow}(\mathfrak{f},\mathcal{V}_{\mathfrak{M},s[e/x]}(\varsigma(x))) = \mathfrak{v} \end{aligned} \tag{4.101}$$

En definitiva, en el conjunto $\{\mathcal{V}_{\mathfrak{M},s[e/x]}(x=\tau\rightarrow\varsigma(x))/e \in |\,\mathfrak{M}_y\,| \cap dom(E)\}$ solo hay dos elementos: $\mathfrak{v}$ y $\mathcal{V}_{\mathfrak{M},s}(\varsigma(\tau))$. Como, además, la función $C_{\forall}$ es la función *mín* y $\mathfrak{v}$ es el *mayor* de los elementos de W, tendremos:

$$\begin{aligned} &C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(x=\tau\rightarrow\varsigma(x))/e \in |\,\mathfrak{M}_y\,| \cap dom(E)\}) = \\ &= C_{\forall}(\{\mathfrak{v},\mathcal{V}_{\mathfrak{M},s}(\varsigma(\tau))\}) = \mathcal{V}_{\mathfrak{M},s}(\varsigma(\tau)) \end{aligned} \tag{4.102}$$

De (4.96) y (4.102) resulta:

$$\mathcal{V}_{\mathfrak{M},s}(\forall x(x=\tau\rightarrow\varsigma(x))) = \mathcal{V}_{\mathfrak{M},s}(\varsigma(\tau)) \tag{4.103}$$

Como conclusión (definición 4.25, p. 125), la oración $\forall x(x=\tau\rightarrow\varsigma(x))$ es fuertemente equivalente a $\varsigma(\tau)$. ■

Bajo las premisas de la proposición anterior, podemos afirmar que la oración $\forall x(x = \ulcorner Pf_{T}x\urcorner \rightarrow Pf_{T}x)$ es fuertemente equivalente a $Pf_{T}\ulcorner Pf_{T}x\urcorner$ (aquí la fórmula $\varsigma(x)$ de dicha proposición es $Pf_{T}x$, y el término τ es $\ulcorner Pf_{T}x\urcorner$).

La tarskificación deja claro que no es necesario que el lenguaje sea fuertemente autorreferencial para que una oración afirme algo de sí misma. Basta con que el lenguaje sea autorreferencial en el sentido siguiente:

Definición 4.31 (lenguaje autorreferencial). *Un lenguaje interpretado con capacidad de cita es autorreferencial ssi para toda fórmula con una única variable libre, $\sigma(x)$, existen un término ν y una oración φ tales que $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$ y φ es fuertemente equivalente a $\sigma(\nu)$.*

Una consecuencia es que $\sigma(\nu)$ resulta ser una oración autorreferencial (v. definición 4.30, p. 130).

Otra consecuencia es que si un lenguaje interpretado con capacidad de cita es fuertemente autorreferencial también es autorreferencial. En efecto, al ser fuertemente autorreferencial, para toda fórmula con una única variable libre, $\sigma(x)$, existe un término, ν, tal que $\mathcal{V}_{\mathfrak{M}}(\nu) = \sigma(\nu)$ y, evidentemente, $\sigma(\nu)$ es fuertemente equivalente a $\sigma(\nu)$. Por ello, el lenguaje es autorreferencial.

La siguiente proposición muestra que, con una pequeña restricción, un lenguaje con capacidad de cita y tarskificación es autorreferencial.

Proposición 4.11. *Sea un lenguaje interpretado con capacidad de cita en el que se cumple: a) dada cualquier fórmula, $\varsigma(x)$, con una única variable libre, la oración $\forall x(x = \ulcorner\varsigma(x)\urcorner \rightarrow \varsigma(x))$ es fuertemente equivalente a $\varsigma(\ulcorner\varsigma(x)\urcorner)$ y b) f_{T} representa una función de tarskificación. Entonces ese lenguaje es autorreferencial.*

Demostración. Dada una fórmula con una única variable libre, $\sigma(x)$:

(1) $\forall x(x = \ulcorner\sigma(f_{T}x)\urcorner \rightarrow \sigma(f_{T}x))$ es fuertemente equivalente a $\sigma(f_{T}\ulcorner\sigma(f_{T}x)\urcorner)$
; hipótesis a) cuando la fórmula $\varsigma(x)$ es $\sigma(f_{T}x)$

(2) $\mathcal{V}_{\mathfrak{M}}(f_{T}\ulcorner\sigma(f_{T}x)\urcorner) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\forall x(x = \ulcorner\sigma(f_{T}x)\urcorner \rightarrow \sigma(f_{T}x))\urcorner)$
; definición de tarskificación (def. 4.29, p. 129)

Si llamamos ν al término $f_{T}\ulcorner\sigma(f_{T}x)\urcorner$ y φ a la fórmula $\forall x(x = \ulcorner\sigma(f_{T}x)\urcorner \rightarrow \sigma(f_{T}x))$, los asertos (1) y (2) pueden ser rescritos como:

(1') φ es fuertemente equivalente a $\sigma(\nu)$

(2') $\mathcal{V}_{\mathfrak{M}}(\nu) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\varphi\urcorner)$

Teniendo en cuenta que $\mathcal{V}_{\mathfrak{M}}(\ulcorner\varphi\urcorner) = \varphi$, queda de manifiesto que existen un término ν y una oración φ tales que $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$ y φ es fuertemente equivalente a $\sigma(\nu)$. Por tanto, el lenguaje es autorreferencial. ■

Obsérvese que si el cuantificador $\forall$ y la conectiva $\rightarrow$ se interpretan según la lógica trivaluada fuerte de Kleene, la proposición 4.10 (p. 131) nos garantiza que se cumple la condición a). Por tanto, este tipo de lenguajes interpretados será autorreferencial en cuanto posea una función de tarskificación.

Nuestra siguiente conclusión importante es que, como cabía presumir, no es preciso que un lenguaje sea fuertemente autorreferencial para que, en un sistema composicional, se produzca la incompatibilidad con la existencia de un predicado de desentrecomillado. Basta con que el lenguaje sea autorreferencial.

En efecto, tomemos una fórmula de la forma $\rho(Dx)$. Si el lenguaje es autorreferencial, existirán una oración ψ y un término ν tales que $\mathcal{V}_{\mathfrak{M}}(\nu) = \psi$ y ψ es fuertemente equivalente a $\rho(D\nu)$. Entonces:

$$\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\rho(D\nu)) \tag{4.104}$$

y, por la composicionalidad de $\mathcal{V}_{\mathfrak{M}}$, existirá una función C_ρ definida entre W y W, tal que $\mathcal{V}_{\mathfrak{M}}(\rho(D\nu)) = C_\rho(\mathcal{V}_{\mathfrak{M}}(D\nu))$. Si D es un predicado de desentrecomillado, $\mathcal{V}_{\mathfrak{M}}(D\nu) = C_D(\mathcal{V}_{\mathfrak{M}}(\psi))$. Como conclusión:

$$\mathcal{V}_{\mathfrak{M}}(\psi) = C_\rho(C_D(\mathcal{V}_{\mathfrak{M}}(\psi))) \tag{4.105}$$

que es imposible cuando la función $(C_\rho \circ C_D)$ no tiene ningún punto fijo.

Pero nuestro enfoque no solo nos dice que hay un problema cuando alguna de las funciones $(C_\rho \circ C_D)$ —o $(C_D \circ C_t)$, en el caso de los términos— no tienen punto fijo. También hay un problema cuando alguna tiene varios puntos fijos —como ocurre si intentamos formalizar la oración del veraz— e incluso, como veremos, cuando tienen un único punto fijo.

4.2.3. Enfoque general del problema

En el lenguaje natural existen predicados de desentrecomillado (como verdadero, falso) y oraciones autorreferenciales, y queremos que, dentro de lo posible, ambas características se conserven en el lenguaje formal. Por otra parte, toda oración enunciativa o bien es verdadera o bien es falsa o bien no es verdadera ni falsa,[51] es decir, en una formalización trivaluada debe tener un único valor del verdad perteneciente al conjunto $\{\mathfrak{f}, \mathfrak{i}, \mathfrak{v}\}$.

El problema que hemos visto es que si en un lenguaje autorreferencial, hubiese un predicado de desentrecomillado, D; a partir de una fórmula de la forma $\rho(Dx)$, se podría construir una oración, ψ_ρ, con un valor de verdad, $\mathcal{V}_{\mathfrak{M}}(\psi_\rho)$, que debería cumplir $\mathcal{V}_{\mathfrak{M}}(\psi_\rho) = (C_\rho \circ C_D)(\mathcal{V}_{\mathfrak{M}}(\psi_\rho))$. Pero eso es imposible si escogemos $\rho(Dx)$ de modo que $(C_\rho \circ C_D)$ no tenga ningún punto fijo. (En un lenguaje fuertemente

[51] Suponer que los portadores de verdad genuinos son los contenidos enunciativos no impide calificar a las oraciones (cuando podemos hacer abstracción del contexto en que se emiten) de verdaderas o falsas, si se entiende que llamamos verdadera/falsa a aquella oración con un contenido enunciativo verdadero/falso.

autorreferencial con desentrecomillado podemos encontrar un problema similar para los términos, es decir, habría términos cuya referencia debería cumplir una condición imposible).

Es muy importante destacar que el problema que acabamos de señalar se describe en los mismos términos independientemente de si usamos un lenguaje enunciativo o uno puramente extensional. En cualquier caso llegamos a la conclusión de que debería cumplirse $\mathcal{V}_{\mathfrak{M}}(\psi_\rho) = (C_\rho \circ C_D)(\mathcal{V}_{\mathfrak{M}}(\psi_\rho))$, lo cual es problemático cuando $(C_\rho \circ C_D)$ no tenga un único punto fijo.

Cuando $(C_\rho \circ C_D)$ tiene varios puntos fijos ya no hay problema técnico pues en un lenguaje interpretado se puede asignar a la oración ψ_ρ uno (y solo uno) de entre varios valores de verdad y hacerlo sin contradicción. Pero la situación sigue siendo problemática porque no hay ningún criterio que nos permita decir que un valor de verdad está mejor escogido que otro. Se trata de un dilema que ya aparece en un análisis informal de la oración del veraz ("esta oración es verdadera") porque admite sin contradicción que se considere verdadera y que se considere falsa pero no hay más motivo para una elección que para la otra.

¿Qué decir del caso en que $(C_\rho \circ C_D)$ tiene un único punto fijo? Tomemos un ejemplo: la oración "esta oración carece de contenido enunciativo (completo)". Con objeto de formalizarla en un lenguaje interpretado enunciativo trivaluado, utilizamos un predicado de existencia de contenido enunciativo, R, caracterizado porque dados una oración φ y un término ν tales que $\mathcal{V}_{\mathfrak{M}}(\nu) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\varphi\urcorner)$, se cumple:

$$\mathcal{U}_{\mathfrak{M}}(R\nu) \equiv (\mathcal{V}_{\mathfrak{M}}(\varphi) \in \{\mathfrak{f}, \mathfrak{v}\}) \tag{4.106}$$

(Suponemos que, en el metalenguaje, $\mathcal{V}_{\mathfrak{M}}(\varphi) \in \{\mathfrak{f}, \mathfrak{v}\}$ representa el contenido enunciativo de "la valuación de φ, en el lenguaje interpretado $(\mathcal{L}, \xi, \mathfrak{M}, E)_{etr}$, es $\mathfrak{f}$ o es $\mathfrak{v}$"). Por tanto, si $\mathcal{V}_{\mathfrak{M}}(\nu) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\varphi\urcorner)$, el predicado R cumplirá:

$$\mathcal{V}_{\mathfrak{M}}(R\nu) = C_R(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{4.107}$$

donde la función C_R viene dada por: $C_R(\mathfrak{f}) = C_R(\mathfrak{v}) = \mathfrak{v}$, $C_R(\mathfrak{i}) = \mathfrak{f}$.

Por otra parte, la formalización de "esta oración carece de contenido enunciativo (completo)" será una oración autorreferencial de la forma $\neg R\tau$ donde τ es un

término que tiene como referencia la propia oración, es decir:

$$\mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner \neg R\tau \urcorner) \tag{4.108}$$

Si tenemos en cuenta (4.107), resulta que:

$$\mathcal{V}_{\mathfrak{M}}(R\tau) = \mathcal{V}_{\mathfrak{M}}(R\ulcorner \neg R\tau \urcorner) = C_R(\mathcal{V}_{\mathfrak{M}}(\neg R\tau)) \tag{4.109}$$

Al considerar también que $\mathcal{V}_{\mathfrak{M}}(\neg R\tau) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(R\tau))$, se deduce:

$$\mathcal{V}_{\mathfrak{M}}(R\tau) = C_R(C_{\neg}(\mathcal{V}_{\mathfrak{M}}(R\tau)) = (C_R \circ C_{\neg})(\mathcal{V}_{\mathfrak{M}}(R\tau)) \tag{4.110}$$

Puesto que el grafo de la función[52] $C_{\neg}$ es $\{(\mathfrak{f}, \mathfrak{v}), (\mathfrak{v}, \mathfrak{f}), (\mathfrak{i}, \mathfrak{i})\}$, y el de C_R es $\{(\mathfrak{f}, \mathfrak{v}), (\mathfrak{v}, \mathfrak{v}), (\mathfrak{i}, \mathfrak{f})\}$, la función compuesta de $C_{\neg}$ y C_R, $(C_R \circ C_{\neg})$, tendrá como grafo $\{(\mathfrak{f}, \mathfrak{v}), (\mathfrak{v}, \mathfrak{v}), (\mathfrak{i}, \mathfrak{f})\}$ y, por tanto, solo tiene un punto fijo, el valor de verdad $\mathfrak{v}$. ¿Podemos entonces concluir que la oración $R\tau$ es verdadera y, que por tanto, $\neg R\tau$ es falsa? En esencia, hay dos planteamientos ante esta pregunta, los cuales conducen a respuestas distintas.

Un planteamiento lo podemos ejemplificar en Mackie (1973, p. 291) quien responde negativamente a la pregunta anterior o, más exactamente, a la pregunta similar de si la oración "this sentence, standarly construed, is indeterminate" es falsa.[53] Para él, la oración tiene un tipo de autorreferencia que la hace indeterminada: hay una propiedad derivativa —en este caso, "ser indeterminada"— que aparece en la oración como derivativa de sí misma. En casos de este tipo, un intento de desempaquetar el contenido de la oración nos daría una expansión infinita (es indeterminado que es indeterminado que es indeterminado que ...) lo que, según Mackie, indica una falta de contenido de la oración en cuestión y, por tanto, indica que la oración es indeterminada.[54] Además, se pregunta en qué consistiría que fuese falsa y señala que la única respuesta es que consistiría en que no es indeterminada porque es falsa. En resumen: "It's false because it's false because it's false because ...".[55] De esta forma nos dice que su supuesta falsedad no está bien justificada; de hecho, se justifica igual que la veracidad o la falsedad de la oración

[52] El grafo de una función h es el conjunto de pares $\{(x, y)/h(x) = y\}$.
[53] Mackie denomina indeterminadas a las oraciones que no son verdaderas ni falsas.
[54] Ibíd., p. 286.
[55] Ibíd., p. 291.

del veraz. En cambio, opina que la reflexión crítica que lleva a la conclusión de que la oración es indeterminada debe prevalecer.

Una intuición similar a la de Mackie, por la que las propiedades derivativas pueden dar lugar a oraciones sin contenido, parece inspirar a Kripke (1975) su noción de oración fundamentada. La idea es que para establecer que una oración de la forma $T^{\ulcorner}\varphi^{\urcorner}$ es verdadera o establecer que es falsa, es preciso establecer previamente que φ es verdadera o que es falsa, respectivamente. Por consiguiente, la verdad o falsedad de una oración donde aparece el predicado de verdad,T, es derivado, en última instancia, del valor de verdad de oraciones en que no aparece T. Si esa derivación no es posible, la oración es no fundamentada. Formalmente, una oración es fundamentada si es verdadera o es falsa en el lenguaje interpretado que corresponde al menor punto fijo de la función que toma una interpretación del predicado de verdad y la extiende (añadiendo a la extensión del predicado aquellos códigos de oraciones que se evalúan como verdaderas y a la antiextensión los de las que se evalúan como falsas). Una consecuencia importante de esta definición es que el que una oración sea o no fundamentada puede depender del modo en que se definan en el lenguaje interpretado las operaciones con valores de verdad que corresponden a las conectivas lógicas. Tomando un ejemplo de Kripke (1975),[56] usando la valuación fuerte de Kleene, la disyunción de "la nieve es blanca" con una oración del mentiroso será verdadera (porque $C_{\vee}(\mathfrak{v},\mathfrak{i}) = \mathfrak{v}$). Pero con una valuación en que $C_{\vee}(\mathfrak{v},\mathfrak{i}) = \mathfrak{i}$, la disyunción anterior constituiría una oración no fundamentada. No obstante, hay oraciones como la del veraz y la del mentiroso que siempre resultan no fundamentadas.

El segundo planteamiento es que aquellas oraciones a las que se puede asignar un único valor de verdad coherentemente, tienen ese valor de verdad, sean fundamentadas o no. Así lo defiende Gupta (1982) implícitamente en su crítica a Kripke (1975). Simplificando el ejemplo de Gupta (1982, ejemplo (3) en parte IV), consideremos que A dice únicamente:

(a1) todo lo que dice B es verdadero;

(a2) algo de lo que dice B no es verdadero

y B dice únicamente:

(b1) como máximo, una afirmación de A es verdadera

[56] Martin (1984, p. 64, nota al pie 17).

Gupta considera que el siguiente razonamiento es natural aunque la explicación de Kripke lo invalidaría: (a1) y (a2) no pueden ser ambas verdaderas porque se contradicen, luego, es verdad que, como máximo, una afirmación de A es verdadera, es decir, (b1) es verdadera. Como (b1) es lo único que dice B, (a1) es verdadera y (a2) es falsa. Estamos pues en un caso en que a las afirmaciones (a1), (a2) y (b1) —en el contexto del ejemplo— se les puede asignar un valor de verdad coherentemente. Se trata de un caso similar al de la oración "esta oración carece de contenido enunciativo (completo)" teniendo en cuenta que el siguiente razonamiento informal permite decir coherentemente que la oración es falsa: si la oración fuese verdadera carecería de contenido enunciativo (completo) y su valor de verdad sería $\mathfrak{i}$ y si careciera de contenido enunciativo, sería verdadera; por lo que ambos supuestos son descartables. Solo queda la posibilidad de que sea falsa, en cuyo caso, la oración tiene contenido enunciativo y, por ello, lo que dice es falso.

La similitud entre ambos ejemplos sugiere que, probablemente, las oraciones (a1), (a2) y (b1) del ejemplo anterior serían indeterminadas para Mackie. Por su parte, Gupta señala que, en la formalización de Kripke, las tres oraciones son no fundamentadas, es decir, no son verdaderas ni falsas en el lenguaje interpretado correspondiente al menor punto fijo. Sin embargo, lo que realmente ocurre es que, ese lenguaje interpretado no tiene suficiente capacidad expresiva para formalizar las oraciones (a1), (a2) y (b1) tal como se entienden en lenguaje natural. Porque en la interpretación de Kripke del predicado T, la oración $T\ulcorner\varphi\urcorner$ no es verdadera ni falsa si φ no es verdadera ni falsa, mientras que, en lenguaje natural, decir de una oración que es verdadera, cuando la oración no es verdadera ni falsa, es decir algo falso. Por eso, del hecho de que (a1) y (a2) no pueden ser ambas verdaderas deducimos, en lenguaje natural, que (b1) es verdadera; mientras que del hecho de que (a1) y (a2) no puedan estar ambas en la extensión de T no se deduce que (b1) tenga que estar en dicha extensión.

La formalización de Kripke resulta por tanto insuficiente para caracterizar la noción intuitiva de oración fundamentada. Él mismo reconoce que hay nociones, como la de ser fundamentada, que no son expresables dentro del lenguaje formal, sino solamente en el metalenguaje. Por tanto, la oración "esta oración es no fundamentada" no puede formalizarse según la propuesta de Kripke, aunque según las ideas de Mackie —dada su similitud con la oración "this sentence, standarly construed, is indeterminate"—, se trataría de una oración sin contenido y por ello, de una oración no verdadera ni falsa.

En resumen, tenemos dos planteamientos:

G) Aquellas oraciones a las que se puede asignar un único valor de verdad coherentemente, tienen ese valor de verdad (parece ser el planteamiento de Gupta).

MK) Una oración a la que se puede asignar un único valor de verdad coherentemente no puede ser verdadera ni puede ser falsa a menos que tenga un contenido enunciativo (completo). Planteamiento más parecido a los de Mackie y Kripke.

La opción MK requiere establecer un criterio para determinar, si dada una oración, tiene o no un contenido enunciativo. Las ideas de Mackie y Kripke no constituyen una explicación suficiente, en este sentido. Mackie (1973) no establece un criterio riguroso. Kripke (1975), sí lo establece pero usando lenguajes formales de insuficiente capacidad expresiva: la simple oración que Mackie usa como ejemplo no puede ser formalizada con la propuesta de Kripke.

No intentaremos en este trabajo desarrollar la opción MK ni dilucidar si una de las dos opciones debe ser descartada, aunque ambos esfuerzos son, en mi opinión, de gran interés. Por el contrario, mantendremos, en lo sucesivo, abiertas ambas posibilidades.

Como conclusión, el resultado de nuestro análisis consiste en que, partiendo del objetivo de explicar coherentemente cuál es el valor de verdad de oraciones como las del mentiroso o del veraz, llegamos al problema de encontrar y justificar posibles fallos en las definiciones de los predicados de desentrecomillado y/o en las relativas a la autorreferencialidad, de modo que, con ciertas modificaciones, la autorreferencialidad y el desentrecomillado puedan coexistir en un lenguaje interpretado y, a la vez, la capacidad de este para expresar contenidos enunciativos no se vea reducida. Dependiendo de si se elige la opción G o la MK, expuestas más arriba, el problema planteado por ciertas oraciones podrá resolverse de modos distintos.[57]

[57] Incluso, dentro de la opción MK, puede haber diversas alternativas. Es más, la opción G podría considerarse un caso particular de la MK: aquel en que se considere que toda oración a la que se puede asignar coherentemente un único valor de verdad tiene un contenido enunciativo cuando ese valor de verdad es $\mathfrak{v}$ o es $\mathfrak{f}$.

4.2.4. La perspectiva de los sistemas de ecuaciones

Para analizar si se puede asignar un único valor de verdad coherentemente a una oración, un medio eficaz es plantear un sistema de ecuaciones de valores de verdad. La idea es sencilla: tomemos, por ejemplo, las oraciones siguientes:

(1) (2) es verdadera

(2) (1) no es verdadera

Si, (suponiendo por sencillez, en este ejemplo, una formalización extensional en que los portadores de verdad son las oraciones) llamamos x e y a los valores de verdad de las oraciones (1) y (2), respectivamente, y C_T y $C_\neg$ son las funciones, entre W y W, que corresponden al predicado "verdadero" y al operador negación, el sistema de ecuaciones que habrían de satisfacer x e y sería:

$$x = C_T(y);\ y = C_\neg(C_T(x)) \tag{4.111}$$

El uso de un sistema de ecuaciones permitirá plantear de modo más claro el problema con una oración autorreferencial de la forma $\rho(D\tau)$, donde D es un predicado de desentrecomillado y el término τ tiene como referencia la oración completa, es decir, $\mathcal{V}_\mathfrak{M}(\tau) = \mathcal{V}_\mathfrak{M}(\ulcorner\rho(D\tau)\urcorner)$:

$\mathcal{V}_\mathfrak{M}(D\tau) = C_D(\mathcal{V}_\mathfrak{M}(\rho(D\tau)))$; $\mathcal{V}_\mathfrak{M}(\tau) = \mathcal{V}_\mathfrak{M}(\ulcorner\rho(D\tau)\urcorner)$ y def. de D
$\mathcal{V}_\mathfrak{M}(\rho(D\tau)) = C_\rho(\mathcal{V}_\mathfrak{M}(D\tau))$; composicionalidad

Si llamamos x e y a los valores de verdad $\mathcal{V}_\mathfrak{M}(D\tau)$ y $\mathcal{V}_\mathfrak{M}(\rho(D\tau))$, respectivamente, obtendremos el siguiente sistema de ecuaciones:

$$\left.\begin{array}{l} x = C_D(y) \\ y = C_\rho(x) \end{array}\right\} \tag{4.112}$$

que, al igual que un sistema de ecuaciones numérico podrá ser determinado (una única solución), indeterminado (más de una solución) o incompatible (ninguna solución).

En principio, es inaceptable que el sistema sea incompatible porque entonces no podríamos asignar un valor de verdad a la oración autorreferencial $\rho(D\tau)$. El problema cuando el sistema es indeterminado es que no hay ningún otro criterio

que nos permita decidir cuál de las soluciones es mejor, como ocurre con la oración del veraz. En caso de que el sistema sea determinado, tendremos una forma coherente de asignar valores de verdad a las oraciones correspondientes a las incógnitas (en el ejemplo anterior, las oraciones $D\tau$ y $\rho(D\tau)$). Aunque, como hemos visto en el apartado 4.2.3, puede discutirse que esos valores de verdad, que pueden asignarse, sean los auténticos valores de verdad de las oraciones.

A modo de ejemplo, analicemos desde la perspectiva de los sistemas de ecuaciones de valores de verdad la oración "esta oración carece de contenido enunciativo (completo)" que habíamos formalizado como $\neg R\tau$ donde la autorreferencia radica en que $\mathcal{V}_{\mathfrak{M}}(\tau) = \neg R\tau$. Aquí el predicado de desentrecomillado es R y C_ρ es $C_\neg$, por lo que, si llamamos x e y a los valores de verdad $\mathcal{V}_{\mathfrak{M}}(R\tau)$ y $\mathcal{V}_{\mathfrak{M}}(\neg R\tau)$ respectivamente, el sistema de ecuaciones por resolver será:

$$\left.\begin{array}{l} x = C_R(y) \\ y = C_\neg(x) \end{array}\right\} \tag{4.113}$$

Puesto que el grafo de la función $C_\neg$ es $\{(\mathfrak{f},\mathfrak{v}),(\mathfrak{v},\mathfrak{f}),(\mathfrak{i},\mathfrak{i})\}$, y el de la función C_R es $\{(\mathfrak{f},\mathfrak{v}),(\mathfrak{v},\mathfrak{v}),(\mathfrak{i},\mathfrak{f})\}$, la única solución al sistema de ecuaciones anterior es $x = \mathfrak{v}$, $y = \mathfrak{f}$, o expresada vectorialmente, $(x, y) = (\mathfrak{v},\mathfrak{f})$.

Llegados a este punto, para algunos el problema estará resuelto y el valor veritativo de $\neg R\tau$ será $\mathfrak{f}$. Para otros, faltará por ver si $\neg R\tau$ tiene o no contenido enunciativo y, solo en caso positivo, el valor veritativo será $\mathfrak{f}$. Mackie, por ejemplo, respondería negativamente, porque para él la oración carece de contenido. Entonces tendríamos que asignar a $\neg R\tau$ el valor veritativo $\mathfrak{i}$ y tendríamos que explicar cómo es que, siendo la oración sin contenido enunciativo (completo) y diciendo de sí misma que no tiene contenido enunciativo (completo), no resulta ser verdadera.

Un modo de estudiar si $\neg R\tau$ tiene contenido enunciativo (completo) consiste en trabajar en un lenguaje interpretado enunciativo con contenidos enunciativos en vez de hacerlo con valores de verdad.

Teniendo en cuenta que la identidad de contenidos enunciativos $\Phi_1 \equiv \Phi_2$ implica $E(\Phi_1) = E(\Phi_2)$, pero no a la inversa, las ecuaciones de valores de verdad serían derivadas de "ecuaciones de contenidos enunciativos". Mas el que una ecuación de valores de verdad tenga una única solución no implica que la ecuación de contenidos enunciativos de la que se deriva haya de tener también una solución.

En el caso que nos ocupa, el sistema de ecuaciones de valores de verdad (4.113) sería derivado del siguiente sistema de ecuaciones de contenidos enunciativos:

$$\left.\begin{aligned}\Phi_x &\equiv \mathfrak{C}_R(\Phi_y)\\ \Phi_y &\equiv \mathfrak{C}_\neg(\Phi_x)\end{aligned}\right\} \tag{4.114}$$

donde las incógnitas Φ_x y Φ_y corresponden a $\mathcal{U}_{\mathfrak{M}}(R\tau)$ y $\mathcal{U}_{\mathfrak{M}}(\neg R\tau)$ respectivamente. La derivación se obtendría por aplicación de la función de extensionalización, E, dado que: a) $\Phi_1 \equiv \Phi_2$ implica $E(\Phi_1) = E(\Phi_2)$; b) $E(\Phi_x) = x$, $E(\Phi_y) = y$, $E(\mathfrak{C}_\neg(\Phi_x)) = C_\neg(x)$ y $E(\mathfrak{C}_R(\Phi_y)) = C_R(y)$.[58]

No pretendemos, en este trabajo, profundizar en la caracterización y estudio de los sistemas de ecuaciones de contenidos enunciativos, sino, solamente, sugerir un posible enfoque intensional del problema.

Este enfoque permite dar cabida a la idea de que la oración $\neg R\tau$ carece de contenido enunciativo; bastaría con encontrar una justificación de que el sistema (4.114) no tiene solución — dado que una de las incógnitas, Φ_y, representa el contenido enunciativo, $\mathcal{U}_{\mathfrak{M}}(\neg R\tau)$, de la oración—. Conviene recalcar que el hecho de que el sistema derivado (4.113) tenga solución no implica que (4.114) haya de tenerla, porque el que dos valores de verdad sean iguales no garantiza que correspondan al mismo contenido enunciativo. En cambio, el que dos contenidos enunciativos sean iguales sí garantiza que tienen el mismo valor de verdad, por lo cual, si (4.114) tiene una solución, (4.113) también la tendrá —bastará con aplicar la función de extensionalización a la solución de (4.114) para obtener una solución de (4.113)—.

Naturalmente, no pretendemos afirmar que el problema de la oración "esta oración carece de contenido enunciativo" quede resuelto con la explicación esbozada. La paradoja es que si la oración carece de contenido enunciativo, entonces afirmar la oración es afirmar algo verdadero y por tanto con contenido enunciativo. O usando nuestra formalización: si $\neg R\tau$ carece de contenido enunciativo su valor veritativo debe ser $\mathfrak{i}$, pero en el sistema de ecuaciones de valores veritativos (4.113) la única solución posible para el valor veritativo de $\neg R\tau$ es $\mathfrak{f}$. El problema de cuál es el auténtico valor veritativo de la oración sigue vivo.

No obstante, esto no implica que este último planteamiento sea peor. La opción alternativa de considerar que aquellas oraciones a las que se puede asignar un único valor de verdad coherentemente, tienen ese valor de verdad, no está libre de

[58] Que $E(\Phi_x) = x$ resulta de $E(\Phi_x) = E(\mathcal{U}_{\mathfrak{M}}(R\tau)) = \mathcal{V}_{\mathfrak{M}}(R\tau) = x$ y, análogamente se obtiene $E(\Phi_y) = y$. Que $E(\mathfrak{C}_\neg(\Phi_x)) = C_\neg(x)$ resulta de $E(\mathfrak{C}_\neg(\Phi_x)) = C_\neg(E(\Phi_x)) = C_\neg(x)$ y, análogamente, se obtiene $E(\mathfrak{C}_R(\Phi_y)) = C_R(y)$ si llamamos C_R a $E(\mathfrak{C}_R)$.

problemas similares, porque no está libre de casos, como la oración del mentiroso reforzada, a los que, en principio, no se puede asignar ningún valor de verdad coherentemente.

5

En busca de una solución

5.1. Abordando el problema

5.1.1. Introducción

Una vez que hemos decidido centrar nuestra atención en la incompatibilidad entre la autorreferencialidad y el desentrecomillado surge la cuestión de si alguno de estos aspectos es más responsable que el otro de la paradoja del mentiroso y de los problemas emparentados con ella (como el de la oración del veraz, la paradoja de Curry, etc.). Sin embargo, tanto el uso en lenguaje natural de la autorreferencia y del predicado "verdadero" como el hecho de que en un lenguaje formal se puede tener autorreferencia o un predicado de verdad —aunque no ambas cosas a la vez— sin mayor problema, no sugieren, en principio, ninguna respuesta a la cuestión anterior. A pesar de ello, se ha dedicado más atención al predicado de verdad (prototipo de predicado de desentrecomillado) que a la autorreferencia. Varios motivos pueden explicar este fenómeno. Uno es que el concepto de referencia parece menos problemático que el de verdad, lo que sugeriría que este último resulte más propenso a ocultar algún error que explicase las paradojas semánticas. Otro es que en la formalización clásica de la aritmética, codificando las fórmulas mediante números de Gödel, la función diagonalización es representable y el predicado de verdad, no; lo cual dirige la atención hacia este predicado y plantea la cuestión de si, con ciertas modificaciones, podría también representarse.

Nosotros partiremos de nuestro análisis formal e indagaremos sin prejuicios sobre las dos posibilidades —modificar los predicados de desentrecomillado o modificar la autorreferencialidad— para evitar los problemas que plantean las para-

dojas. También será necesario comprobar si, más allá de la solución formal, esas modificaciones se pueden justificar.

Para comenzar a abordar el problema, consideremos una oración de la forma $\rho(D\tau)$, que supondremos autorreferencial, es decir, la referencia del término τ será una oración, ψ, fuertemente equivalente a $\rho(D\tau)$. Entonces deberá cumplirse: $\mathcal{V}_{\mathfrak{M}}(\tau) = \psi = \mathcal{V}_{\mathfrak{M}}(\ulcorner\psi\urcorner)$ y $\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\rho(D\tau))$ y, además:

$$\begin{aligned} &\mathcal{V}_{\mathfrak{M}}(D\tau) = C_D(\mathcal{V}_{\mathfrak{M}}(\psi)) && ;\ \mathcal{V}_{\mathfrak{M}}(\tau) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\psi\urcorner) \text{ y def. de } D \\ &\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = C_\rho(\mathcal{V}_{\mathfrak{M}}(D\tau)) && ;\ \text{composicionalidad} \end{aligned}$$

Si llamamos x e y a los valores de verdad $\mathcal{V}_{\mathfrak{M}}(D\tau)$ y $\mathcal{V}_{\mathfrak{M}}(\psi)$, respectivamente, y tenemos en cuenta que $\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\rho(D\tau))$, obtendremos que el sistema de ecuaciones que condiciona el valor de verdad de la oración ψ es:

$$\left.\begin{aligned} x &= C_D(y) \\ y &= C_\rho(x) \end{aligned}\right\} \tag{4.112}$$

Es trivial comprobar que el anterior sistema de ecuaciones equivale a:

$$\left.\begin{aligned} x &= C_D(y) \\ y &= C_\rho(C_D(y)) \end{aligned}\right\} \tag{5.1}$$

También es elemental ver que el problema que plantea este sistema de ecuaciones es equivalente al problema planteado por la ecuación

$$y = C_\rho(C_D(y)) \tag{5.2}$$

en el sentido de que cada solución de (5.1) determina una solución de (5.2) y viceversa.

Como conclusión, la ecuación (5.2) o cualquiera de los sistemas de ecuaciones (4.112) o (5.1) será determinado, indeterminado o incompatible si y solo si $(C_\rho \circ C_D)$ tiene un único punto fijo, varios puntos fijos o ningún punto fijo, respectivamente.

No podemos aceptar los casos donde $(C_\rho \circ C_D)$ no tiene un único punto fijo porque entonces no habría respuesta a la pregunta por el valor de verdad de las oraciones ψ y $D\tau$. Incluso, como hemos visto (apartado 4.2.3), en los casos donde $(C_\rho \circ C_D)$ tiene un único punto fijo, puede discutirse que la solución de la ecuación (5.2) proporcione el auténtico valor de verdad de la oración ψ.

En los próximos apartados, estudiaremos las posibilidades de modificar justificadamente los predicados de desentrecomillado o la autorreferencialidad para solucionar la situación que plantea un sistema de ecuaciones como (4.112).

5.1.2. Modificación de los predicados de desentrecomillado

Supongamos, para empezar, que mantenemos la autorreferencialidad del lenguaje interpretado, ¿cómo habría que modificar cualquier predicado de desentrecomillado D? Como ahora sabemos, en los casos en que una oración de la forma $\rho(D\nu)$ es fuertemente equivalente a una oración ψ, $\mathcal{V}_{\mathfrak{M}}(\nu) = \psi$ y $(C_\rho \circ C_D)$ no tiene ningún punto fijo, la aplicación de la propiedad con la que hemos definido D, $\mathcal{V}_{\mathfrak{M}}(D\nu) = C_D(\mathcal{V}_{\mathfrak{M}}(\psi))$, permite deducir una contradicción. Por tanto, la restricción mínima que podemos hacer a la definición de los predicados de desentrecomillado es cambiarla para esos casos. Tampoco es muy conveniente mantenerla en los casos cuya única diferencia es que $(C_\rho \circ C_D)$ tiene varios puntos fijos, porque el valor de verdad de la oración ψ no quedaría fijado. Incluso cuando $(C_\rho \circ C_D)$ tiene un único punto fijo, es discutible que ese punto fijo sea el auténtico valor de verdad de la oración ψ.

Llamemos —tentativamente— *no desentrecomillable* a cualquier oración, ψ, fuertemente equivalente a una oración de la forma $\rho(D\nu)$, donde $\mathcal{V}_{\mathfrak{M}}(\nu) = \psi$. Se trataría entonces de restringir la propiedad con la que hemos definido D, $\mathcal{V}_{\mathfrak{M}}(D\nu) = C_D(\mathcal{V}_{\mathfrak{M}}(\varphi))$ (cuando $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$), a los casos en que φ sea una oración *desentrecomillable*. Pero, teniendo en cuenta que, en los lenguajes interpretados enunciativos trivaluados, la función de valuación $\mathcal{V}_{\mathfrak{M}}$ está definida para toda oración, debe cumplirse $\mathcal{V}_{\mathfrak{M}}(D\nu) \in \{\mathfrak{f}, \mathfrak{i}, \mathfrak{v}\}$. Llamemos φ_D al valor $\mathcal{V}_{\mathfrak{M}}(D\nu)$ cuando φ sea no desentrecomillable; entonces la caracterización del predicado D sería: dados una oración φ y un término ν tales que $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$:

$$\mathcal{V}_{\mathfrak{M}}(D\nu) = \begin{cases} C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)) & \textit{si } \varphi \textit{ es desentrecomillable} \\ \varphi_D \in W & \textit{si } \varphi \textit{ no es desentrecomillable} \end{cases} \tag{5.3}$$

Para una oración no desentrecomillable φ, es decir, fuertemente equivalente a una oración de la forma $\rho(D\tau)$ donde $\mathcal{V}_{\mathfrak{M}}(\tau) = \varphi$, tendremos:

$\mathcal{V}_{\mathfrak{M}}(D\tau) = \varphi_D$; $\mathcal{V}_{\mathfrak{M}}(\tau) = \varphi$ y definición (5.3) de D
$\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = C_\rho(\mathcal{V}_{\mathfrak{M}}(D\tau))$; composicionalidad

Si llamamos x e y a los valores de verdad $\mathcal{V}_{\mathfrak{M}}(D\tau)$ y $\mathcal{V}_{\mathfrak{M}}(\rho(D\tau))$ respectivamente, se obtiene el siguiente sistema de ecuaciones de valores de verdad:

$$\left.\begin{array}{l} x = \varphi_D \\ y = C_\rho(x) \end{array}\right\} \tag{5.4}$$

Se trata de un sistema determinado cuya única solución es $x = \varphi_D$, $y = C_\rho(\varphi_D)$. Aunque esta solución es técnicamente correcta, no nos dice nada acerca de cuál debe ser el valor φ_D. Como en la paradoja del veraz, no tenemos ningún criterio que nos diga a qué elemento de W tiene que ser igual φ_D, y por tanto, no tenemos ningún criterio que nos diga cuál es el auténtico valor de verdad de $D\tau$.

Más importante que la crítica anterior, es el hecho de que el predicado de verdad, así como otros predicados de desentrecomillado lo que pretenden expresar es que, en su caso, la oración referenciada por su argumento tiene un valor de verdad que pertenece a un determinado subconjunto de W.[1] Es decir, dados una oración φ y un término ν tales que $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$, si D es un predicado de desentrecomillado, se debe cumplir

$$\mathcal{U}_{\mathfrak{M}}(D\nu) \equiv (\mathcal{V}_{\mathfrak{M}}(\varphi) \in A_D) \tag{5.5}$$

siendo $A_D \subset W$. Como consecuencia:

$$\mathcal{V}_{\mathfrak{M}}(D\nu) = \begin{cases} \mathfrak{v} & si\ \mathcal{V}_{\mathfrak{M}}(\varphi) \in A_D \\ \mathfrak{f} & si\ \mathcal{V}_{\mathfrak{M}}(\varphi) \notin A_D \end{cases} \tag{5.6}$$

Pero esto equivale a caracterizar D mediante $\mathcal{V}_{\mathfrak{M}}(D\nu) = C_D(\mathcal{V}_{\mathfrak{M}}(\varphi))$ ¡sin restricción alguna en cuanto al tipo de oración φ referenciada por ν!.[2] Consiguientemente, la restricción propuesta en (5.3) supondría una restricción inaceptable a la capacidad expresiva del lenguaje formal, lo que, como sabemos, es un error muy común, quizás el más común, en los intentos de solucionar las paradojas semánticas: no reflejar en el lenguaje formal algunas características del lenguaje natural relevantes para la aparición de las paradojas.

[1] Véase, por ejemplo la definición (4.106) del predicado R en p. 135.
[2] Basta con definir C_D de forma que si $x \in A_D$, $C_D(x) = \mathfrak{v}$ y si $x \notin A_D$, $C_D(x) = \mathfrak{f}$.

5.1.3. Modificación de la autorreferencia

Por una parte, acabamos de reafirmar que, dados una oración,φ, y un término, ν, tales que $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi$, un predicado de desentrecomillado, D, debe cumplir:

$$\mathcal{V}_{\mathfrak{M}}(D\nu) = C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)) \tag{5.7}$$

Por otra, consideremos la oración autorreferencial $\sigma(\tau)$ en la que el término τ tiene como referencia una oración ψ fuertemente equivalente a $\sigma(\tau)$. Tendremos $\mathcal{V}_{\mathfrak{M}}(\tau) = \psi$ y, debido a la equivalencia fuerte entre ψ y $\sigma(\tau)$, $\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\sigma(\tau))$. Además:

$\mathcal{V}_{\mathfrak{M}}(D\tau) = C_D(\mathcal{V}_{\mathfrak{M}}(\psi))$; $\mathcal{V}_{\mathfrak{M}}(\tau) = \psi$ y definición de D
$\mathcal{V}_{\mathfrak{M}}(\sigma(\tau)) = C_\sigma(\mathcal{V}_{\mathfrak{M}}(\tau))$; composicionalidad

Llamando x e y a los valores de verdad $\mathcal{V}_{\mathfrak{M}}(D\tau)$ y $\mathcal{V}_{\mathfrak{M}}(\psi)$ respectivamente, y teniendo en cuenta que $\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\sigma(\tau))$, el sistema de ecuaciones correspondiente es:

$$\left.\begin{aligned} x &= C_D(y) \\ y &= C_\sigma(\psi) \end{aligned}\right\} \tag{5.8}$$

Obsérvese que C_σ es una función entre el universo extensional,$|\ \mathfrak{M}_x\ |$, y el conjunto de valores de verdad, W. Pero lo más importante es constatar que si $C_\sigma(\psi)$ es independiente de $\mathcal{V}_{\mathfrak{M}}(D\tau)$, y de $\mathcal{V}_{\mathfrak{M}}(\psi)$ (es decir, de x e y), el sistema (5.8) tiene una única solución: $y = C_\sigma(\psi)$, $x = C_D(C_\sigma(\psi))$.

Por eso, la oración autorreferencial $\sigma(\tau)$ no crea problemas si en ella no aparece ningún predicado de desentrecomillado, puesto que entonces, al afirmar la oración autorreferencial $\sigma(\tau)$ se afirma algo sobre los aspectos formales de la oración (por ejemplo, cuántos símbolos tiene, si es demostrable en un sistema formal...) pero no se dice nada sobre su propio contenido enunciativo —$C_\sigma(\psi)$ será independiente de $\mathcal{V}_{\mathfrak{M}}(D\tau)$ y $\mathcal{V}_{\mathfrak{M}}(\psi)$—.

El problema surge cuando $\mathcal{V}_{\mathfrak{M}}(\tau) = \psi$ y, al afirmar $\sigma(\tau)$, pretendemos afirmar algo sobre el valor de verdad o, más en general, sobre el contenido enunciativo de una oración, ψ, fuertemente equivalente a $\sigma(\tau)$. Esto solo podrá suceder cuando τ aparezca en $\sigma(\tau)$ precedido de un predicado de desentrecomillado, D; porque afirmar $D\tau$ es la forma de afirmar algo sobre el contenido enunciativo de la oración referenciada por τ, que es, precisamente, ψ. Si $D\tau$ es una

subfórmula de $\sigma(\tau)$, esta oración será de la forma $\rho(D\tau)$ y $\mathcal{V}_{\mathfrak{M}}(\sigma(\tau))$ dependerá de $\mathcal{V}_{\mathfrak{M}}(D\tau)$, es decir, existirá una función C_ρ definida entre W y W tal que $\mathcal{V}_{\mathfrak{M}}(\sigma(\tau)) = \mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = C_\rho(\mathcal{V}_{\mathfrak{M}}(D\tau))$. Con otras palabras, cuando $\sigma(\tau)$ sea de la forma particular $\rho(D\tau)$, se tendrá $C_\sigma(\psi) = C_\sigma(\sigma(\tau)) = C_\rho(\mathcal{V}_{\mathfrak{M}}(D\tau))$ y, teniendo en cuenta que $\mathcal{V}_{\mathfrak{M}}(D\tau)$ corresponde a la variable x, cambiaremos en (5.8) $C_\sigma(\psi)$ por $C_\rho(x)$. Así obtendremos, como particularización de (5.8) para el caso en que $\sigma(\tau)$ sea de la forma $\rho(D\tau)$, el sistema de ecuaciones:

$$\left.\begin{array}{l} x = C_D(y) \\ y = C_\rho(x) \end{array}\right\} \tag{4.112}$$

que, como sabemos, tiene tantas soluciones como puntos fijos tenga la función $(C_\rho \circ C_D)$.

En resumen, nuestra formalización resulta útil para distinguir la autorreferencia inocua de la *peligrosa*. Podemos aceptar sin problemas que un término τ tenga como referencia una oración $\sigma(\tau)$ cuando $\sigma(\tau)$ no contiene ningún predicado de desentrecomillado. En cambio, no es admisible que la referencia de τ sea una oración fuertemente equivalente a $\rho(D\tau)$ cuando $(C_\rho \circ C_D)$ no tenga ningún punto fijo (e incluso si tiene uno o varios puntos fijos, la situación es problemática). Esto supone una restricción de las oraciones que pueden ser referidas por un término.

La cuestión clave es ahora ¿está justificada tal restricción de la referencia?

Consideremos la oración del mentiroso reforzada. En principio se trata de una oración, ψ, de la forma $\neg T\tau$, donde el término τ referencia la propia oración y los grafos que definen las funciones $C_\neg$ y C_T son, respectivamente, $\{(\mathfrak{f},\mathfrak{v}),(\mathfrak{v},\mathfrak{f}),(\mathfrak{i},\mathfrak{i})\}$ y $\{(\mathfrak{f},\mathfrak{f}),(\mathfrak{v},\mathfrak{v}),(\mathfrak{i},\mathfrak{f})\}$. Pero la restricción anterior impide que τ referencie la oración, puesto que: a) $T\tau$ es una subfórmula de ψ; b) T es un predicado de desentrecomillado; c) C_ρ es en este caso $C_\neg$, porque $\mathcal{V}_{\mathfrak{M}}(\psi) = \mathcal{V}_{\mathfrak{M}}(\neg T\tau) = C_\neg(\mathcal{V}_{\mathfrak{M}}(T\tau))$; d) el grafo de la función compuesta $(C_\neg \circ C_T)$ es $\{(\mathfrak{f},\mathfrak{v}),(\mathfrak{v},\mathfrak{f}),(\mathfrak{i},\mathfrak{v})\}$, por lo que, dicha función no tiene ningún punto fijo.

Hay una razón sencilla y poderosa para impedir que τ sea un nombre de la oración $\neg T\tau$. Si τ es el nombre de una oración verdadera, $\neg T\tau$ será una oración falsa, luego τ no podrá ser el nombre de $\neg T\tau$. Y si τ es el nombre de una oración no verdadera, $\neg T\tau$ será una oración verdadera, luego tampoco τ podrá ser el nombre de $\neg T\tau$.[3] Este razonamiento se puede trasladar al caso general de una oración

[3] Un razonamiento similar, pero aplicado a nombres de enunciados, puede verse en Goldstein (2000, p. 55).

ψ fuertemente equivalente a una oración de la forma $\rho(D\tau)$, cuando $(C_\rho \circ C_D)$ no tenga ningún punto fijo: si τ tiene como referencia una oración φ, es decir, $\mathcal{V}_\mathfrak{M}(\tau) = \varphi$, se cumplirá:

$\mathcal{V}_\mathfrak{M}(D\tau) = C_D(\mathcal{V}_\mathfrak{M}(\varphi))$; $\mathcal{V}_\mathfrak{M}(\tau) = \varphi$ y definición de D
$\mathcal{V}_\mathfrak{M}(\rho(D\tau)) = C_\rho(\mathcal{V}_\mathfrak{M}(D\tau))$; composicionalidad

Llamando w, x, y a los valores de verdad $\mathcal{V}_\mathfrak{M}(\varphi)$, $\mathcal{V}_\mathfrak{M}(D\tau)$ y $\mathcal{V}_\mathfrak{M}(\rho(D\tau))$ respectivamente, obtenemos el sistema

$$\left.\begin{array}{l} x = C_D(w) \\ y = C_\rho(x) \end{array}\right\} \tag{5.9}$$

que equivale a

$$\left.\begin{array}{l} x = C_D(w) \\ y = (C_\rho \circ C_D)(w) \end{array}\right\} \tag{5.10}$$

Cuando $(C_\rho \circ C_D)$ no tiene ningún punto fijo, $y \neq w$, es decir, $\mathcal{V}_\mathfrak{M}(\rho(D\tau)) \neq \mathcal{V}_\mathfrak{M}(\varphi)$, y, por consiguiente, φ y $\rho(D\tau)$ no pueden ser fuertemente equivalentes. Ahora bien, ψ es fuertemente equivalente a $\rho(D\tau)$, luego ψ y φ no pueden ser la misma oración. Y como φ es la oración referenciada por τ, τ no puede ser un nombre de la oración ψ.

La segunda ecuación de (5.10) corresponde a $\mathcal{V}_\mathfrak{M}(\rho(D\tau)) = (C_\rho \circ C_D)(\mathcal{V}_\mathfrak{M}(\varphi))$. Por eso, en general, podemos conocer $\mathcal{V}_\mathfrak{M}(\rho(D\tau))$ a partir de $\mathcal{V}_\mathfrak{M}(\varphi)$. Pero cuando φ es fuertemente equivalente a $\rho(D\tau)$ y $(C_\rho \circ C_D)$ tiene varios puntos fijos, el valor $\mathcal{V}_\mathfrak{M}(\rho(D\tau))$ no quedaría determinado por el lenguaje interpretado. Un ejemplo de esta situación lo proporciona la oración del veraz, es decir, una oración de la forma $T\tau$, donde el término τ referencia la propia oración.

5.1.4. Conclusiones

Después de analizar posibles restricciones a los predicados de desentrecomillado o a la autorreferencia, conviene resaltar algunas importantes conclusiones:

1. Hemos distinguido la autorreferencia problemática de la que no lo es: si en la oración autorreferencial $\sigma(\tau)$ no hay una aparición de τ —referencial respecto a $\sigma(\tau)$— en ninguna expresión de la forma $D\tau$, el hecho de que τ tenga como referencia la propia oración $\sigma(\tau)$ no crea ningún problema (al

afirmar la oración autorreferencial $\sigma(\tau)$ se afirma algo sobre los aspectos formales de la propia oración, no sobre su valor de verdad).

2. Incluso la mínima restricción de los predicados de desentrecomillado necesaria para evitar conclusiones contradictorias se muestra problemática e injustificada. La tentativa de redefinir este tipo de predicados mediante el esquema (5.3) —p. 147— supondría una restricción inadmisible a la capacidad expresiva del lenguaje formal. Adicionalmente plantea el problema de que no tenemos ningún criterio que nos diga qué elemento de W tiene que ser φ_D ($\mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\varphi^{\urcorner})$ cuando φ no es desentrecomillable).

3. La propuesta de restringir la autorreferencia evitando que un término τ haga referencia a una oración ψ fuertemente equivalente a una oración de la forma $\rho(D\tau)$, cuando la función $(C_\rho \circ C_D)$ no tenga ningún punto fijo, está, a mi juicio, plenamente justificada: según hemos visto en el apartado anterior, τ no puede ser el nombre de una oración fuertemente equivalente a $\rho(D\tau)$ cuando $(C_\rho \circ C_D)$ no tiene ningún punto fijo.

5.1.5. La paradoja del mentiroso y los contextos referencialmente anuladores

En cuanto a la paradoja del mentiroso reforzada y, en general, a las paradojas asociadas a oraciones de la forma $\rho(D\tau)$ para las que $(C_\rho \circ C_D)$ no tiene ningún punto fijo, hemos llegado a la siguiente conclusión: el término τ no puede tener como referencia ninguna oración fuertemente equivalente a $\rho(D\tau)$.

Siguiendo el planteamiento de Martin podemos decir que hemos eliminado la paradoja del mentiroso reforzada justificando que la siguiente afirmación no es verdadera:

(S) Hay una oración que dice de sí misma únicamente que no es verdadera

En el lenguaje formal podemos representar la oración del mentiroso reforzada como $\neg T\tau$, pero hemos visto que la pretensión de que el término τ tenga como referencia la oración $\neg T\tau$ es vana. Podemos parafrasear el razonamiento que impide que τ tenga como referencia la oración $\neg T\tau$ cambiando el lenguaje formal por el natural: si A es el nombre de una oración verdadera, "A no es verdadera" será una oración falsa, y si A es el nombre de una oración no verdadera, "A no es verdadera"

será una oración verdadera, luego A no podrá ser el nombre de la oración "A no es verdadera". Si cambiamos A por "esta oración", llegamos a la conclusión de que la expresión "esta oración", en su aparición dentro de la oración "esta oración no es verdadera" no puede tener como referencia dicha oración. Ahora bien, como tampoco tiene sentido que tenga como referencia otra oración distinta a aquella de la que forma parte, debemos concluir que la expresión "esta oración" —en su aparición dentro de la oración "esta oración no es verdadera"— carece de referencia. Así pues si formalizamos la oración del mentiroso reforzada mediante $\neg T\tau$, el término τ no debe tener referencia alguna. Estamos ahora en condiciones de conocer el valor de verdad de esta oración: al ser τ un término sin referencia, $\mathcal{V}_{\mathfrak{M}}(T\tau) = \mathfrak{i}$,[4] luego $\mathcal{V}_{\mathfrak{M}}(\neg T\tau) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(T\tau)) = C_{\neg}(\mathfrak{i})$. Si, vaya por caso, usamos la lógica trivaluada fuerte de Kleene, $C_{\neg}(\mathfrak{i}) = \mathfrak{i}$, y entonces $\mathcal{V}_{\mathfrak{M}}(\neg T\tau) = \mathfrak{i}$.

Quine (1976, p. 7) alega que la explicación de que "esta oración", en su aparición dentro de la oración del mentiroso, carece de referencia no es aplicable a la oración siguiente:[5]

(Q) "añadida a su propia cita, es una oración no verdadera" añadida a su propia cita, es una oración no verdadera

Aquí la autorreferencia se consigue gracias a la operación de añadir una secuencia de caracteres a su propia cita. Como nos muestra Smullyan (1996), en un lenguaje formal se puede conseguir autorreferencia de varias formas entre las que hemos destacado la diagonalización, la tarskificación y la norma (añadir a una secuencia de caracteres su propia cita). Todas ellas, así como la operación de Quine, tienen algo en común: son funciones que toman como argumento una secuencia de caracteres y pueden obtener como valor una oración. Por su carácter puramente sintáctico parece difícil negar que, por razones semánticas, puedan aplicarse siempre a cualquier argumento. Sin embargo, en ciertos contextos, como el de la oración del mentiroso, así ocurre.

En efecto, tomemos, por ejemplo, el caso de la diagonalización. La hemos definido como una función, d, que asocia la oración $\sigma(\ulcorner\sigma(x)\urcorner)$ a una fórmula, $\sigma(x)$, con una única variable libre, x, es decir, el término $d\ulcorner\sigma(x)\urcorner$ tiene como

[4] Recuérdese que esto es cierto tanto en los lenguajes trivaluados extensionales como en los enunciativos trivaluados (ver proposición 4.8, p. 124).

[5] Hemos hecho una traducción un tanto libre de la oración del texto original (*'Yields a falsehood when appended to its own quotation' yields a falsehood when appended to its own quotation*) pero respetando lo esencial de su estructura y significado paradójico.

referencia la oración $\sigma(\ulcorner\sigma(x)\urcorner)$. Pero entonces, el término $d\ulcorner\neg Tdx\urcorner$ tendría como referencia la oración $\neg Td\ulcorner\neg Tdx\urcorner$; lo cual contradice nuestra conclusión de que, en el supuesto de que la función $(C_\neg \circ C_T)$ no tiene ningún punto fijo, un término τ no puede tener como referencia la oración $\neg T\tau$.

Debemos admitir que el término $d\ulcorner\neg Tdx\urcorner$ no puede tener como referencia la oración $\neg Td\ulcorner\neg Tdx\urcorner$ en un lenguaje interpretado en el que $(C_\neg \circ C_T)$ no tenga ningún punto fijo. Esto puede ir contra nuestras intuiciones, pero, como ya se ha señalado en este trabajo, precisamente es necesario revisar los principios que normalmente aceptamos, incluidas nuestras intuiciones, para resolver una paradoja.

Sin embargo, contra lo que a primera vista pudiese parecer, no hay una diferencia sustancial entre rechazar que el término $d\ulcorner\neg Tdx\urcorner$ pueda tener como referencia la oración $\neg Td\ulcorner\neg Tdx\urcorner$ o rechazar que un nombre A pueda ser el nombre de la oración "A no es verdadera" o que la expresión "esta oración", en su aparición dentro de la oración "esta oración no es verdadera" pueda tener como referencia dicha oración. En los tres casos tenemos un nombre y una oración que, debido a su significado, no puede ser la referencia de ese nombre. El que un nombre sea explícito (A), o construido ($d\ulcorner\neg Tdx\urcorner$, "esta oración") no afecta al razonamiento anterior.

Conviene recordar que, en nuestros lenguajes formales, si f es un símbolo de función n-ádica y $t_1, t_2, ..., t_n$ son términos cerrados, $f\,t_1, t_2, ..., t_n$ es un término cerrado, un nombre (en sentido amplio). Así se usan también los símbolos de función en ámbitos menos formales e incluso en lenguaje natural: con el nombre de la función y los de sus argumentos se forma un nombre del valor que la función asocia a esos argumentos. Por ejemplo, "$sen(\pi/3)$" puede verse como un nombre del número $0,5$ y "el hijo de Juan y María" puede ser, en un contexto determinado, una forma correcta de referirnos a una persona concreta.

Consiguientemente, cuando se define una función, cuando se establece qué valor asigna la función a cada posible secuencia de sus argumentos, estamos al mismo tiempo estableciendo un nombre para cada uno de los valores del rango de la función: el nombre que resulta de combinar, con la sintaxis adecuada, el nombre de la función y un nombre para cada argumento. En ese sentido definir una función puede verse como una forma de crear muchos nombres de una sola vez.

Si rechazamos que un nombre A pueda ser el nombre de la oración "A no es verdadera" también debemos rechazar definir una función cuando eso suponga la

creación de algún nombre cuya referencia sea una oración que dice que la oración a la que se refiere ese nombre no es verdadera. Entonces —en un lenguaje interpretado en el que la negación, $\neg$, y el predicado de verdad, T, se interpretan mediante las funciones $C_\neg$ y C_T de modo que $(C_\neg \circ C_T)$ no tiene ningún punto fijo— debemos rechazar definir la función de diagonalización de la forma que lo hemos hecho, porque el término $d\ulcorner\neg Tdx\urcorner$ tendría como referencia la oración $\neg Td\ulcorner\neg Tdx\urcorner$ que *dice* que la oración a la que se refiere ese término no es verdadera. Pero esta misma visión es aplicable a otras funciones que pueden usarse para conseguir la autorreferencia como son la tarskificación y la norma. La norma (añadir a una secuencia de caracteres su propia cita) es, en un lenguaje formal, una función muy parecida a la que se usa en lenguaje natural para construir la oración (Q) de Quine (añadir una secuencia de caracteres a su propia cita). Por tanto la explicación que sirve para la diagonalización y para la norma sirve también para la oración de Quine. En este caso la función que se utiliza podríamos indicarla mediante: "() añadida a su propia cita". A algunos argumentos esta función asocia oraciones: por ejemplo, si el argumento es "está compuesta por cinco palabras" la función asocia como valor la oración verdadera:

> "está compuesta por cinco palabras" está compuesta por cinco palabras

a la que podemos referirnos mediante el nombre (en forma de descripción):

> "está compuesta por cinco palabras" añadida a su propia cita

De forma análoga, el nombre (en forma de descripción)

> "añadida a su propia cita es una oración no verdadera" añadida a su propia cita

tiene aparentemente como referencia la oración de Quine:

(Q) "añadida a su propia cita es una oración no verdadera" añadida a su propia cita es una oración no verdadera

Llegamos así a una situación de perplejidad típica de una buena paradoja. Por una parte, puesto que negamos que un nombre pueda tener como referencia una oración que dice que la oración a la que se refiere ese nombre no es verdadera, tenemos que negar que el nombre

> "añadida a su propia cita es una oración no verdadera" añadida a su propia cita

tenga como referencia la oración (Q). Es decir, en lenguaje castellano, la función "() añadida a su propia cita" no puede entenderse de modo que asocie al argumento "añadida a su propia cita es una oración no verdadera" la oración (Q), del mismo modo que la función diagonalización, d, no puede interpretarse de modo que asocie a la fórmula $\neg Tdx$ la oración $\neg Td^{\ulcorner}\neg Tdx^{\urcorner}$. Por otra parte, la operación de añadir algo a su propia cita es tan clara y sencilla que parece innegable que si añadimos "añadida a su propia cita es una oración no verdadera" a su propia cita obtendremos la oración (Q). Prueba de ello es que no habría problema en mecanizar esa operación, pues es fácil obtener un programa de ordenador que tome como entrada una secuencia de caracteres y obtenga como salida el resultado de añadir a su cita dicha secuencia. Y si a ese programa le damos como entrada la cadena "añadida a su propia cita es una oración no verdadera" obtendría como salida la oración (Q). Además, podríamos obtener oraciones no problemáticas que incluyesen el nombre

> "añadida a su propia cita es una oración no verdadera" añadida a su propia cita

como nombre de la oración (Q). Por ejemplo:

> "añadida a su propia cita es una oración no verdadera" añadida a su propia cita es una oración con más de diez palabras

que equivaldría a:

> (Q) es una oración con más de diez palabras

Por tanto, fieles al principio de no restringir más de lo necesario tendremos que aceptar que, normalmente, el nombre

> "añadida a su propia cita es una oración no verdadera" añadida a su propia cita

tiene como referencia la oración (Q) de Quine. Pero no podemos aceptar que esto ocurra en la aparición del nombre como parte de la propia oración (Q) de Quine. Tenemos pues que admitir que un mismo nombre tiene referencia en muchos casos pero no en todos: cuando $(C_\rho \circ C_D)$ no tiene ningún punto fijo, el nombre τ, en su aparición en la oración $\rho(D\tau)$, no puede tener como referencia una oración fuertemente equivalente a $\rho(D\tau)$.

De este modo, la paradoja se resuelve, sin restringir más de lo necesario, a cambio de renunciar al principio de que la referencia de un nombre es independiente de la oración en la que aparece: ahora tenemos que aceptar que un nombre, que habitualmente tiene una determinada referencia, deja de tener referencia alguna cuando aparece formando parte de cierta oración. Ello supone aceptar cierta contextualidad en la interpretación de los nombres y, por tanto, cierta pérdida de composicionalidad en la interpretación de las expresiones bien formadas (términos y fórmulas).

Sin embargo, la pérdida de composicionalidad producida por dicha contextualidad es mínima dado que no altera la forma en que la referencia de (la aparición) de un término o su falta de referencia interviene en la determinación de los valores de verdad de las oraciones. Esto se debe a que, en nuestros lenguajes interpretados trivaluados, ya teníamos un criterio para asignar valores de verdad a oraciones que contienen nombres sin referencia —téngase en cuenta que si R es un símbolo de relación n-ádica y alguno de los términos $t_1, t_2, ..., t_n$ carece de referencia, $\mathcal{V}_{\mathfrak{M}}(R\,t_1, t_2, ..., t_n) = \mathfrak{i}$—.

La contextualidad en la interpretación de los nombres a que aquí nos referimos es una contextualidad que podemos calificar de semántica. Porque lo que hace que, por ejemplo, un nombre A, en su aparición en la oración "A no es verdadera", no pueda tener como referencia la propia oración, es que las reglas semánticas determinan que, si A tiene como referencia una oración, el valor de verdad de "A no es verdadera" depende del valor de verdad de la oración referenciada por A (en su aparición en dicha oración) de forma que ambos valores son siempre distintos.

Por tanto, en nuestro análisis, el contexto clave en el problema de la paradoja del mentiroso, no es el contexto de emisión de la oración (una oración como "esta oración es falsa" que parece encerrarse en sí misma es quizá de las más independientes del contexto de emisión que podamos encontrar). No estamos pues ante una propuesta del estilo de las de Barwise y Etchemendy (1987) o Simmons (1993)

que consideran que la explicación de la paradoja del mentiroso radica en que el significado del predicado "verdadero" es sensible al contexto de emisión.

Nuestra propuesta sugiere que, referencialmente, no solo existen contextos transparentes y opacos —por usar la terminología de Quine— sino también lo que podríamos denominar contextos referencialmente anuladores, contextos en los que un término no tiene su referencia aparente o habitual ni ninguna otra. Por ejemplo, la oración del mentiroso reforzada crea un contexto anulador de la referencia de su término sujeto. También es un contexto generalmente anulador el entrecomillado. Pero, en este caso, el contexto se establece sintácticamente lo que hace más sencilla su detección, mientras que en la oración del mentiroso y afines, se establece semánticamente.[6]

Sin embargo, es importante señalar, los términos de cita no se ven afectados por esta contextualidad. Sabemos que si D es un predicado de desentrecomillado y φ es una oración, se cumple $\mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\varphi^{\urcorner}) = C_D(\mathcal{V}_{\mathfrak{M}}(\varphi))$ luego, una vez encontrado el valor $\mathcal{V}_{\mathfrak{M}}(\varphi)$, que siempre existe, podremos calcular $\mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\varphi^{\urcorner})$ sin problema.

Para comprobar que nunca hay necesidad de que el término de cita $^{\ulcorner}\varphi^{\urcorner}$ pierda su referencia en un oración de la forma $D^{\ulcorner}\varphi^{\urcorner}$ —que, como hemos estudiado, es el tipo de oraciones en que un término puede perder su referencia— podríamos introducir en el lenguaje formal interpretado una conectiva monádica, Δ, que se interpretase mediante una función C_Δ idéntica a C_D. Entonces la interpretación de $\Delta\varphi$ y la interpretación de $D^{\ulcorner}\varphi^{\urcorner}$ cuando el término de cita $^{\ulcorner}\varphi^{\urcorner}$ conserva su referencia, serían iguales, ya que:

$$\mathcal{V}_{\mathfrak{M}}(\Delta\varphi) = C_\Delta(\mathcal{V}_{\mathfrak{M}}(\varphi)) = C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)) = \mathcal{V}_{\mathfrak{M}}(D^{\ulcorner}\varphi^{\urcorner}) \tag{5.11}$$

Ahora bien, una vez evaluada φ, no habrá problema en evaluar $\Delta\varphi$. Mas si $\Delta\varphi$ se ha evaluado correctamente, la evaluación de $D^{\ulcorner}\varphi^{\urcorner}$ bajo el supuesto de que $^{\ulcorner}\varphi^{\urcorner}$ conserva su referencia, también será correcta, dado que es la misma que la de $\Delta\varphi$. Así pues, que el término de cita $^{\ulcorner}\varphi^{\urcorner}$ conserve su referencia no es ningún obstáculo o problema para evaluar $D^{\ulcorner}\varphi^{\urcorner}$. De hecho, hay una oración, $\Delta\varphi$, con el mismo valor de verdad que $D^{\ulcorner}\varphi^{\urcorner}$, donde el término de cita $^{\ulcorner}\varphi^{\urcorner}$ no aparece.

Al ser nombres explícitos, nombres de los que forma parte lo nombrado, los términos de cita no pueden nombrar la oración que los contiene, no pueden provo-

[6] Cierta similitud existe en los contextos referencialmente opacos. El entrecomillado es uno de ellos (sintáctico). Otros contextos opacos como los contextos modales o contextos de la forma "sabe que...", "dice que...", etc. son de naturaleza semántica.

car por sí mismos autorreferencia directa ni indirecta. Consideremos por ejemplo la oración "esta oración es falsa". Si cambiamos el nombre "esta oración" por el término de cita ""esta oración es falsa"" obtenemos ""esta oración es falsa" es falsa" pero aquí el sujeto no se refiere a la oración de la que forma parte, a diferencia de lo que ocurre en la oración "esta oración es falsa".

Es muy importante destacar que el uso de términos de cita y predicados de desentrecomillado nos permite afirmar sin problemas y dentro del lenguaje si determinada oración es verdadera, si es falsa, o si no es verdadera ni falsa. Recuérdese que en muchas propuestas de solución a la paradoja del mentiroso, ello solo es posible en el metalenguaje.

Tomemos, por ejemplo, la oración del mentiroso reforzada $\neg T\tau$, donde T es el predicado "verdadero" y τ es un término cuya referencia (habitual) es la oración $\neg T\tau$. ¿Cómo podemos afirmar en el lenguaje que la oración $\neg T\tau$ no es verdadera? Es claro que no podemos hacerlo mediante $\neg T\tau$, porque en esta oración el término τ carece de referencia. Pero el uso de un término de cita nos da una sencilla solución: la oración $\neg T^{\ulcorner}\neg T\tau^{\urcorner}$. A diferencia de $\neg T\tau$, que no es verdadera ni falsa, la oración $\neg T^{\ulcorner}\neg T\tau^{\urcorner}$ es verdadera, pues su evaluación es:

$$\begin{aligned}&\mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\neg T\tau^{\urcorner}) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(T^{\ulcorner}\neg T\tau^{\urcorner})) = C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\neg T\tau))) = \\ &= C_{\neg}(C_T(\mathfrak{i})) = C_{\neg}(\mathfrak{f}) = \mathfrak{v}\end{aligned} \tag{5.12}$$

En general, donde hasta ahora hablábamos de la referencia de un término, ahora deberemos hablar de la referencia de una aparición de un término. Para ello cambiaremos el término t por

$$t^{[A]} \tag{5.13}$$

donde A es una especificación de una aparición de t en una oración (o en un término cerrado).

Por ejemplo, para indicar que la referencia del término cerrado τ en su aparición en la oración $P\tau$, es la propia oración, podemos escribir:

$$\mathcal{V}_{\mathfrak{M}}(\tau^{[P\tau]}) = P\tau \tag{5.14}$$

Para indicar que el término $d^{\ulcorner}\neg Tdx^{\urcorner}$ no tiene referencia en su aparición en la oración $\neg Td^{\ulcorner}\neg Tdx^{\urcorner}$, podríamos escribir:

$$d^{\ulcorner}\neg Tdx^{\urcorner[\neg Td^{\ulcorner}\neg Tdx^{\urcorner}]} \notin dom(\mathcal{V}_{\mathfrak{M}})$$

En cambio, si en su aparición en la oración $Pd^{\ulcorner}\neg Tdx^{\urcorner}$, ese mismo término tiene como referencia la oración $\neg Td^{\ulcorner}\neg Tdx^{\urcorner}$, escribiríamos:

$$\mathcal{V}_{\mathfrak{M}}(d^{\ulcorner}\neg Tdx^{\urcorner[Pd^{\ulcorner}\neg Tdx^{\urcorner}]}) = \neg Td^{\ulcorner}\neg Tdx^{\urcorner} \tag{5.15}$$

En general, cuando dos apariciones de un mismo término, $t^{[A]}$ y $t^{[B]}$, tienen referencia (respecto a una asignación de variables), ambas tienen la misma referencia (respecto a esa asignación de variables), es decir, $\mathcal{V}_{\mathfrak{M},s}(t^{[A]}) = \mathcal{V}_{\mathfrak{M},s}(t^{[B]})$. A esa referencia la llamaremos la referencia *habitual* del término t (respecto a esa asignación de variables), que designaremos mediante:

$$\mathcal{V}^{h}_{\mathfrak{M},s}(t) \tag{5.16}$$

Nótese que la referencia habitual de un término (respecto a una asignación de variables) es la única referencia posible de ese término (respecto a esa asignación de variables).

Cuando t es un término cerrado su referencia habitual no depende de ninguna asignación de variables, por lo que la denotaremos mediante:

$$\mathcal{V}^{h}_{\mathfrak{M}}(t) \tag{5.17}$$

5.1.6. Otras oraciones autorreferenciales y contextos referencialmente anuladores

¿Es generalizable el planteamiento de los contextos referencialmente anuladores a las oraciones de la forma $\rho(D\tau)$ cuando la función $(C_\rho \circ C_D)$ tiene uno o varios puntos fijos?

5.1.6.1. Un único punto fijo

Como hemos visto en el apartado 4.2.3, cuando $(C_\rho \circ C_D)$ tiene un único punto fijo caben dos planteamientos diferentes:

1. Se considera que el punto fijo de $(C_\rho \circ C_D)$ es el valor de verdad de la oración autorreferencial $\rho(D\tau)$. Con este planteamiento no hay motivo para que la referencia del término τ quede anulada: no hay problema en que la referencia de τ sea una oración fuertemente equivalente a $\rho(D\tau)$. Llamamos G

a este planteamiento, por estar, en buena medida de acuerdo con el análisis de Gupta (1982, ejemplo (3) en parte IV). También podemos llamarlo planteamiento extensional.

2. Se considera que, cualquiera que sea el punto fijo de $(C_\rho \circ C_D)$, la oración no puede ser verdadera ni falsa a menos que tenga un contenido enunciativo (completo). Con diferentes estilos, los planteamientos de Mackie (1973) y Kripke (1975) sugieren que si una oración atómica, $\alpha(D\tau)$, en la que aparecen predicados de desentrecomillado[7] posee un contenido enunciativo, dicho contenido será reducible al de otra oración atómica, φ, en la que no aparecen predicados de ese tipo: $\mathcal{U}_{\mathfrak{M}}(\alpha(D\tau)) \equiv \mathcal{U}_{\mathfrak{M}}(\varphi)$. En cuanto a las oraciones compuestas, su valor de verdad se deducirá según las definiciones de las conectivas y cuantificadores que se establezcan;[8] (por simplicidad, también convendré en decir que una oración compuesta tiene contenido enunciativo completo solo si es verdadera o es falsa). Llamamos MK a este planteamiento, por estar, en buena medida de acuerdo con ideas de Mackie y Kripke. También podemos llamarlo planteamiento intensional.

El planteamiento MK tiene la limitación de que si el término τ hace referencia a una oración fuertemente equivalente a $\rho(D\tau)$ no se puede asignar a $\rho(D\tau)$ otro valor de verdad que el punto fijo de $(C_\rho \circ C_D)$, a no ser que consideremos el contexto de la oración $\rho(D\tau)$ como anulador de la referencia habitual de τ. Así pues, si de algún modo se establece que $\rho(D\tau)$ no tiene contenido enunciativo completo, es decir, que $\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = \mathfrak{i}$, el contexto de la oración $\rho(D\tau)$ anulará la referencia habitual de τ, a menos que el único punto fijo de $(C_\rho \circ C_D)$ sea $\mathfrak{i}$.

Un ejemplo informal de esta situación lo tenemos con la oración "esta oración carece de contenido enunciativo (completo)". Si siguiendo a Mackie, aceptamos que dicha oración no tiene contenido y, por tanto, al afirmarla no se afirma nada verdadero ni falso; nos vemos avocados a aceptar que la expresión "esta oración", en su aparición en la oración entrecomillada anterior, no tiene como referencia la oración (si la tuviese, afirmar la oración sería afirmar algo verdadero). Tendría-

[7] Lo que nosotros llamamos predicados de desentrecomillado representan lo que Mackie (1973, p. 286) denomina propiedades derivativas. Por otra parte, el predicado de verdad de Kripke (1975) es un predicado de desentrecomillado.

[8] Por ejemplo, Mackie (1973, p. 289) señala que la afirmación de Epiménides (todo lo que afirme un cretense es falso) será falsa si algún cretense afirma algo verdadero, pero no intenta reducirla a una oración en la que no aparezca el predicado *falso*.

mos que admitir que la expresión "esta oración", en su aparición en la oración entrecomillada anterior, no tiene referencia (a pesar de las apariencias).

En general, si la referencia habitual de un término τ es una oración fuertemente equivalente a una oración de la forma $\rho(D\tau)$, pero establecemos que esta oración no tiene contenido enunciativo completo y sabemos que el único punto fijo de $(C_\rho \circ C_D)$ no es $\mathfrak{i}$, tendremos que considerar que el término τ, en su aparición en la oración $\rho(D\tau)$, no tiene referencia. Si ν es un término que nunca tiene referencia, $\nu \notin dom(\mathcal{V}_\mathfrak{M})$, podremos decir que $\mathcal{V}_\mathfrak{M}(\rho(D\tau)) = \mathcal{V}_\mathfrak{M}(\rho(D\nu))$. Además: $\nu \notin dom(\mathcal{V}_\mathfrak{M}) \Rightarrow \mathcal{V}_\mathfrak{M}(D\nu) = \mathfrak{i}$. Finalmente deducimos:

$$\mathcal{V}_\mathfrak{M}(\rho(D\tau)) = \mathcal{V}_\mathfrak{M}(\rho(D\nu)) = C_\rho(\mathcal{V}_\mathfrak{M}(D\nu)) = C_\rho(\mathfrak{i})$$

Como $\mathcal{V}_\mathfrak{M}(\rho(D\tau)) = \mathfrak{i}$, se concluye que $C_\rho(\mathfrak{i}) = \mathfrak{i}$. Llegamos así a un interesante condicionamiento del planteamiento MK: cualquiera que sea el criterio para determinar si la oración $\rho(D\tau)$ tiene contenido enunciativo, la respuesta ha de ser afirmativa cuando $C_\rho(\mathfrak{i}) \neq \mathfrak{i}$ y $(C_\rho \circ C_D)$ tenga un único punto fijo perteneciente a $\{\mathfrak{f}, \mathfrak{v}\}$.

5.1.6.2. Varios puntos fijos

Cuando la función $(C_\rho \circ C_D)$ tenga varios puntos fijos, el lenguaje interpretado dejaría sin determinar el valor de verdad de la oración autorreferencial $\rho(D\tau)$, es decir, la función de valuación $\mathcal{V}_\mathfrak{M}$ no quedaría completamente definida. El mismo problema tenemos en el lenguaje natural: así la oración del veraz puede ser considerada verdadera o ser considerada falsa pero la simetría entre ambas posibilidades no nos permite decantarnos claramente por una de ellas.

En cualquier caso, mantenemos el muy razonable principio de que toda oración (emitida en un determinado contexto del que nos abstraemos en el lenguaje formal) es verdadera o es falsa o no es verdadera ni falsa y que cada una de estas tres posibilidades excluye las otras dos.[9] Por eso queremos que $\mathcal{V}_\mathfrak{M}$ quede completamente definida. Hay varias formas de conseguirlo.[10] Cada una de ellas

[9] Rechazar este principio, no solo es antiintuitivo, sino que, seguramente crearía más problemas que los que supuestamente pueda resolver. Así lo hemos comprobado, por ejemplo, en la crítica en este trabajo a las propuestas *vagas* y a las propuestas *inconsistentes* de resolución de la paradoja del mentiroso.

[10] La situación presenta cierta similitud con la planteada en matemáticas ante las sucesivas ampliaciones del concepto de número y las relaciones y operaciones aritméticas: cuando, vaya por caso, se introduce el concepto de número complejo no hay una única manera obvia de

corresponde a un modo de extender los conceptos de verdadero y falso a oraciones autorreferenciales para las que nuestras intuiciones convencionales no permiten clasificarlas claramente en lo que respecta a su valor de verdad.

Distinguiré dos tipos de planteamientos, los coherentes con la opción G (o extensional) y los coherentes con la opción MK (o intensional) del apartado anterior.

1. Opción G. En principio caben dos posibilidades dentro de esta opción. Supongamos que la referencia habitual del término τ es una oración fuertemente equivalente a $\rho(D\tau)$ —y, por supuesto, que $(C_\rho \circ C_D)$ tiene varios puntos fijos—. Una posibilidad es que la aparición del término τ en la oración $\rho(D\tau)$ conserve su referencia habitual; la otra es que pierda su referencia. Sin embargo, la primera posibilidad tiene graves inconvenientes. Tenemos varios valores de verdad posibles para la oración $\rho(D\tau)$ y ningún criterio que nos obligue a elegir uno de ellos. Por tanto si queremos definir completamente el valor de verdad de toda oración tenemos que establecer, no sin gran arbitrariedad, un criterio por convenio.[11] Se puede alegar que, situaciones parecidas, son normales en el ámbito matemático sin que causen extrañeza: por ejemplo el resto de la división entera de -8 entre 3 puede tomarse como -2 (y el cociente sería -2) o como 1 (y el cociente sería -3) dependiendo del criterio que, por convenio, elijamos para dividir. Ahora bien, en el caso que nos ocupa, los problemas para asignar un valor de verdad son mayores si consideramos sistemas de oraciones que admiten varias asignaciones de valores de verdad, como el siguiente

$$F\tau_2; \quad F\tau_1 \tag{5.18}$$

(supóngase que la referencia habitual de τ_1 es la oración $F\tau_2$, que la referencia habitual de τ_2 es $F\tau_1$ y que F representa el predicado "falso"): En este caso, una oración será verdadera y la otra falsa, pero el problema es ¿qué valor de verdad damos a cada una? Forzando mucho las cosas podríamos establecer un criterio por el cual diéramos cierta prioridad a la primera oración (a la

generalizar la relación "menor que" existente en el conjunto de los números reales y se ha de recurrir a criterios como la sencillez, la conservación del mayor número de propiedades posible, etc. para elegir una de las alternativas.

[11] Por ejemplo, se podría decidir elegir siempre el valor *posterior* de los posibles según la relación de orden total, $\mathfrak{f} \prec \mathfrak{i} \prec \mathfrak{v}$, que hemos establecido por convenio en el conjunto W. Con este criterio, la oración del veraz tomaría el valor de verdad $\mathfrak{v}$ dado que el otro valor posible es $\mathfrak{f}$ y $\mathfrak{f} \prec \mathfrak{v}$.

que, por ejemplo, asignaríamos primero un valor de verdad según el criterio que ya teníamos para el caso de una única oración como la del veraz).[12] Pero ni siquiera esto es posible en el caso de una tarjeta que en cada una de sus caras tiene escritas únicamente la oración "la oración del otro lado de esta tarjeta es falsa"; la situación es tan simétrica que parece inevitable que las dos oraciones tengan que tener el mismo valor de verdad. Por todas estas razones, en lo sucesivo, descartaremos esta posibilidad.

Considero mucho más razonable la posibilidad alternativa: que el término τ en su aparición en la oración $\rho(D\tau)$ no pueda tener como referencia una oración fuertemente equivalente a $\rho(D\tau)$ cuando $(C_\rho \circ C_D)$ tiene varios puntos fijos. De esta forma se da un mismo tratamiento al caso en que $(C_\rho \circ C_D)$ tiene varios puntos fijos que al caso en que $(C_\rho \circ C_D)$ no tiene ningún punto fijo y, por ende, el mismo tratamiento a la paradoja del veraz que a la del mentiroso. Es más, podemos decir que hay un tratamiento común para todos los casos en que la referencia habitual de un término τ es una oración fuertemente equivalente a $\rho(D\tau)$. Es el siguiente:

> la aparición del término τ en la oración $\rho(D\tau)$ tiene referencia ssi la función $(C_\rho \circ C_D)$ tiene un único punto fijo

o, en términos de ecuaciones,

> la aparición del término τ en la oración $\rho(D\tau)$ tiene referencia ssi la ecuación $x = (C_\rho \circ C_D)(x)$ tiene una única solución

Otra ventaja de este planteamiento es que es fácil de extender a sistemas de oraciones como (5.18). Si llamamos x e y a los valores de verdad de $F\tau_2$ y $F\tau_1$, respectivamente, y C_F a la función de W en W asociada al predicado "falso", el sistema de ecuaciones de valores de verdad asociado al sistema de oraciones será:

$$x = C_F(y);\ y = C_F(x)) \tag{5.19}$$

y dado que no tiene una única solución (tiene dos) concluiremos que las apariciones de los términos τ_1 y τ_2 en las oraciones $F\tau_1$ y $F\tau_2$ carecen de referencia. Aplicado al ejemplo de la tarjeta —del que (5.18) es una formalización—,

[12] Con el criterio señalado en la nota al pie anterior, la oración $F\tau_2$ tomaría el valor de verdad $\mathfrak{v}$ y, como consecuencia, la oración $F\tau_1$ sería falsa.

el resultado es que las expresiones "la oración del otro lado de esta tarjeta" que aparecen en ella carecen de referencia debido al contexto en que se hallan. Por tanto, ninguna de las dos oraciones de la tarjeta será verdadera ni falsa (ambas tendrán el mismo valor de verdad: $\mathfrak{i}$).

En resumen, la extensión más razonable de la opción G al caso en que $(C_\rho \circ C_D)$ tenga varios puntos fijos, consiste en que el término τ, en su aparición en la oración $\rho(D\tau)$, no pueda tener como referencia una oración fuertemente equivalente a $\rho(D\tau)$.

2. Opción MK. Es claro que la oración autorreferencial $\rho(D\tau)$ no puede reducirse a una oración φ sin predicados de desentrecomillado, porque, dado que $(C_\rho \circ C_D)$ tiene varios puntos fijos, $\rho(D\tau)$ admite coherentemente varios valores de verdad, en tanto que φ solo admite uno por carecer de predicados de desentrecomillado. Por consiguiente, desde el punto de vista MK, la oración $\rho(D\tau)$ carece de contenido enunciativo. De hecho Mackie (1973, p. 260) considera que la pronunciación declarativa de la oración del veraz ("lo que ahora digo es verdadero"), ejemplo paradigmático de este tipo de oraciones, no constituye un enunciado verdadero ni uno falso porque realmente no afirma nada.

   Lo natural en este planteamiento es pues considerar que $\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = \mathfrak{i}$. Cuando $\mathfrak{i}$ sea un punto fijo de $(C_\rho \circ C_D)$ podrá mantenerse la referencia habitual de τ pero, en caso contrario, habría que admitir que el término τ no tiene referencia en su aparición en la oración $\rho(D\tau)$. Trasladado al ejemplo informal de la oración del veraz anterior, la expresión "lo que ahora digo" carecería de referencia lo cual explicaría que la oración no se pueda considerar verdadera ni falsa.

   Ahora bien, como hemos visto más arriba, si τ no tiene referencia en su aparición en la oración $\rho(D\tau)$, se cumple $\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = C_\rho(\mathfrak{i})$. Por eso, un condicionante de este planteamiento en el que $\mathcal{V}_{\mathfrak{M}}(\rho(D\tau)) = \mathfrak{i}$, es que debe cumplirse $C_\rho(\mathfrak{i}) = \mathfrak{i}$ (cuando $\mathfrak{i}$ no es un punto fijo de $(C_\rho \circ C_D)$).

Conviene observar algunas similitudes y diferencias entre las dos opciones que hemos denominado G y MK. En ambas, se acaba resolviendo del mismo modo la oración del mentiroso y, en general, las oraciones de la forma $\rho(D\tau)$ para las que $(C_\rho \circ C_D)$ no tiene ningún punto fijo: el término τ, en su aparición en la

oración $\rho(D\tau)$ no puede tener como referencia una oración fuertemente equivalente a $\rho(D\tau)$. La misma conclusión también es compartida por ambas opciones en los casos en que el conjunto de puntos fijos de $(C_\rho \circ C_D)$ es $\{\mathfrak{f}, \mathfrak{v}\}$ (es decir, hay varios puntos fijos pero $\mathfrak{i}$ no es uno de ellos).

La diferencia esencial es que la opción MK tiene un condicionante importante (que no tiene la opción G): cuando $\mathfrak{i}$ no es un punto fijo de $(C_\rho \circ C_D)$ y la aparición de τ en la oración $\rho(D\tau)$ no tiene referencia, debe cumplirse, como se ha visto más arriba, $C_\rho(\mathfrak{i}) = \mathfrak{i}$. En principio esto hace imposible dar solución a la oración $\sim F\tau$, si la referencia habitual de τ es la propia oración —algo fácil de conseguir usando por ejemplo la diagonalización—, F es el predicado "falso" ($C_F(\mathfrak{f}) = \mathfrak{v}$, $C_F(\mathfrak{v}) = C_F(\mathfrak{i}) = \mathfrak{f}$) y $\sim$ es la negación exclusiva (es decir, $C_\sim(\mathfrak{v}) = \mathfrak{f}$, $C_\sim(\mathfrak{f}) = \mathfrak{v}$, $C_\sim(\mathfrak{i}) = \mathfrak{v}$). En este caso particular la función $(C_\rho \circ C_D)$ pasa a ser $(C_\sim \circ C_F)$ que tiene dos puntos fijos: los valores de verdad $\mathfrak{f}$ y $\mathfrak{v}$. En consecuencia, la oración $\sim F\tau$ admite coherentemente cualquiera de esos dos valores de verdad por lo que, según el planteamiento MK, no tiene contenido enunciativo. Pero entonces,

$$\mathcal{V}_\mathfrak{M}(\sim F\tau) = \mathfrak{i} \tag{5.20}$$

y, como $\mathfrak{i}$ no es un punto fijo de $(C_\sim \circ C_F)$, la aparición de τ en $\sim F\tau$ no puede tener referencia. Sin embargo, esto nos lleva a que $\mathcal{V}_\mathfrak{M}(F\tau) = \mathfrak{i}$ y:

$$\mathcal{V}_\mathfrak{M}(\sim F\tau) = C_\sim(\mathcal{V}_\mathfrak{M}(F\tau)) = C_\sim(\mathfrak{i}) = \mathfrak{v} \tag{5.21}$$

lo cual se contradice con (5.20).

Con la opción G la evaluación de la oración $\sim F\tau$ no presenta problemas: como $(C_\sim \circ C_F)$ tiene varios puntos fijos, la aparición de τ en $\sim F\tau$ no tiene referencia y la valuación correcta es (5.21), es decir, la oración es verdadera.

Aunque las posibilidades de la opción MK requieren un estudio más profundo, la opción G nos resulta inicialmente más sencilla y menos problemática por lo que será la que usaremos en el resto de este trabajo.

5.2. Análisis de sistemas de oraciones mediante sistemas de ecuaciones de valores de verdad

En los próximos apartados estudiaremos sistemas de oraciones, paradójicos o no, e intentaremos generalizar el planteamiento G a tales sistemas. Para ello asociaremos a los sistemas de oraciones sistemas de ecuaciones de valores de verdad y determinaremos qué apariciones de términos pierden su referencia habitual teniendo en cuenta el número de soluciones de los sistemas de ecuaciones. Una vez que se sabe qué apariciones de términos pierden su referencia habitual y cuáles no, la evaluación de las oraciones deja de ser problemática.

5.2.1. La paradoja del mentiroso como sistema de dos ecuaciones

Una versión de la paradoja del mentiroso se obtiene mediante el siguiente par de oraciones

(1) (2) es verdadera

(2) (1) no es verdadera

Podríamos formalizar este sistema mediante

$$T\tau_2;\ \neg T\tau_1 \tag{5.22}$$

siempre que, de alguna forma, se establezca que la referencia habitual de τ_1 es la oración $T\tau_2$ y la referencia habitual de τ_2, la oración $\neg T\tau_1$; es decir, $\mathcal{V}^h_{\mathfrak{M}}(\tau_1) = T\tau_2$ y $\mathcal{V}^h_{\mathfrak{M}}(\tau_2) = \neg T\tau_1$.[13] En el supuesto de que τ_1 y τ_2 también tienen su referencia habitual cuando aparecen en las oraciones $T\tau_2$ y $\neg T\tau_1$, tendremos:

$$\left.\begin{array}{l} \mathcal{V}_{\mathfrak{M}}(T\tau_2) = C_T(\mathcal{V}_{\mathfrak{M}}(\neg T\tau_1)) \\ \mathcal{V}_{\mathfrak{M}}(\neg T\tau_1) = C_\neg(\mathcal{V}_{\mathfrak{M}}(T\tau_1)) = C_\neg(C_T(\mathcal{V}_{\mathfrak{M}}(T\tau_2))) \end{array}\right\} \tag{5.23}$$

[13] Véase, por ejemplo, Smullyan (1996, p. 9) o simplemente, establézcase directamente la referencia habitual de cada término en el modelo.

Si llamamos x_1y x_2 a los valores de verdad de las oraciones $T\tau_2$ y $\neg T\tau_1$, respectivamente, basándonos en el sistema anterior, podemos obtener:

$$\left.\begin{array}{l} x_1 = C_T(x_2) \\ x_2 = C_\neg(C_T(x_1)) \end{array}\right\} \tag{5.24}$$

Mientras no se indique lo contrario, la negación y el predicado verdadero los entenderemos de modo que el grafo de $C_\neg$ sea $\{(\mathfrak{f},\mathfrak{v}),(\mathfrak{v},\mathfrak{f}),(\mathfrak{i},\mathfrak{i})\}$ y el de C_T sea $\{(\mathfrak{f},\mathfrak{f}),(\mathfrak{v},\mathfrak{v}),(\mathfrak{i},\mathfrak{f})\}$. Entonces, es fácil comprobar que el sistema de ecuaciones anterior no tiene solución. Por tanto, el supuesto de que τ_1 y τ_2 tienen su referencia habitual cuando aparecen en las oraciones $T\tau_2$ y $\neg T\tau_1$ es falso. Caben entonces tres posibilidades en cuanto a la referencia de τ_1 y τ_2 cuando aparecen en las oraciones $T\tau_2$ y $\neg T\tau_1$:

1. $\mathcal{V}_\mathfrak{M}(\tau_1^{[\neg T\tau_1]}) = T\tau_2$ pero τ_2 carece de referencia ($\tau_2^{[T\tau_2]} \notin dom(\mathcal{V}_\mathfrak{M})$). Entonces:

$$\left.\begin{array}{l} \mathcal{V}_\mathfrak{M}(T\tau_2) = \mathfrak{i} \\ \mathcal{V}_\mathfrak{M}(\neg T\tau_1) = C_\neg(C_T(\mathcal{V}_\mathfrak{M}(T\tau_2))) = C_\neg(C_T(\mathfrak{i})) = \mathfrak{v} \end{array}\right\} \tag{5.25}$$

2. $\mathcal{V}_\mathfrak{M}(\tau_2^{[T\tau_2]}) = \neg T\tau_1$ pero τ_1 carece de referencia ($\tau_1^{[\neg T\tau_1]} \notin dom(\mathcal{V}_\mathfrak{M})$). En tal caso:

$$\left.\begin{array}{l} \mathcal{V}_\mathfrak{M}(\neg T\tau_1) = C_\neg(\mathcal{V}_\mathfrak{M}(T\tau_1)) = C_\neg(\mathfrak{i}) = \mathfrak{i} \\ \mathcal{V}_\mathfrak{M}(T\tau_2) = C_T(\mathcal{V}_\mathfrak{M}(\neg T\tau_1)) = C_T(\mathfrak{i}) = \mathfrak{f} \end{array}\right\} \tag{5.26}$$

3. Tanto τ_1 como τ_2 carecen de referencia ($\tau_2^{[T\tau_2]} \notin dom(\mathcal{V}_\mathfrak{M})$, $\tau_1^{[\neg T\tau_1]} \notin dom(\mathcal{V}_\mathfrak{M})$). Entonces:

$$\left.\begin{array}{l} \mathcal{V}_\mathfrak{M}(T\tau_2) = \mathfrak{i} \\ \mathcal{V}_\mathfrak{M}(\neg T\tau_1) = C_\neg(\mathcal{V}_\mathfrak{M}(T\tau_1)) = C_\neg(\mathfrak{i}) = \mathfrak{i} \end{array}\right\} \tag{5.27}$$

Es claro que no hay mayor motivo para afirmar que τ_1 carece de referencia que para afirmar que τ_2 carece de referencia ni al contrario. Por tanto, las opciones *1* y *2* resultan arbitrarias y no hay razón para preferir la opción *1* a la *2* ni la *2* a la *1*. En cambio, un principio de simetría nos haría decantarnos por la opción *3*. Además, esta opción está más en consonancia con la alternativa G que hemos escogido: recordemos que se evitaba elegir un valor de verdad para una oración cuando no había ningún motivo para preferirlo a otro, lo cual también puede verse como la aplicación de un principio de simetría. También hay que recordar que en algunos ejemplos similares, como el de la tarjeta que por ambas caras tiene una

misma oración (por ejemplo, la oración "lo escrito en la otra cara de esta tarjeta es falso"), la aplicación del principio de simetría es inevitable.

Incluso en el planteamiento MK, la opción *3* es la única posible: ninguna de las dos oraciones de esta paradoja son reducibles a oraciones sin predicados de desentrecomillado, por tanto carecen de contenido enunciativo lo que conlleva que su valor de verdad es $\mathfrak{i}$, algo que solo ocurre en la opción *3*.

5.2.2. La paradoja del veraz como sistema de dos ecuaciones

Una versión de la paradoja del veraz se obtiene mediante el siguiente par de oraciones

(1) (2) es verdadera

(2) (1) es verdadera

Podemos formalizar este sistema mediante

$$T\tau_2;\ T\tau_1 \tag{5.28}$$

siempre que se establezca que la referencia habitual de τ_1 es la oración $T\tau_2$ y la referencia habitual de τ_2, la oración $T\tau_1$. Siguiendo el mismo procedimiento que en el apartado anterior, si llamamos x_1 y x_2 a los valores de verdad de las oraciones $T\tau_2$ y $T\tau_1$, respectivamente, podemos obtener el sistema de ecuaciones:

$$\left.\begin{aligned} x_1 &= C_T(x_2) \\ x_2 &= C_T(x_1) \end{aligned}\right\} \tag{5.29}$$

Este sistema está basado en el supuesto de que τ_1 y τ_2 tienen su referencia habitual cuando aparecen en las oraciones $T\tau_2$ y $T\tau_1$. Pero ahora el sistema no es incompatible sino indeterminado; en concreto, tiene dos soluciones: a) $x_1 = x_2 = \mathfrak{v}$; b) $x_1 = x_2 = \mathfrak{f}$. Sin embargo, no hay motivo para decantarnos por una de ellas a costa de la otra. Tenemos el criterio de exigir que quede determinado inequívocamente un único valor de verdad para toda oración, por lo cual no podemos admitir que τ_1 y τ_2 tengan su referencia habitual cuando aparecen en el sistema

de oraciones (5.28). Finalmente, el criterio de simetría nos lleva a concluir que ni τ_1 ni τ_2 tienen su referencia habitual en ese contexto.[14]

También es claro que, en el planteamiento MK ninguna de las dos oraciones de esta paradoja son reducibles a oraciones sin predicados de desentrecomillado, por tanto carecen de contenido enunciativo lo que conlleva que su valor de verdad es $\mathfrak{i}$, algo que solo ocurre cuando ni τ_1 ni τ_2 tienen su referencia habitual en el contexto del sistema (5.28).

5.2.3. La paradoja de Löb/Curry

La paradoja de Löb puede presentarse mediante la siguiente oración:

(1) Si (1) es verdadera entonces Dios existe

Como sabemos esta paradoja permite demostrar que Dios existe o cualquier otra afirmación que pongamos en su lugar. ¿Cuál es ahora nuestro análisis? La oración es de la forma $(T\tau \rightarrow p)$, donde la referencia habitual de τ es la propia oración y p representa la proposición "Dios existe". El valor de verdad de la oración es $\mathcal{V}_{\mathfrak{M}}((T\tau \rightarrow p)) = C_{\rightarrow}(\mathcal{V}_{\mathfrak{M}}(T\tau), \mathcal{V}_{\mathfrak{M}}(p))$. Ahora bien, si τ tiene aquí su referencia habitual, es decir, si $\mathcal{V}_{\mathfrak{M}}(\tau^{[(T\tau \rightarrow p)]}) = (T\tau \rightarrow p)$, entonces $\mathcal{V}_{\mathfrak{M}}(T\tau) = C_T(\mathcal{V}_{\mathfrak{M}}((T\tau \rightarrow p)))$. Es fácil comprobar que, si llamamos x al valor de verdad de $(T\tau \rightarrow p)$, obtendremos la siguiente ecuación:

$$x = C_{\rightarrow}(C_T(x), \mathcal{V}_{\mathfrak{M}}(p)) \tag{5.30}$$

que solo tiene solución cuando $\mathcal{V}_{\mathfrak{M}}(p) = \mathfrak{v}$ (entonces la solución es $x = \mathfrak{v}$).

Por tanto, nuestra conclusión es que si la referencia habitual de τ es la oración $(T\tau \rightarrow p)$, la oración constituye un contexto anulador de la referencia de τ solo cuando p no es verdadera. Esto simplemente confirma que los motivos por los que algunos términos dejan de tener su referencia habitual en el contexto de ciertas oraciones son de naturaleza semántica e incluso empírica.

Finalmente, es importante observar que si p es verdadera y usamos la lógica trivaluada fuerte de Kleene, $\mathcal{V}_{\mathfrak{M}}((T\tau \rightarrow p)) = \mathfrak{v}$ independientemente de si τ tiene

[14] Una vez más la necesidad de respetar el criterio de simetría es evidente si consideramos una tarjeta que en cada una de sus caras tiene escrita la oración "la oración de la otra cara es verdadera". El problema se formalizaría de igual modo pero la simetría de la situación es más evidente.

o no referencia. Por eso, incluso en el planteamiento MK tendríamos que admitir que $\mathcal{V}_{\mathfrak{M}}((T\tau \to p)) = \mathfrak{v}$.

5.2.4. El ejemplo de Gupta

Tomemos el ejemplo de Gupta (1982, ejemplo (3) en parte IV), simplificado como se indica a continuación:

(a1) (b) es verdadero;

(a2) (b) no es verdadero

(b) (a1) no es verdadero o (a2) no es verdadero

Podemos formalizar las oraciones mediante

$$T\beta;\ \neg T\beta;\ (\neg T\alpha_1 \vee \neg T\alpha_2) \tag{5.31}$$

suponiendo que las referencias habituales de α_1, α_2 y β son, respectivamente, las oraciones $T\beta$, $\neg T\beta$ y $(\neg T\alpha_1 \vee \neg T\alpha_2)$. Si llamamos x, y, z a los valores de verdad de las respectivas oraciones, y suponemos que α_1, α_2 y β tienen su referencia habitual cuando aparecen en ellas, obtenemos el siguiente sistema:

$$\left.\begin{array}{l} x = C_T(z) \\ y = C_\neg(C_T(z)) \\ z = C_\vee(C_\neg(C_T(x)), C_\neg(C_T(y))) \end{array}\right\} \tag{5.32}$$

El sistema tiene una única solución: $z = x = \mathfrak{v}$, $y = \mathfrak{f}$, lo que refleja el análisis informal por el cual Gupta afirmaría que las oraciones (a1) y (b) son verdaderas y (a2) es falsa.

En cambio, con el planteamiento MK, parece bastante claro que ninguna de las oraciones tiene contenido —unas oraciones dicen algo sobre las otras pero ninguna dice nada empírico, ninguna está fundamentada según el concepto de Kripke—. En tal caso, el único valor de verdad que podríamos asignarle a cada oración es$\mathfrak{i}$, lo cual solo es posible si el sistema (5.31) constituye un contexto anulador de la referencia de los términos α_1, α_2 y β.

5.2.5. Análisis general de sistemas finitos de oraciones no cuantificadas

En los siguientes apartados generalizaremos el análisis de los sistemas de oraciones anteriores al caso general de sistemas finitos de oraciones no cuantificadas.

5.2.5.1. ¿Oraciones-tipo u oraciones-caso?

En los lenguajes formales que hemos utilizado, las oraciones han sido tratadas como oraciones-tipo, es decir, no hemos distinguido diversas instancias de una misma oración.

Sin embargo, ejemplos como el siguiente nos llevan a plantear la cuestión de si debe realizarse tal distinción.

Consideremos un aula con dos pizarras, una verde y otra negra. En cada una de ellas lo único escrito es: "la oración escrita en la pizarra negra no es verdadera". Tenemos dos instancias de una misma oración-tipo, es decir, dos oraciones-caso y una única oración-tipo. La oración-caso de la pizarra negra es paradójica: según nuestro análisis, su término sujeto carece de referencia y la oración no es verdadera ni falsa. Si las oraciones-caso solo fueran maneras de ejemplificar una misma oración-tipo, el término sujeto de la oración-caso de la pizarra verde también debería carecer de referencia. Pero no parece que haya ningún motivo independiente que lo justifique. Por el contrario, resulta más natural entender que el contenido de la oración-caso de la pizarra verde es que la oración-caso escrita en la pizarra negra no es verdadera, lo cual, según acabamos de ver, es algo verdadero.

Tenemos pues dos alternativas: a) permitir que distintas oraciones-caso de una misma oración-tipo, puedan tener distintos valores de verdad; b) partir del supuesto de que cuando atribuimos a una oración un valor de verdad lo estamos atribuyendo siempre a una oración-tipo.

Para formalizar el ejemplo anterior es claro que usaríamos una oración-tipo de la forma $\neg T\tau$. En la primera alternativa, podemos distinguir las dos oraciones-caso mediante una etiqueta añadida a la oración-tipo. Por ejemplo, la oración-caso de la pizarra verde podría representarse mediante $[verde; \neg T\tau]$ y la de la pizarra negra mediante $[negra; \neg T\tau]$ siempre y cuando la referencia habitual del término τ sea la oración-caso $[negra; \neg T\tau]$. Ahora, la función de valuación, $\mathcal{V}_{\mathfrak{M}}$, no la podríamos aplicar a la oración-tipo, $\neg T\tau$, pero sí a las oraciones-caso. Según nuestro análisis informal, debe cumplirse: $\mathcal{V}_{\mathfrak{M}}([negra; \neg T\tau]) = \mathfrak{i}$ y $\mathcal{V}_{\mathfrak{M}}([verde; \neg T\tau]) = \mathfrak{v}$.

Para referirnos a una suboración de una oración-caso, podemos utilizar la misma etiqueta que para la oración-caso y una coma (en vez de un punto y coma) para separar la etiqueta de la suboración. Por ejemplo, $[verde, T\tau]$ representa la suboración $T\tau$ dentro de la oración-caso $[verde; \neg T\tau]$. Así podemos expresar cómo se calcula el valor de verdad de una oración-caso en función del de sus suboraciones. Para la oración-caso $[negra; \neg T\tau]$ tendríamos:

$$\mathcal{V}_{\mathfrak{M}}([negra; \neg T\tau]) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}([negra, T\tau])) = C_{\neg}(\mathfrak{i}) = \mathfrak{i} \tag{5.33}$$

donde hemos tenido en cuenta que el término τ carece de referencia dentro de la oración-caso $[negra; \neg T\tau]$, y por tanto dentro de la suboración $[negra, T\tau]$, lo que hace que $\mathcal{V}_{\mathfrak{M}}([negra, T\tau]) = \mathfrak{i}$. La evaluación de la oración-caso $[verde; \neg T\tau]$ será:

$$\begin{aligned} &\mathcal{V}_{\mathfrak{M}}([verde; \neg T\tau]) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}([verde, T\tau])) = \\ &= C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}([negra; \neg T\tau]))) = C_{\neg}(C_T((\mathfrak{i})) = \mathfrak{v} \end{aligned} \tag{5.34}$$

Aquí el término τ dentro de la suboración $[verde, T\tau]$ tiene su referencia habitual: la oración-caso $[negra; \neg T\tau]$. Por eso, $\mathcal{V}_{\mathfrak{M}}([verde, T\tau]) = C_T(\mathcal{V}_{\mathfrak{M}}([negra; \neg T\tau]))$.

Este etiquetado de oraciones añade sin duda complejidad a la sintaxis y semántica del lenguaje formal, aunque afortunadamente la distinción entre instancias de una misma oración-tipo solo será necesaria en aquellas oraciones-tipo que contengan términos susceptibles de perder su referencia habitual, puesto que, como en el ejemplo anterior, podría haber oraciones-caso donde un término perdiera la referencia habitual y otras donde no. Como ejemplo de oración-tipo donde no es necesaria (para nuestro propósito) la distinción entre sus diversas instancias, sirve cualquier oración del lenguaje formal, φ, carente de predicados de desentrecomillado. Si $[1; \varphi]$ y $[2; \varphi]$ fuesen dos instancias cualesquiera de la oración φ, entonces $\mathcal{V}_{\mathfrak{M}}([1; \varphi]) = \mathcal{V}_{\mathfrak{M}}([2; \varphi])$. Podríamos pues, referirnos al valor de verdad de cualquier instancia de la oración φ, sin especificar una concreta, escribiendo $\mathcal{V}_{\mathfrak{M}}([*; \varphi])$ —el asterisco se interpretaría como "cualquier instancia de"— o, incluso, simplemente escribiendo $\mathcal{V}_{\mathfrak{M}}(\varphi)$.

A pesar de que, posiblemente, el "problema de las dos pizarras" sugiera la conveniencia de permitir que distintas oraciones-caso de una misma oración-tipo puedan tener distintos valores de verdad; también puede abordarse partiendo del supuesto de que cuando atribuimos a una oración un valor de verdad lo estamos atribuyendo siempre a una oración-tipo. En efecto, llamemos simplemente oración

a la oración-tipo. Las oraciones-caso de cada pizarra serán simplemente instancias de una misma oración cuya formalización será $\neg T\tau$ —siempre que la referencia habitual del término τ sea la propia oración $\neg T\tau$—. Y, como sabemos, según nuestro análisis, la aparición del término τ en la oración $\neg T\tau$ carecerá de referencia. Desde el punto de vista informal, en las dos pizarras hay una instancia de una misma oración: no hay dos oraciones y, por eso, no tiene sentido decir que una es paradójica y otra no o que en una el término sujeto tiene referencia y en otra no. Solo hay una oración y es paradójica.

Como sabemos algunas propuestas de solución presentan el problema de que las propias conclusiones no pueden expresarse en el lenguaje objeto sino en un metalenguaje donde las paradojas siguen sin resolverse. Cabe plantearse si con esta segunda alternativa se pierde la posibilidad de expresar, en el propio lenguaje formal, que la oración $\neg T\tau$ no es verdadera. Si fuera así, no podría formar parte de una solución satisfactoria. Pero afortunadamente, es muy sencillo expresar, en el propio lenguaje formal, que la oración $\neg T\tau$ no es verdadera: basta con usar la oración $\neg T\ulcorner\neg T\tau\urcorner$. La clave está en usar términos de cita puesto que sabemos que nunca pierden su referencia.

En este trabajo, continuaremos con el supuesto de que cuando atribuimos a una oración un valor de verdad lo estamos atribuyendo siempre a una oración-tipo. Pero aunque no desarrollaremos la evaluación de oraciones desde el planteamiento de permitir que distintas oraciones-caso de una misma oración-tipo, puedan tener distintos valores de verdad, sí haremos comentarios puntuales al respecto cuando resulten relevantes.

5.2.5.2. La herramienta de los grafos

Nos valdremos del uso de grafos orientados para el análisis de sistemas de oraciones. Revisemos brevemente parte de la nomenclatura habitual para este tipo de grafos.

Un grafo (orientado) es un conjunto de pares (ordenados) de elementos. Cada elemento es un *vértice* del grafo. Cada par ordenado de vértices es un *arco*. Un *camino* es una secuencia de arcos en la que el vértice final de cualquier arco es el mismo que el vértice inicial del siguiente. Un *ciclo* es un camino en el que el vértice inicial del primer arco es el mismo que el final del último arco.

Un grafo es *fuertemente conexo* ssi para todo par de vértices distintos (u, v) existe un camino de u a v y un camino de v a u.

Un grafo orientado es *conexo* ssi se convierte en fuertemente conexo cuando para cada arco (u, v) del grafo añadimos, si no existía, el arco (v, u).

Dado un conjunto, Θ, al establecer en él una relación binaria, R,[15] se establece un grafo que designaremos como $G(R, \Theta)$:

$$G(R, \Theta) =_{def} \{(u, v)/(u, v) \in \Theta \times \Theta \wedge (uRv)\} \tag{5.35}$$

En el conjunto de vértices, Θ, de un grafo orientado, $G(R, \Theta)$, podemos establecer la relación de equivalencia $\sim_R$: dos vértices, u y v, están relacionados, $u \sim_R v$, solo cuando son iguales o, en $G(R, \Theta)$, hay un camino de u a v y un camino de v a u. Es casi trivial comprobar que la relación $\sim_R$ es de equivalencia, es decir, cumple las propiedades reflexiva, simétrica y transitiva. Como toda relación de equivalencia establece una partición del conjunto Θ en clases de equivalencia. Representamos la clase de equivalencia del elemento u mediante $[u]_{\sim_R}$ y el conjunto de todas las clases de equivalencia mediante $\Theta/\sim_R$. Lo interesante es que cada uno de los subgrafos $G(R, [u]_{\sim_R})$ es un grafo fuertemente conexo. En efecto, si u_1y u_2 son dos elementos distintos pertenecientes a $[u]_{\sim_R}$, se cumple, $u_1 \sim_R u_2$, es decir, en $G(R, \Theta)$, hay un camino de u_1 a u_2 y de u_2 a u_1. Por la propiedad transitiva de $\sim_R$, los vértices de estos caminos pertenecen todos a $[u]_{\sim_R}$luego, podemos concluir que en $G(R, [u]_{\sim_R})$ hay un camino de u_1 a u_2 y de u_2 a u_1, es decir, que $G(R, [u]_{\sim_R})$ es un grafo fuertemente conexo.

Sobre el conjunto de clases de equivalencia, $\Theta/\sim_R$, definimos la siguiente relación binaria:

$$[u]_{\sim_R} \prec_R [v]_{\sim_R} \quad ssi \quad [u]_{\sim_R} \neq [v]_{\sim_R} \wedge \exists\, a, b\, (a \in [u]_{\sim_R} \wedge b \in [v]_{\sim_R} \wedge aRb) \tag{5.36}$$

Es obvio que se trata de una relación irreflexiva.[16] También es fácil ver que es asimétrica. En efecto, cuando $[u]_{\sim_R} \prec_R [v]_{\sim_R}$ existen a, b tales que $a \in [u]_{\sim_R}$, $b \in [v]_{\sim_R}$ y aRb; por tanto, hay un camino en el grafo $G(R, \Theta)$ desde cualquier elemento de $[u]_{\sim_R}$ hasta a (dado que $u \sim_R a$), un arco de a a b (dado que aRb) y un camino desde b a cualquier elemento de $[v]_{\sim_R}$. En definitiva, si $[u]_{\sim_R} \prec_R [v]_{\sim_R}$ hay en el grafo $G(R, \Theta)$ un camino desde cualquier elemento de $[u]_{\sim_R}$ a cualquier elemento de $[v]_{\sim_R}$. Por tanto, no puede ser $[v]_{\sim_R} \prec_R [u]_{\sim_R}$ ya que entonces también habría un camino en el grafo $G(R, \Theta)$ entre, por ejemplo, v y u. Al haber un camino

[15] Ver apéndice A, apartado A.3.

[16] La nomenclatura usada acerca de relaciones binarias está en el apéndice A, apartado A.3.

entre u y v y otro entre v y u, se cumpliría $u \sim_R v$ y, por tanto, $[u]_{\sim_R} = [v]_{\sim_R}$. Esto último contradice la hipótesis de que $[u]_{\sim_R} \prec_R [v]_{\sim_R}$.

5.2.5.3. Definiciones

Con objeto de poder establecer un método general de análisis de sistemas de oraciones es conveniente establecer algunas definiciones.

Definición 5.1 (referencia semántica). *La expresión cerrada α contiene una referencia semántica aparente a la expresión cerrada β, mediante el término cerrado τ, si y solo si en α hay una aparición de una expresión de la forma $D\tau$, donde D es un predicado o una función de desentrecomillado, la aparición de τ en $D\tau$ es referencial respecto a α*[17] *y la referencia habitual de τ es la expresión β. Si τ es un término de cita o la aparición de τ en $D\tau$ no está en un contexto referencialmente anulador, la expresión α contiene una referencia semántica (real) a la expresión β (mediante el término τ).*

Nótese que las referencias semánticas reales son un caso particular de las referencias semánticas aparentes.

Definición 5.2 (referencia semántica sin término de cita). *Entre dos expresiones formales, α y β, de un lenguaje interpretado se da la relación S, $\alpha S \beta$, si y solo si α contiene una referencia semántica aparente a la expresión β mediante un término que no es de cita.*

Aunque las definiciones anteriores son aplicables a expresiones formales en general —sean oraciones o términos cerrados—, en lo sucesivo, ignoraremos los términos que tienen referencias semánticas aparentes a otros términos y nos limitaremos a las oraciones que tienen referencias semánticas aparentes a otras oraciones.

La importancia de la relación S es que, en el supuesto de que los términos, no de cita, mediante los que una oración contiene referencias semánticas aparentes a determinadas oraciones, tengan su referencia habitual, el valor de verdad de la oración puede expresarse en función de los valores de verdad de esas oraciones y de ninguna otra. Por ejemplo, consideremos la oración φ_1:

$$(\neg T\tau_1 \vee T^{\ulcorner} F\tau_2 ^{\urcorner}) \tag{5.37}$$

[17] Es decir, al sustituir la aparición de τ por otro término con la misma referencia, α se convierte en otra expresión, α', con la misma referencia que α. Véase definición 4.10 en p. 79.

y supongamos que T y F son predicados de desentrecomillado y las referencias habituales de los términos cerrados τ_1 y τ_2 son respectivamente la oración φ_1 y una oración φ_2. Entonces tenemos $\varphi_1 S \varphi_1$ y $\varphi_1 S \varphi_2$ lo que indicará que el valor de verdad de φ_1 puede expresarse en función de los valores de verdad de φ_1 y de φ_2 (y de ninguna otra oración). En efecto, si abreviamos los valores de verdad φ_1 y φ_2 mediante las incógnitas x_1 y x_2, respectivamente, basta aplicar la noción de predicado de desentrecomillado y la composicionalidad para obtener:

$$x_1 = C_\vee(C_\neg(C_T(x_1)), C_T(C_F(x_2))) \tag{5.38}$$

El valor de verdad de φ_1 también puede descomponerse de otras formas, como

$$\mathcal{V}_\mathfrak{M}(\varphi_1) = C_\vee(\mathcal{V}_\mathfrak{M}(\neg T\tau_1), \mathcal{V}_\mathfrak{M}(T^\ulcorner F\tau_2{}^\urcorner)) \tag{5.39}$$

donde un argumento de la función $C_\vee$ es el valor de verdad de la oración $\neg T\tau_1$, pero en el supuesto de que τ_1 conserva su referencia habitual, el valor de verdad de $\neg T\tau_1$ depende, en última instancia del de φ_1.

Definición 5.3 (conjunto de oraciones S-cerrado). *Un conjunto de oraciones Φ de un lenguaje es S-cerrado si y solo si no existen dos oraciones del lenguaje φ y ψ tales que $(\varphi \in \Phi) \wedge (\varphi S \psi) \wedge (\psi \notin \Phi)$.*

En principio, un sistema de oraciones está incompleto si el conjunto de oraciones no es S-cerrado, porque, en el supuesto de que todos los términos tengan su referencia habitual, el valor de verdad de algunas oraciones dependerá del de otras que no están en el sistema. Así pues, ante un conjunto de oraciones que no es S-cerrado, hay que añadir las que faltan para que lo sea. Solo entonces debemos proceder a evaluarlas. Siguiendo con el ejemplo, el hecho de que $\varphi_1 S \varphi_2$ indica que el conjunto $\{\varphi_1\}$ no es S-cerrado y, al mismo tiempo, que no se puede evaluar φ_1 aisladamente sino como parte de un sistema de oraciones al que también debe pertenecer la oración φ_2.

Definición 5.4 (oración fundamentada). *Una oración φ perteneciente a un conjunto de oraciones S-cerrado, Φ, está S-fundamentada si y solo si no hay un camino infinito a partir de φ en el grafo $G(S, \Phi)$.*

Se pueden distinguir dos clases de caminos infinitos en un grafo: los que pasan por un número finito de vértices y los que contienen un número infinito de vértices.

Obsérvese que en un grafo en el que hay un camino circular también hay un camino infinito —obtenido repitiendo infinitas veces el camino circular—.

5.2.5.4. En busca de un procedimiento de evaluación

Observando ahora algunos de los ejemplos estudiados (el ejemplo de Gupta o las paradojas del veraz y del mentiroso como sistemas de oraciones), comprobamos que para calcular el valor de verdad de sus oraciones —con el planteamiento G— nos hemos valido de: 1/ el sistema de ecuaciones de valores de verdad obtenido bajo el supuesto de que todos los términos tienen su referencia habitual; 2/ un principio de simetría por el que si el sistema de ecuaciones tiene una única solución todos los términos tienen su referencia habitual, pero, en caso contrario, ninguna de las referencias semánticas aparentes a oraciones del sistema es real (salvo las realizadas mediante términos de cita que nunca pueden dejar de ser reales).

Una vez que se sabe qué apariciones de términos pierden su referencia habitual y cuáles no, la evaluación de las oraciones deja de ser problemática.

Estas observaciones serán fundamentales para resolver nuestro problema: dado un conjunto finito de oraciones no cuantificadas, Φ, S-cerrado, evaluar sus oraciones —desde el punto de vista que hemos dado en llamar G—.

Hemos visto en el apartado 5.2.5.2 que, a partir de la relación S, podemos definir sobre Φ la relación de equivalencia $\sim_S$: dos oraciones, u y v, están relacionadas, $u \sim_S v$, solo cuando son iguales o, en $G(S, \Phi)$, hay un camino de u a v y un camino de v a u. Es fácil comprobar que en cualquiera de los ejemplos antes mencionados (el ejemplo de Gupta o las paradojas del veraz y del mentiroso como sistemas de oraciones) la relación $\sim_S$ solo produce una clase de equivalencia. Evidentemente puede haber conjuntos de oraciones donde la relación $\sim_S$ genere varias clases de equivalencia. Por tanto, la forma en que hemos solucionado aquellos ejemplos no es generalizable sin más a cualquier sistema de oraciones.

También hemos visto en el apartado 5.2.5.2 que, sobre el conjunto de clases de equivalencia $\Phi/\sim_S$, podemos definir una relación binaria $\prec_S$. Recordemos que

$$[u]_{\sim_S} \prec_S [v]_{\sim_S} \quad ssi \quad [u]_{\sim_S} \neq [v]_{\sim_S} \wedge \exists\, a, b\, (a \in [u]_{\sim_S} \wedge b \in [v]_{\sim_S} \wedge aSb) \quad (5.40)$$

debido a lo cual, cuando $[u]_{\sim_S} \prec_S [v]_{\sim_S}$ existen dos oraciones, a y b, tales que aSb, $a \in [u]_{\sim_S}$ y $b \in [v]_{\sim_S}$. Consiguientemente, cuando $[u]_{\sim_S} \prec_S [v]_{\sim_S}$ —y en el supuesto de que todos las referencias semánticas aparentes a oraciones de $[v]_{\sim_S}$

sean reales—, al menos una oración de $[u]_{\sim_S}$ —la oración a— tiene un valor de verdad que depende del de una oración de $[v]_{\sim_S}$ —la oración b—. Sin embargo, el valor de verdad de cualquier oración de $[v]_{\sim_S}$ puede expresarse de modo independiente del valor de verdad de cualquier oración de $[u]_{\sim_S}$ debido a que, en el grafo $G(S,\Phi)$, no puede haber ningún camino entre una oración de $[v]_{\sim_S}$ y una de $[u]_{\sim_S}$ —si lo hubiera, $[u]_{\sim_S}$ y $[v]_{\sim_S}$ serían la misma clase porque ya hay un camino, que pasa por (a,b), entre cualquier oración de $[u]_{\sim_S}$ y cualquier oración de $[v]_{\sim_S}$—.

La conclusión es sencilla: cuando $[u]_{\sim_S} \prec_S [v]_{\sim_S}$ deben evaluarse las oraciones de $[v]_{\sim_S}$ antes que las de $[u]_{\sim_S}$ y al evaluar estas ha de tenerse en cuenta como se evaluaron las de $[v]_{\sim_S}$.

La siguiente cuestión es cómo evaluar las oraciones de una cualquiera de las clases de equivalencia $[u]_{\sim_S}$. Aquí sí podemos aplicar el procedimiento seguido en ejemplos como el de Gupta en el que todas las oraciones formaban una única clase de equivalencia. Pero con una salvedad: en general hay que tener en cuenta cómo se han evaluado las oraciones de las clases de equivalencia cuyas oraciones se debían evaluar antes.

Consideremos el siguiente ejemplo:

$$\left.\begin{array}{ll} (\varphi_1) & (T\tau_3 \vee T^{\ulcorner}\neg T\tau_1{}^{\urcorner}) \\ (\varphi_2) & \neg T\tau_1 \\ (\varphi_3) & (\neg T\tau_3 \wedge T\tau_1) \\ (\varphi_4) & H\tau_1 \end{array}\right\} \tag{5.41}$$

con los siguientes supuestos: $1/$ φ_1 a φ_4 son nombres con los que en el metalenguaje identificamos las oraciones; $2/$ la referencia habitual de τ_i, es la oración φ_i, para $1 \leq i \leq 4$; $3/$ la interpretación de las conectivas lógicas es la de la lógica trivaluada fuerte de Kleene; $4/$ los predicados T y H son predicados de desentrecomillado cuyas funciones asociadas C_T y C_H, respectivamente, vienen dadas por $C_T(\mathfrak{f}) = \mathfrak{f}$, $C_T(\mathfrak{i}) = \mathfrak{f}$, $C_T(\mathfrak{v}) = \mathfrak{v}$, $C_H(\mathfrak{f}) = \mathfrak{v}$, $C_H(\mathfrak{i}) = \mathfrak{v}$, $C_H(\mathfrak{v}) = \mathfrak{f}$.

Entonces $\Phi = \{\varphi_1, \varphi_2, \varphi_3, \varphi_4\}$, $\varphi_1 S \varphi_1$, porque φ_1 tiene una referencia semántica aparente a φ_1 mediante τ_1; $\varphi_1 S \varphi_3$, porque φ_1 tiene una referencia semántica aparente a φ_3 mediante τ_3, etc. Así podemos comprobar que:

$$G(S,\Phi) = \{(\varphi_1,\varphi_1),(\varphi_1,\varphi_3),(\varphi_2,\varphi_1),(\varphi_3,\varphi_1),(\varphi_3,\varphi_3),(\varphi_4,\varphi_1)\} \tag{5.42}$$

También vemos que $\varphi_1 \sim_S \varphi_3$ porque en el grafo anterior hay un camino de φ_1 a φ_3 y viceversa. Si seguimos analizando el grafo $G(S, \Phi)$ veremos que las clases de equivalencia formadas por la relación $\sim_S$ son $\{\varphi_1, \varphi_3\}$, $\{\varphi_2\}$ y $\{\varphi_4\}$, es decir:

$$\Phi/\sim_S = \{\{\varphi_1, \varphi_3\}, \{\varphi_2\}, \{\varphi_4\}\} \tag{5.43}$$

Al establecer en $\Phi/\sim_S$ la relación $\prec_S$ tenemos que $\{\varphi_2\} \prec_S \{\varphi_1, \varphi_3\}$ porque $\varphi_2 S \varphi_1$ y $\{\varphi_4\} \prec_S \{\varphi_1, \varphi_3\}$ porque $\varphi_4 S \varphi_1$. Por tanto debemos evaluar primero las oraciones de la clase de equivalencia $\{\varphi_1, \varphi_3\}$ y, después, las de $\{\varphi_2\}$ y $\{\varphi_4\}$ —no importa en qué orden—. Para evaluar las oraciones de $\{\varphi_1, \varphi_3\}$, obtenemos su sistema de ecuaciones de valores de verdad (bajo el supuesto de que τ_1 y τ_3 tienen su referencia habitual). Si los valores de verdad que tendrían φ_1 y φ_3, respectivamente, en caso de que τ_1 y τ_3 no perdieran su referencia, los representamos mediante las incógnitas x_1 y x_3, el sistema de ecuaciones de valores de verdad que podemos obtener, teniendo en cuenta la composicionalidad y que T es un predicado de desentrecomillado, es:

$$\left.\begin{array}{l} x_1 = C_\vee(C_T(x_3), C_T(C_\neg(C_T(x_1)))) \\ x_3 = C_\wedge(C_\neg(C_T(x_3)), C_T(x_1)) \end{array}\right\} \tag{5.44}$$

Se trata de un sistema sin solución, por lo que la aplicación del principio de simetría nos lleva a concluir que las apariciones de los términos τ_1 y τ_3 en las oraciones φ_1 y φ_3 carecen de referencia. Esta conclusión nos permite evaluar las oraciones φ_1 y φ_3:

$$\begin{array}{l} \mathcal{V}_\mathfrak{M}(\varphi_1) = \mathcal{V}_\mathfrak{M}((T\tau_3 \vee T^\ulcorner \neg T\tau_1 ^\urcorner)) = C_\vee(\mathcal{V}_\mathfrak{M}(T\tau_3), C_T(\mathcal{V}_\mathfrak{M}(\neg T\tau_1))) = \\ = C_\vee(\mathcal{V}_\mathfrak{M}(T\tau_3), C_T(C_\neg(\mathcal{V}_\mathfrak{M}(T\tau_1)))) = C_\vee(\mathfrak{i}, C_T(C_\neg(\mathfrak{i}))) = C_\vee(\mathfrak{i}, \mathfrak{f}) = \mathfrak{i} \end{array} \tag{5.45}$$

$$\mathcal{V}_\mathfrak{M}(\varphi_3) = C_\wedge(C_\neg(\mathcal{V}_\mathfrak{M}(T\tau_3)), \mathcal{V}_\mathfrak{M}(T\tau_1)) = C_\wedge(C_\neg(\mathfrak{i}), \mathfrak{i}) = C_\wedge(\mathfrak{i}, \mathfrak{i}) = \mathfrak{i} \tag{5.46}$$

El siguiente paso es evaluar las oraciones de $\{\varphi_2\}$ y de $\{\varphi_4\}$ teniendo en cuenta cómo se han evaluado las de $\{\varphi_1, \varphi_3\}$. Para ello debemos responder la siguiente cuestión: ¿pierde su referencia habitual la aparición del término τ_1 en las oraciones φ_2 y φ_4?

Veremos que la respuesta a la pregunta puede ser diferente si distinguimos, o no, diversas instancias de una misma oración. Si hacemos esa distinción, es decir, si distinguimos oraciones-caso, no hay motivo para que el término τ_1 pierda su

referencia en las oraciones φ_2 y φ_4. La evaluación de φ_2 y φ_4 sería:

$$\mathcal{V}_{\mathfrak{M}}(\varphi_2) = C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\varphi_1))) = C_{\neg}(C_T(\mathfrak{i})) = C_{\neg}(\mathfrak{f}) = \mathfrak{v} \tag{5.47}$$

$$\mathcal{V}_{\mathfrak{M}}(\varphi_4) = C_H(\mathcal{V}_{\mathfrak{M}}(\varphi_1)) = C_H(\mathfrak{i}) = \mathfrak{v} \tag{5.48}$$

(por simplicidad, no usamos las notaciones introducidas en el apartado 5.2.5.1).

En cambio, si los valores de verdad no los atribuimos a oraciones-caso sino a oraciones-tipo, el término τ_1 carecerá de referencia en la oración φ_2. Si no fuera así tendríamos $\mathcal{V}_{\mathfrak{M}}(\varphi_2) = \mathcal{V}_{\mathfrak{M}}(\neg T\tau_1) = \mathfrak{v}$ como en (5.47), al tiempo que, dada la forma de φ_1, $\mathcal{V}_{\mathfrak{M}}(\varphi_1) = C_{\vee}(\mathcal{V}_{\mathfrak{M}}(T\tau_3), C_T(\mathcal{V}_{\mathfrak{M}}(\neg T\tau_1)))$. Pero entonces,

$$\mathcal{V}_{\mathfrak{M}}(\varphi_1) = C_{\vee}(\mathcal{V}_{\mathfrak{M}}(T\tau_3), C_T(\mathfrak{v})) = \mathfrak{v} \tag{5.49}$$

lo que contradice a (5.45). El motivo es que al evaluar $\neg T\tau_1$ en (5.45) como parte de la oración φ_1 hemos considerado que τ_1 carece de referencia, por lo que no podemos considerar lo contrario en ninguna otra evaluación de $\neg T\tau_1$ (ya que se trata de la misma oración-tipo).

Así pues, si los valores de verdad los atribuimos a oraciones-tipo, la evaluación correcta de φ_2 es:

$$\mathcal{V}_{\mathfrak{M}}(\varphi_2) = \mathcal{V}_{\mathfrak{M}}(\neg T\tau_1) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(T\tau_1)) = C_{\neg}(\mathfrak{i}) = \mathfrak{i} \tag{5.50}$$

Finalmente, ¿debe perder la aparición de τ_1 en la oración φ_4 su referencia habitual? En principio, no, porque la oración $H\tau_1$ no es una parte propia de φ_1 ni de φ_3. Pero conviene señalar una consecuencia de ello. La evaluación de $H\tau_1$ será

$$\mathcal{V}_{\mathfrak{M}}(H\tau_1) = C_H(\mathcal{V}_{\mathfrak{M}}(\varphi_1) = C_H(\mathfrak{i}) = \mathfrak{v} \tag{5.51}$$

por lo que $\mathcal{V}_{\mathfrak{M}}(H\tau_1) \neq \mathcal{V}_{\mathfrak{M}}(\neg T\tau_1)$. Este ejemplo muestra que, a pesar de que C_H es la misma función que $(C_{\neg} \circ C_T)$, no se podrá sustituir H por $\neg T$ en una oración con garantías de que conserve su valor de verdad.

5.2.5.5. Procedimiento de evaluación

El estudio del apartado anterior nos permite explicitar el siguiente procedimiento de evaluación de sistemas finitos de oraciones no cuantificadas. En la

evaluación de las oraciones supondremos que no distinguimos entre diferentes instancias de una misma oración.

1. Añadir las oraciones necesarias para formar un conjunto de oraciones S-cerrado, Φ.

2. Establecer el orden en que se recorrerán los elementos de $\Phi/\sim_S$ con el siguiente criterio: empezar por los vértices (clases de equivalencia establecidas por la relación $\sim_S$ en el conjunto de oraciones Φ) que no tienen sucesor en el grafo $G(\prec_S, \Phi/\sim_S)$ y continuar en sentido inverso al que establece la relación $\prec_S$.

3. Recorrer las clases de equivalencia de $\Phi/\sim_S$ en el orden establecido para evaluar sus oraciones. Con cada clase de equivalencia, $[u]_{\sim_S}$ hay que hacer lo siguiente:

   a) Hallar el sistema de ecuaciones de valores de verdad asociado:

      1) Sean $\varphi_1, \varphi_2, ..., \varphi_n$ las oraciones de $[u]_{\sim_S}$. Al valor de verdad de cada oración, φ_i, $(1 \leq i \leq n)$, se asignará una incógnita, x_i y una ecuación de valores de verdad de la forma $x_i = ...$

      2) Sean $A_i = \{\psi/\psi \in [u]_{\sim_S} \wedge \varphi_i S \psi\}$, $B_i = \{\psi/\psi \notin [u]_{\sim_S} \wedge \varphi_i S \psi\}$. Para obtener el segundo miembro de la ecuación de valores de verdad:

         a' Téngase en cuenta que la aparición de un término en φ_i pierde su referencia cuando se da en una oración (igual a φ_i o a una parte propia de φ_i) que ya ha sido evaluada (una oración del sistema o una parte propia de una oración ya evaluada) con el resultado de que la aparición del término perdía su referencia.

         b' Supóngase que ningún otro término de φ_i pierde su referencia y aplíquese a la oración la composicionalidad y la definición de los predicados de desentrecomillado hasta que su valor de verdad quede expresado exclusivamente en función de los valores de verdad de oraciones ya evaluadas y de los valores de verdad de oraciones de A_i.[18]

[18] Conseguir expresar el valor de verdad de φ_i en la forma indicada siempre es posible dado que: a) los términos de φ_i que pierden su referencia forman parte de oraciones ya evaluadas; b) las oraciones para las que φ_i contiene alguna referencia semántica mediante términos no

Obsérvese que si llamamos $\varphi_{i_1}, \varphi_{i_2}, ..., \varphi_{i_k}$ a los elementos de A_i, la ecuación asociada a la oración φ_i tendrá la forma

$$x_i = f_i(x_{i_1}, x_{i_2}, ..., x_{i_k}) \tag{5.52}$$

b) Aplicar el principio de simetría para resolver las referencias semánticas aparentes (es decir, decidir si son reales o no): si el sistema de ecuaciones de valores de verdad tiene una única solución, todas las referencias semánticas aparentes a oraciones de $[u]_{\sim_S}$ (contenidas en oraciones de $[u]_{\sim_S}$) son reales; en caso contrario, de esas referencias, solo son reales las realizadas mediante términos de cita.[19]

c) Evaluar las oraciones de $[u]_{\sim_S}$. Si el sistema de ecuaciones de valores de verdad tuvo solución, la evaluación está hecha. Si no, ya sabemos qué referencias semánticas aparentes sin término de cita son reales y cuales no y, en el caso de las reales, conocemos el valor de verdad de las oraciones referidas (serán oraciones ya evaluadas). Por tanto, siempre podremos evaluar las oraciones de $[u]_{\sim_S}$.

En el apéndice B, con objeto de facilitar la comprensión del procedimiento de evaluación, se aplica a un ejemplo.

5.3. Cuantificación y autorreferencia

5.3.1. Introducción

Una oración autorreferencial no cuantificada tiene un término propio cuya referencia es una oración fuertemente equivalente a la oración autorreferencial en cuestión. Otra forma de autorreferencia se tiene en oraciones cuantificadas.

de cita son oraciones de $A_i \cup B_i$; c) las oraciones de B_i ya han sido evaluadas (si $\psi \in B_i$, entonces $[u]_{\sim_S} \prec_S [\psi]_{\sim_S}$ porque $\varphi_i S \psi$ y $[\varphi_i]_{\sim_S} = [u]_{\sim_S}$; luego las oraciones de $[\psi]_{\sim_S}$ ya han sido evaluadas).

[19] Encontrar el número de soluciones del sistema de ecuaciones de valores de verdad siempre es posible porque cada incógnita solo admite un número finito de valores (tres) y hay un número finito de incógnitas.

Tomemos por ejemplo la oración $\forall x\, Px$. Su valor de verdad es[20]

$$\mathcal{V}_{\mathfrak{M}}(\forall x\, Px) =_{def} C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(Px)/e \in\mid \mathfrak{M} \mid\})$$

Pero si la oración está en el universo del modelo (lo que ocurre en los lenguajes interpretados con capacidad de cita), es decir, si $(\forall x\, Px) \in\mid \mathfrak{M} \mid$, entonces $\mathcal{V}_{\mathfrak{M},s[(\forall x Px)/x]}(Px) = \mathcal{V}_{\mathfrak{M}}(P^{\ulcorner}\forall x\, Px^{\urcorner})$, luego:

$$\mathcal{V}_{\mathfrak{M}}(P^{\ulcorner}\forall x\, Px^{\urcorner}) \in \{\mathcal{V}_{\mathfrak{M},s[e/x]}(Px)/e \in\mid \mathfrak{M} \mid\}$$

Por tanto, la interpretación de $P^{\ulcorner}\forall x\, Px^{\urcorner}$ interviene en la interpretación de $\forall x\, Px$. En particular, si como es habitual, la interpretación del cuantificador universal se entiende como una conjunción generalizada, la oración $\forall x\, Px$ resulta ser fuertemente equivalente a $P^{\ulcorner}\forall x\, Px^{\urcorner} \wedge \forall x(x \neq {\ulcorner}\forall x\, Px^{\urcorner} \rightarrow Px)$, es decir, en un sentido extensional, afirma, entre otras cosas, que ella misma tiene la propiedad P.

Otro punto de vista, es que toda variable ligada hace referencia a todo elemento del universo $\mid \mathfrak{M} \mid$, por lo que si la oración cuantificada de la que forma parte es un elemento de $\mid \mathfrak{M} \mid$, la oración hace referencia a sí misma, es autorreferencial. Como, en los lenguajes con capacidad de cita, todas las oraciones pertenecen a $\mid \mathfrak{M} \mid$, en estos lenguajes, todas las oraciones de la forma $\forall x\varphi(x)$ o $\exists x\varphi(x)$ son autorreferenciales. Este conclusión está recogida en la definición 4.30 (p. 130) de oración autorreferencial.[21]

Cuando en $\varphi(x)$ aparecen predicados de desentrecomillado podemos tener oraciones cuantificadas paradójicas, como es el caso de la oración de Epiménides.

5.3.2. Paradoja de Epiménides

A Epiménides el cretense se atribuye la afirmación "todos los cretenses son mentirosos" (debiéndose entender por mentiroso a aquel que nunca afirma nada verdadero). Para formalizar la afirmación de Epiménides supondremos que el pre-

[20] Por simplicidad expositiva, en el apartado 5.3, utilizo un lenguaje trivaluado extensional en lugar de un lenguaje enunciativo trivaluado. Recordemos que el problema de la paradoja del mentiroso y similares se describe en los mismos términos independientemente de si usamos un lenguaje enunciativo o uno puramente extensional.

[21] De acuerdo con la definición también son autorreferenciales oraciones de las que$\forall x\varphi(x)$ o $\exists x\varphi(x)$ es una suboración. Un ejemplo, $(p \wedge \exists x\varphi(x))$. La justificación de que estas oraciones son autorreferenciales es muy similar: la oración —por ejemplo$(p \wedge \exists x\varphi(x))$— pertenece al universo $\mid \mathfrak{M} \mid$ y la variable ligada, x, hace referencia a todo elemento del universo.

dicado *cretense* lo formalizamos mediante C y la relación binaria "x ha afirmado y alguna vez" mediante Ax, y. Entonces "algún cretense ha afirmado x alguna vez" puede formalizarse mediante:

$$\exists y\,(Cy \wedge Ay, x) \tag{5.53}$$

A la fórmula (5.53) la llamaremos κ y, como es costumbre, escribiremos $\kappa(x)$ para señalar que la variable x es libre en κ. Con estas notaciones, la versión formal de la oración de Epiménides se puede representar mediante:

$$\forall x\,(\kappa(x) \to \neg Tx) \tag{5.54}$$

Puesto que Epiménides era cretense, es verdadero que algún cretense ha afirmado la oración de Epiménides, por lo que debe cumplirse

$$\mathcal{V}_{\mathfrak{M}}(\kappa(\ulcorner\forall x(\kappa(x) \to \neg Tx)\urcorner)) = \mathfrak{v} \tag{5.55}$$

Abreviaremos $\ulcorner\forall x(\kappa(x) \to \neg Tx)\urcorner$ mediante τ, y $\forall x(x \neq \tau \to (\kappa(x) \to \neg Tx))$ mediante p. Entonces,

$$\mathcal{V}_{\mathfrak{M}}(\kappa(\tau)) = \mathfrak{v} \tag{5.56}$$

Tomemos la lógica trivaluada fuerte de Kleene. Aceptando la equivalencia fuerte entre $\forall x(\kappa(x) \to \neg Tx)$ y $(\kappa(\tau) \to \neg T\tau) \wedge p$ tendremos

$$\mathcal{V}_{\mathfrak{M}}(\forall x(\kappa(x) \to \neg Tx)) = \mathcal{V}_{\mathfrak{M}}((\kappa(\tau) \to \neg T\tau) \wedge p) \tag{5.57}$$

donde

$$\mathcal{V}_{\mathfrak{M}}((\kappa(\tau) \to \neg T\tau) \wedge p) = C_{\wedge}(\mathcal{V}_{\mathfrak{M}}(\kappa(\tau) \to \neg T\tau), \mathcal{V}_{\mathfrak{M}}(p)) \tag{5.58}$$

$$\mathcal{V}_{\mathfrak{M}}(\kappa(\tau) \to \neg T\tau) = C_{\to}(\mathcal{V}_{\mathfrak{M}}(\kappa(\tau)), \mathcal{V}_{\mathfrak{M}}(\neg T\tau)) \tag{5.59}$$

Teniendo en cuenta (5.56) y (5.59):

$$\mathcal{V}_{\mathfrak{M}}(\kappa(\tau) \to \neg T\tau) = C_{\to}(\mathfrak{v}, \mathcal{V}_{\mathfrak{M}}(\neg T\tau)) = \mathcal{V}_{\mathfrak{M}}(\neg T\tau) \tag{5.60}$$

De (5.57), (5.58) y (5.60) se obtiene:

$$\mathcal{V}_{\mathfrak{M}}(\forall x(\kappa(x) \to \neg Tx)) = C_{\wedge}(\mathcal{V}_{\mathfrak{M}}(\neg T\tau), \mathcal{V}_{\mathfrak{M}}(p)) \tag{5.61}$$

Es fácil ver que el que haya paradoja o no depende del valor de verdad de p. Si $\mathcal{V}_{\mathfrak{M}}(p) = \mathfrak{f}$, entonces

$$\mathcal{V}_{\mathfrak{M}}(\forall x(\kappa(x) \rightarrow \neg Tx)) = C_{\wedge}(\mathcal{V}_{\mathfrak{M}}(\neg T\tau), \mathfrak{f}) = \mathfrak{f} \tag{5.62}$$

es decir, la oración de Epiménides es falsa y no hay paradoja. Pero si $\mathcal{V}_{\mathfrak{M}}(p) = \mathfrak{v}$, entonces

$$\mathcal{V}_{\mathfrak{M}}(\forall x(\kappa(x) \rightarrow \neg Tx)) = C_{\wedge}(\mathcal{V}_{\mathfrak{M}}(\neg T\tau), \mathfrak{v}) = \mathcal{V}_{\mathfrak{M}}(\neg T\tau) \tag{5.63}$$

Ahora bien, como τ es $\ulcorner \forall x(\kappa(x) \rightarrow \neg Tx) \urcorner$ y T es un predicado de desentrecomillado:

$$\mathcal{V}_{\mathfrak{M}}(\neg T\tau) = C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\forall x(\kappa(x) \rightarrow \neg Tx)))) \tag{5.64}$$

Finalmente, considerando (5.63) y (5.64), es fácil concluir:

$$\mathcal{V}_{\mathfrak{M}}(\forall x(\kappa(x) \rightarrow \neg Tx)) = (C_{\neg} \circ C_T)(\mathcal{V}_{\mathfrak{M}}(\forall x(\kappa(x) \rightarrow \neg Tx))) \tag{5.65}$$

lo cual es imposible de cumplir puesto que la función $(C_{\neg} \circ C_T)$ no tiene ningún punto fijo.

5.3.3. Contextos referencialmente anuladores en oraciones cuantificadas

¿Cuál es la forma natural de extender a las oraciones cuantificadas la idea de los contextos referencialmente anuladores? Ahora no es un término cerrado el que tiene como referencia aparente la oración de la que forma parte (o una oración fuertemente equivalente a ella) sino una variable ligada con respecto a una asignación de variables particular.

En efecto, según las reglas de valuación,

$$\mathcal{V}_{\mathfrak{M}}(\forall x\sigma(x)) = C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\sigma(x))/e \in\mid \mathfrak{M} \mid\})$$

(donde s es una asignación de variables cualquiera) pero si la oración $\forall x\sigma(x)$ está en el universo del modelo, $\mathcal{V}_{\mathfrak{M},s[(\forall x\sigma(x))/x]}(\sigma(x))$ es un elemento del conjunto $\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\sigma(x))/e \in\mid \mathfrak{M} \mid\}$ y, por tanto, en general, $\mathcal{V}_{\mathfrak{M}}(\forall x\sigma(x))$ depende de

$\mathcal{V}_{\mathfrak{M},s[(\forall x\sigma(x))/x]}(\sigma(x))$. Ahora bien, por composicionalidad,

$$\mathcal{V}_{\mathfrak{M},s[(\forall x\sigma(x))/x]}(\sigma(x)) = C_\sigma(\mathcal{V}_{\mathfrak{M},s[(\forall x\sigma(x))/x]}(x)) \tag{5.66}$$

donde

$$\mathcal{V}_{\mathfrak{M},s[(\forall x\sigma(x))/x]}(x) = s[(\forall x\sigma(x))/x](x) = (\forall x\sigma(x)) = \mathcal{V}_{\mathfrak{M}}(\ulcorner\forall x\sigma(x)\urcorner) \tag{5.67}$$

Es decir, si en una oración no cuantificada, había autorreferencia cuando contenía un término, τ, cuya referencia era la propia oración; en una oración cuantificada, $\forall x\sigma(x)$, hay autorreferencia cuando la referencia de la variable x con respecto a una asignación de variables ($s[(\forall x\sigma(x))/x]$) es la propia oración. Consiguientemente, si en oraciones no cuantificadas la aparición de un término en ciertos contextos suponía que dejase de tener su referencia habitual; la extensión de esa idea a las oraciones cuantificadas es que algunas de dichas oraciones pueden crear contextos en los que la aparición de la variable ligada por el cuantificador deja de tener su referencia habitual con respecto a cierta asignación de variables. Por ejemplo, una oración de la forma $\forall x\sigma(x)$ podría crear un contexto en que determinada aparición de la variable x no esté en el dominio de la asignación de variables $s[(\forall x\sigma(x))/x]$.

La definición 5.1 (p. 176) no contemplaba la posibilidad de oraciones cuantificadas. Para hacerlo debemos completar aquella definición añadiendo lo siguiente:

> *En un lenguaje con capacidad de cita, diremos que una oración* de la forma $\forall x\sigma(x)$ *contiene una referencia semántica aparente a cualquier expresión formal del lenguaje (y en particular a la propia oración), mediante la variable x, si y solo si en dicha oración hay una aparición ligada de x formando parte de una expresión de la forma Dx y D es un predicado o una función de desentrecomillado. Si dicha aparición de la variable x no está en un contexto anulador de su referencia respecto a la asignación de variables $s[\varepsilon/x]$, la oración $\forall x\sigma(x)$ contiene una referencia semántica real a la expresión formal ε.*
>
> *Lo mismo es aplicable a oraciones de la forma $\exists x\sigma(x)$ o a oraciones de las que $\forall x\sigma(x)$* o *$\exists x\sigma(x)$ constituyen una suboración.*

Dada una asignación de variables, r, y una variable, x, no escribiremos $\mathcal{V}_{\mathfrak{M},r}(x)$ sino $\mathcal{V}_{\mathfrak{M},r}(x^{[A]})$ donde A es una especificación de una aparición de la variable x en

una fórmula. También escribiremos $\mathcal{V}^h_{\mathfrak{M},r}(x)$ para indicar la referencia habitual de x con respecto a la asignación de variables r.

Cambiaremos la definición $\mathcal{V}_{\mathfrak{M},r}(x) =_{def} r(x)$ por $\mathcal{V}_{\mathfrak{M},r}(x^{[A]}) =_{def} r(x^{[A]})$. Cuando $x^{[A]} \in dom(\mathcal{V}_{\mathfrak{M},r})$, $\mathcal{V}_{\mathfrak{M},r}(x^{[A]})$ —o $r(x^{[A]})$— es la referencia habitual de x con respecto a r: $\mathcal{V}^h_{\mathfrak{M},r}(x)$. También diremos que $\mathcal{V}^h_{\mathfrak{M},r}(x)$ o que $r(x^{[A]})$ es el valor que la asignación r asocia habitualmente a la variable x, al cual podremos referirnos mediante $r^h(x)$.

Analicemos a modo de ejemplo la siguiente versión cuantificada de la paradoja del mentiroso. Supongamos que en una hoja hay escrita una única oración: "ninguna oración escrita en esta hoja es verdadera". Esta oración puede formalizarse como

$$\forall x(Hx \rightarrow \neg Tx) \tag{5.68}$$

siempre que Hx se interprete como "x es una oración escrita en la hoja". En general, la oración $\forall x(Hx \rightarrow \neg Tx)$ es una versión cuantificada de la oración del mentiroso si en el lenguaje interpretado se cumple $\mathcal{V}_{\mathfrak{M},s[e/x]}(Hx) = \mathfrak{v}$ cuando e es la oración $\forall x(Hx \rightarrow \neg Tx)$ y $\mathcal{V}_{\mathfrak{M},s[e/x]}(Hx) = \mathfrak{f}$ cuando e es otra oración.

Según las reglas de valuación,

$$\mathcal{V}_{\mathfrak{M}}(\forall x(Hx \rightarrow \neg Tx)) = C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(Hx \rightarrow \neg Tx)/e \in |\, \mathfrak{M} \,|\})$$

Usando los mismos supuestos que en el análisis anterior sobre la oración de Epiménides,

$$\begin{aligned}
&C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(Hx \rightarrow \neg Tx)/e \in |\, \mathfrak{M} \,|\}) = \\
&= C_{\forall}(\{\mathfrak{v}, \mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(Hx \rightarrow \neg Tx)\}) = \\
&= C_{\forall}(\{\mathfrak{v}, \mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(\neg Tx)\}) = \\
&= C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(\neg Tx)\}) = \\
&= \mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(\neg Tx) = \\
&= C_{\neg}(\mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(Tx)) = \\
&= C_{\neg}(\mathfrak{I}_{\mathfrak{M}}(T)(\mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(x))) = \\
&= C_{\neg}(\mathfrak{I}_{\mathfrak{M}}(T)(s[(\forall x(Hx \rightarrow \neg Tx))/x](x)))
\end{aligned} \tag{5.69}$$

Ignorando, todavía, la existencia de contextos referencialmente anuladores podríamos continuar del siguiente modo:

$$\begin{aligned}
&C_{\neg}(\mathfrak{I}_{\mathfrak{M}}(T)(s[(\forall x(Hx \rightarrow \neg Tx))/x](x))) = \\
&= C_{\neg}(\mathfrak{I}_{\mathfrak{M}}(T)(\forall x(Hx \rightarrow \neg Tx))) = \\
&= C_{\neg}(\mathfrak{I}_{\mathfrak{M}}(T)(\mathcal{V}_{\mathfrak{M}}(\ulcorner \forall x(Hx \rightarrow \neg Tx \urcorner)))) = \\
&= C_{\neg}(\mathcal{V}_{\mathfrak{M}}(T\ulcorner \forall x(Hx \rightarrow \neg Tx \urcorner))) = \\
&= C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\forall x(Hx \rightarrow \neg Tx)))
\end{aligned} \tag{5.70}$$

y llegaríamos a la siguiente conclusión:

$$\mathcal{V}_{\mathfrak{M}}(\forall x(Hx \rightarrow \neg Tx)) = C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\forall x(Hx \rightarrow \neg Tx))) \tag{5.71}$$

que es, como sabemos, imposible.

Ahora bien, la tercera aparición de la variable x en la oración $\forall x(Hx \rightarrow \neg Tx)$ va precedida de un predicado de desentrecomillado por lo cual la oración tiene una referencia semántica aparente a sí misma. Así pues, esa referencia semántica podría no ser real y el desarrollo anterior, que conduce a contradicción, quedaría bloqueado.

Si representamos la tercera aparición de x en la oración $\forall x(Hx \rightarrow \neg Tx)$, por medio de $x^{[3,\forall x(Hx \rightarrow \neg Tx)]}$, la referencia semántica aparente de dicha oración a sí misma deja de ser real en cuanto se cumpla

$$x^{[3,\forall x(Hx \rightarrow \neg Tx)]} \notin dom(s[(\forall x(Hx \rightarrow \neg Tx))/x])$$

En ese caso, la paradoja desparece y, en consecuencia, podemos determinar el valor de verdad de $\forall x(Hx \rightarrow \neg Tx)$. Este valor de verdad es el mismo que el de la oración del mentiroso reforzada original. En efecto, es fácil comprobar que:

$$\begin{aligned}
&\mathcal{V}_{\mathfrak{M}}(\forall x(Hx \rightarrow \neg Tx)) = \\
&= \mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(Hx \rightarrow \neg Tx^{[3,\forall x(Hx \rightarrow \neg Tx)]}) = \\
&= \mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(\neg Tx^{[3,\forall x(Hx \rightarrow \neg Tx)]}) = \\
&= C_{\neg}(\mathcal{V}_{\mathfrak{M},s[(\forall x(Hx \rightarrow \neg Tx))/x]}(Tx^{[3,\forall x(Hx \rightarrow \neg Tx)]})) = \\
&= C_{\neg}(\mathfrak{i}) = \mathfrak{i}
\end{aligned} \tag{5.72}$$

En realidad, en la segunda línea de (5.72) tendríamos que haber escrito

$$\mathcal{V}_{\mathfrak{M},s[(\forall x(Hx\rightarrow\neg Tx))/x]}(Hx^{[2,\forall x(Hx\rightarrow\neg Tx)]}\rightarrow\neg Tx^{[3,\forall x(Hx\rightarrow\neg Tx)]})$$

en lugar de

$$\mathcal{V}_{\mathfrak{M},s[(\forall x(Hx\rightarrow\neg Tx))/x]}(Hx\rightarrow\neg Tx^{[3,\forall x(Hx\rightarrow\neg Tx)]})$$

para especificar las dos apariciones de x. Por sencillez, prescindiremos de los superíndices cuando ello no provoque ambigüedad (como en este caso, dado que H no es un predicado de desentrecomillado) o cuando queramos hacer un análisis ignorando la existencia de contextos referencialmente anuladores.

5.3.4. La paradoja del mentiroso en un lenguaje sin símbolos de función

Es sabido que en un lenguaje formal podemos suprimir los símbolos de función[22] sin perder capacidad expresiva. Por ejemplo, la función de diagonalización podría representarse mediante una fórmula con dos variables libres, $\Delta(x,y)$:

Definición 5.5 (fórmula para función diagonalización). *La fórmula $\Delta(x,y)$ representa una función de diagonalización en un lenguaje interpretado ssi para toda fórmula, $\sigma(x)$, con una única variable libre, x, y para toda asignación de variables, s, se cumple:*

$$\mathcal{V}_{\mathfrak{M},s}(\Delta(\ulcorner\sigma(x)\urcorner,y))=\mathcal{V}_{\mathfrak{M},s}(y=\ulcorner\sigma(\ulcorner\sigma(x)\urcorner)\urcorner) \tag{5.73}$$

Cabe pensar que, si no perdemos capacidad expresiva, habrá un problema de incompatibilidad entre la existencia de predicados de desentrecomillado y una fórmula que represente una función de diagonalización. Veamos que, en efecto, así es y analicemos después si la solución de los contextos referencialmente anuladores es aplicable.

Para mostrar la incompatibilidad anterior probemos primero la siguiente proposición (en el supuesto de que todas las referencias semánticas aparentes mediante variables ligadas son reales):

[22] Recordemos que nos referimos a funciones que tienen como argumento una secuencia de elementos del dominio y como valor un elemento del dominio.

Proposición 5.1. *Si la fórmula $\Delta(x,y)$ representa una función de diagonalización y el cuantificador $\forall$ y la conectiva $\rightarrow$ se interpretan extensionalmente según la lógica trivaluada fuerte de Kleene, entonces dada una fórmula, $\sigma(x)$, con una única variable libre, existe una oración, φ tal que*

$$\mathcal{V}_{\mathfrak{M}}(\varphi) = \mathcal{V}_{\mathfrak{M}}(\sigma(\ulcorner\varphi\urcorner)) \tag{5.74}$$

Demostración. Sea $\sigma_d(x)$ la fórmula $\forall y\,(\Delta(x,y) \rightarrow \sigma(y))$ y sea φ la oración $\sigma_d(\ulcorner\sigma_d(x)\urcorner)$. Entonces φ es la oración $\forall y\,(\Delta(\ulcorner\sigma_d(x)\urcorner, y) \rightarrow \sigma(y))$. Además, para cualquier asignación de variables, s:

(1) $\mathcal{V}_{\mathfrak{M},s}(\Delta(\ulcorner\sigma_d(x)\urcorner, y)) = \mathcal{V}_{\mathfrak{M},s}(y = \ulcorner\sigma_d(\ulcorner\sigma_d(x)\urcorner)\urcorner)$
; $\Delta(x,y)$ representa una función de diagonalización

(2) $\mathcal{V}_{\mathfrak{M},s}(\Delta(\ulcorner\sigma_d(x)\urcorner, y)) = \mathcal{V}_{\mathfrak{M},s}(y = \ulcorner\varphi\urcorner)$
; 1 y φ es la oración $\sigma_d(\ulcorner\sigma_d(x)\urcorner)$

(3) $\mathcal{V}_{\mathfrak{M},s}(\forall y\,(\Delta(\ulcorner\sigma_d(x)\urcorner, y) \rightarrow \sigma(y))) = \mathcal{V}_{\mathfrak{M},s}(\forall y\,(y = \ulcorner\varphi\urcorner \rightarrow \sigma(y)))$
; 2 y composicionalidad

(4) $\mathcal{V}_{\mathfrak{M},s}(\varphi) = \mathcal{V}_{\mathfrak{M},s}(\forall y\,(y = \ulcorner\varphi\urcorner \rightarrow \sigma(y)))$
; 3 y φ es la oración $\forall y\,(\Delta(\ulcorner\sigma_d(x)\urcorner, y) \rightarrow \sigma(y))$

(5) $\mathcal{V}_{\mathfrak{M},s}(\forall y\,(y = \ulcorner\varphi\urcorner \rightarrow \sigma(y))) = \mathcal{V}_{\mathfrak{M},s}(\sigma(\ulcorner\varphi\urcorner))$
; aplicación de la proposición 4.10 (p. 131)

(6) $\mathcal{V}_{\mathfrak{M},s}(\varphi) = \mathcal{V}_{\mathfrak{M},s}(\sigma(\ulcorner\varphi\urcorner))$; 4 y 5

(7) $\mathcal{V}_{\mathfrak{M}}(\varphi) = \mathcal{V}_{\mathfrak{M}}(\sigma(\ulcorner\varphi\urcorner))$
; 6, dado que φ y $\sigma(\ulcorner\varphi\urcorner)$ no tienen variables libres

■

En una lógica trivaluada fuerte de Kleene, es incompatible la existencia de una fórmula, $\Delta(x,y)$, que represente una función de diagonalización con la existencia de una oración de la forma $\rho(D\tau)$ en la que D sea un predicado de desentrecomillado y para la que $(C_\rho \circ C_D)$ no tenga ningún punto fijo. Si ambas existiesen, la aplicación de la proposición anterior a la fórmula $\rho(Dx)$ nos diría que existe una oración, φ tal que $\mathcal{V}_{\mathfrak{M}}(\varphi) = \mathcal{V}_{\mathfrak{M}}(\rho(D\ulcorner\varphi\urcorner))$. Pero $\mathcal{V}_{\mathfrak{M}}(\rho(D\ulcorner\varphi\urcorner)) = C_\rho(C_D(\mathcal{V}_{\mathfrak{M}}(\varphi)))$. Por tanto tendríamos $\mathcal{V}_{\mathfrak{M}}(\varphi) = (C_\rho \circ C_D)(\mathcal{V}_{\mathfrak{M}}(\varphi))$ lo cual es imposible si $(C_\rho \circ C_D)$ no tiene ningún punto fijo.

La oración φ de la que estamos hablando es, según la proposición anterior, $\sigma_d(\ulcorner\sigma_d(x)\urcorner)$, pero puesto que $\sigma_d(x)$ es la fórmula $\forall y\,(\Delta(x,y) \rightarrow \sigma(y))$ y estamos

suponiendo que $\sigma(x)$ es la fórmula $\rho(Dx)$, la oración φ será:

$$\forall y\,(\Delta(\ulcorner\sigma_d(x)\urcorner, y) \to \rho(Dy)) \tag{5.75}$$

o, con todo detalle:

$$\forall y\,(\Delta(\ulcorner\forall y\,(\Delta(x,y) \to \rho(Dy))\urcorner, y) \to \rho(Dy)) \tag{5.76}$$

Una vez comprobada la incompatibilidad entre la existencia de predicados de desentrecomillado y una fórmula que represente una función de diagonalización, ha llegado el momento de comprobar que la solución de los contextos referencialmente anuladores es aplicable a la oración (5.76). Para hacer más sencilla la explicación tomaremos un caso particular de dicha oración, aunque será patente que no perdemos generalidad. El caso particular consiste en tomar como $\rho(Dx)$ la fórmula $\neg Tx$, siendo T el predicado *verdadero*. En tal caso, la oración (5.76) se convierte en la oración del mentiroso reforzada (a la que llamaremos λ):[23]

$$\forall y\,(\Delta(\ulcorner\forall y\,(\Delta(x,y) \to \neg Ty)\urcorner, y) \to \neg Ty) \tag{5.77}$$

Si llamamos $\alpha(y)$ a la fórmula $\Delta(\ulcorner\forall y\,(\Delta(x,y) \to \neg Ty)\urcorner, y)$, la oración λ se puede representar de un modo más compacto mediante:

$$\forall y\,(\alpha(y) \to \neg Ty) \tag{5.78}$$

Obsérvese la similitud de esta oración con la oración de Epiménides —(5.54) en p. 185— y con la versión cuantificada de la paradoja del mentiroso —(5.68) en p. 188—. Además se tiene:

$$\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y)) = \mathcal{V}_{\mathfrak{M},s[e/y]}(\Delta(\ulcorner\forall y\,(\Delta(x,y) \to \neg Ty)\urcorner, y)) \tag{5.79}$$

y, según la definición 5.5 (p. 190),

$$\begin{aligned}&\mathcal{V}_{\mathfrak{M},s[e/y]}(\Delta(\ulcorner\forall y\,(\Delta(x,y) \to \neg Ty)\urcorner, y)) = \\ &= \mathcal{V}_{\mathfrak{M},s[e/y]}(y = \ulcorner\forall y\,(\Delta(\ulcorner\forall y\,(\Delta(x,y) \to \neg Ty)\urcorner, y) \to \neg Ty)\urcorner)\end{aligned} \tag{5.80}$$

[23] De hecho, según la proposición 5.1, se cumpliría $\mathcal{V}_{\mathfrak{M}}(\lambda) = \mathcal{V}_{\mathfrak{M}}(\neg T\ulcorner\lambda\urcorner)$, es decir, extensionalmente, λ *diría* de sí misma que no es verdadera.

Como hemos llamado λ a la oración (5.77):

$$\begin{array}{l}\mathcal{V}_{\mathfrak{M},s[e/y]}(y = \ulcorner\forall y\,(\Delta(\ulcorner\forall y\,(\Delta(x,y) \to \neg Ty)\urcorner, y) \to \neg Ty)\urcorner) = \\ = \mathcal{V}_{\mathfrak{M},s[e/y]}(y = \ulcorner\lambda\urcorner)\end{array} \tag{5.81}$$

De (5.79), (5.80) y (5.81):

$$\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y)) = \mathcal{V}_{\mathfrak{M},s[e/y]}(y = \ulcorner\lambda\urcorner) \tag{5.82}$$

Por consiguiente, cuando e es la oración λ se cumple $\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y)) = \mathfrak{v}$ y, cuando e es otra oración, se cumple $\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y)) = \mathfrak{f}$. Todo ello nos permite hacer el mismo análisis para λ, es decir, para $\forall y\,(\alpha(y) \to \neg Ty)$ —oración (5.78)—, que el que hicimos para la oración $\forall x(Hx \to \neg Tx)$ —oración (5.68) en p. 188— donde también se cumplía $\mathcal{V}_{\mathfrak{M},s[e/x]}(Hx) = \mathfrak{v}$ cuando e era la oración $\forall x(Hx \to \neg Tx)$ y $\mathcal{V}_{\mathfrak{M},s[e/x]}(Hx) = \mathfrak{f}$ cuando e era otra oración (este era el requisito para que $\forall x(Hx \to \neg Tx)$ fuese una versión cuantificada de la oración del mentiroso).

El análisis que hicimos nos llevó a la conclusión de que cuando x aparecía precedida del predicado de desentrecomillado T en la oración $\forall x(Hx \to \neg Tx)$, no tenía referencia respecto a la asignación de variables $s[(\forall x(Hx \to \neg Tx))/x]$. Por similitud, la solución a la oración λ consistirá en que la aparición de la variable y precedida de T en la oración —aparición que indicaremos mediante $y^{[ult,\lambda]}$ por ser la última de y en λ— no tiene referencia respecto a la asignación de variables $s[\lambda/y]$. Entonces, el valor de verdad de λ será el mismo que el de $\forall x(Hx \to \neg Tx)$. En efecto:

$$\begin{array}{l}\mathcal{V}_{\mathfrak{M}}(\lambda) = \mathcal{V}_{\mathfrak{M}}(\forall y\,(\alpha(y) \to \neg Ty)) = \\ = C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y) \to \neg Ty)/e \in |\,\mathfrak{M}\,|\}) = \\ = C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y) \to \neg Ty)/e \in (|\,\mathfrak{M}\,| - \{\lambda\})\} \cup \\ \cup\{\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y) \to \neg Ty)/e = \lambda\})\end{array} \tag{5.83}$$

Cuando $e \neq \lambda$, $\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y)) = \mathfrak{f}$ y por tanto

$$\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y) \to \neg Ty^{[ult,\lambda]}) = \mathfrak{v}$$

Cuando $e = \lambda$, $\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y)) = \mathfrak{v}$, por lo que

$$\mathcal{V}_{\mathfrak{M},s[e/y]}(\alpha(y) \to \neg Ty^{[ult,\lambda]}) = \mathcal{V}_{\mathfrak{M},s[e/y]}(\neg Ty^{[ult,\lambda]})$$

De estas consideraciones y (5.83) resulta:

$$\begin{aligned}&\mathcal{V}_{\mathfrak{M}}(\lambda) = C_{\forall}(\{\mathfrak{v}\} \cup \{\mathcal{V}_{\mathfrak{M},s[\lambda/y]}(\neg Ty^{[ult,\lambda]})\}) = \\ &= C_{\forall}(\{\mathfrak{v}, \mathcal{V}_{\mathfrak{M},s[\lambda/y]}(\neg Ty^{[ult,\lambda]})\} = \\ &= C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[\lambda/y]}(\neg Ty^{[ult,\lambda]})\}) = \\ &= \mathcal{V}_{\mathfrak{M},s[\lambda/y]}(\neg Ty^{[ult,\lambda]}) = C_{\neg}(\mathcal{V}_{\mathfrak{M},s[\lambda/y]}(Ty^{[ult,\lambda]}))\end{aligned} \tag{5.84}$$

Finalmente, puesto que la aparición $y^{[ult,\lambda]}$ de la variable y no tiene referencia respecto a la asignación de variables $s[\lambda/y]$, $\mathcal{V}_{\mathfrak{M},s[\lambda/y]}(Ty^{[ult,\lambda]}) = \mathfrak{i}$, con lo que llegamos a:

$$\mathcal{V}_{\mathfrak{M}}(\lambda) = C_{\neg}(\mathcal{V}_{\mathfrak{M},s[\lambda/y]}(Ty^{[ult,\lambda]})) = C_{\neg}(\mathfrak{i}) = \mathfrak{i} \tag{5.85}$$

Como conclusión, constatamos: 1/ en un lenguaje carente de símbolos de función las oraciones que hacían uso de ellos pueden reescribirse como oraciones cuantificadas; 2/ el planteamiento de los contextos referencialmente anuladores que habíamos generalizado para que fuese aplicable a oraciones cuantificadas es, por supuesto, aplicable sin que importe que el lenguaje tenga o no símbolos de función. Por eso la oración $\forall y\,(\alpha(y) \rightarrow \neg Ty)$ se resuelve sin mayores problemas que la oración $\forall x(Hx \rightarrow \neg Tx)$ —de hecho ambas son distintas formas de expresar la oración del mentiroso reforzado—.

5.4. Sistemas infinitos de oraciones

Algunos sistemas infinitos de oraciones muestran que puede haber paradoja sin necesidad de autorreferencia directa o indirecta. Uno de los más conocidos es el de Yablo formado por el conjunto de oraciones $\{Y_i / i \in \mathbb{N}\}$ donde la oración Y_i es

$$\forall k > i,\ Y_k\ no\ es\ verdadera \tag{5.86}$$

La paradoja surge porque si existe un $i \in \mathbb{N}$ tal que Y_i es verdadera, entonces

$$\forall k > i,\ Y_k\ no\ es\ verdadera$$

de donde se deduce

$$\forall k > (i+1),\ Y_k\ no\ es\ verdadera$$

que es Y_{i+1}. Así pues, se deduce $Y_{i+1}\, es\, verdadera$, lo cual se contradice con $\forall k > i$, $Y_k\, no\, es\, verdadera$. Por otra parte, si no hay ningún $i \in \mathbb{N}$ tal que Y_i es verdadera, entonces es verdadera la oración

$$\forall k > 1,\, Y_k\, no\, es\, verdadera$$

es decir, es verdadera la oración Y_1, lo cual contradice la hipótesis. En resumen, tanto si existe un $i \in \mathbb{N}$ tal que Y_i es verdadero como si no, llegamos a una contradicción.

Otro ejemplo de sistema infinito y paradójico de oraciones es el de Sorensen obtenido al cambiar el cuantificador universal por el existencial en el sistema de oraciones de Yablo. También hay sistemas infinitos de oraciones que sin ser estrictamente paradójicos permiten distintas evaluaciones, como ocurre con la oración del veraz. Por ejemplo:

- $\{S_i / i \in \mathbb{N}\}$ donde la oración S_i es $S_{i+1}\, es\, verdadera$. Se pueden evaluar todas las oraciones como verdaderas o todas como falsas.
- $\{S_i / i \in \mathbb{N}\}$ donde la oración S_i es $S_{i+1}\, no\, es\, verdadera$. Se pueden evaluar todas las oraciones de índice par como verdaderas y las de índice impar como falsas o viceversa.

Para analizar estos sistemas de oraciones deberíamos generalizar nuestro método de evaluación de sistemas finitos. Aquél método no es aplicable directamente porque, en los sistemas infinitos, en la relación de orden $\prec_S$ definida sobre $\Phi/\sim_S$ puede no haber ningún elemento sin sucesor por el cual empezar la evaluación. No obstante, hay una parte del método que puede y debe conservarse: añadir al conjunto de oraciones las oraciones necesarias para que resulte un conjunto S-cerrado. Además, igual que hicimos en el estudio de sistemas finitos de oraciones, definimos sobre Φ la relación de equivalencia $\sim_S$ y sobre el conjunto de clases de equivalencia $\Phi/\sim_S$ una relación irreflexiva y asimétrica $\prec_S$.

Consideremos el grafo $G(\prec_S, \Phi/\sim_S)$. Como en los sistemas finitos deberíamos comenzar evaluando las oraciones de los vértices[24] del grafo que no tienen sucesor (si los hay) y a continuación evaluaremos las oraciones de aquellos vértices cuyos sucesores ya han sido evaluados. De esta forma evaluaríamos únicamente

[24] Recordemos que un vértice de $G(\prec_S, \Phi/\sim_S)$ es un conjunto de oraciones que constituye una clase de equivalencia según la relación $\sim_S$.

las oraciones de aquellos vértices del grafo $G(\prec_S, \Phi/\sim_S)$ de los que no parta ningún camino infinito. Pero la dificultad es mayor para evaluar las oraciones de los vértices de los que parten caminos infinitos.

En este trabajo no buscaremos un método general de evaluación de sistemas infinitos de oraciones, sino que nos limitaremos a evaluar el sistema de Yablo de forma coherente con las ideas seguidas en la evaluación de sistemas finitos de oraciones. No obstante, la forma de evaluar el sistema de Yablo, puede enseñarnos aspectos importantes acerca de cómo evaluar otros sistemas infinitos de oraciones.

El ejemplo de Yablo nos plantea la cuestión de cómo evaluar Y_1. Si bien la evaluación del conjunto de oraciones $\{Y_2, Y_3, ...Y_n, ...\}$ no depende de la evaluación de Y_1, no es lo mismo evaluar Y_1 posteriormente a $\{Y_2, Y_3, ...Y_n, ...\}$ que simultáneamente con $\{Y_2, Y_3, ...Y_n, ...\}$. La diferencia es que, como veremos, una evaluación conjunta del sistema de Yablo nos lleva a concluir que ninguna de las oraciones es verdadera ni falsa. En cambio, si primero evaluamos conjuntamente $\{Y_2, Y_3, ...Y_n, ...\}$ —con lo que ninguna de las oraciones del conjunto es verdadera ni falsa— y *después* evaluamos Y_1, esta oración resulta verdadera.

La respuesta a nuestra pregunta es afortunadamente sencilla. Debemos evaluar Y_1 simultáneamente con $\{Y_2, Y_3, ...Y_n, ...\}$. Elegir la opción de evaluar Y_1 posteriormente a $\{Y_2, Y_3, ...Y_n, ...\}$ es descartable porque entonces lo coherente sería también evaluar Y_2 posteriormente a $\{Y_3, Y_4, ...Y_n, ...\}$, Y_3 posteriormente a $\{Y_4, Y_5, ...Y_n, ...\}$, etc. lo cual impediría evaluar una sola de las oraciones de Yablo. Evaluar $\{Y_2, Y_3, ...Y_n, ...\}$ conjuntamente y, posteriormente Y_1 no solo es incoherente sino que además produciría el resultado de que Y_1 es verdadera mientras que Y_2 no lo es. Este resultado no es el esperable si tenemos en cuenta la completa similitud entre el sistema de oraciones $\{Y_1, Y_2, ...Y_n, ...\}$ y el sistema $\{Y_2, Y_3, ...Y_n, ...\}$ que nos llevaría a afirmar que el valor de verdad de Y_1 debe ser el mismo que el de Y_2 e incluso, que el valor de verdad de todas las oraciones de $\{Y_1, Y_2, ...Y_n, ...\}$ debe ser el mismo.

Formalicemos el sistema de Yablo en un lenguaje interpretado trivaluado extensional con capacidad de cita. Para ello nuestro lenguaje tendrá: a) un modelo que incluya los números naturales ($\mathbb{N} \subset |\, \mathfrak{M} \,|$); b) para cada número natural, i, un símbolo de constante, c_i, llamado numeral, que se interpretará como el número natural i; c) un símbolo de predicado, N que se interpretará como el predicado "ser un número natural"; d) un símbolo de relación binaria $R_>$ que se interpretará como la relación numérica "mayor que"; e) un símbolo de función monádica, f, que

se interpretará, habitualmente, como la función que asocia al número$i$ la oración que antes llamábamos Y_i, es decir:

$$\mathcal{V}^h_{\mathfrak{M}}(fc_i) = [\forall x(Nx \to (R_{>}x, c_i \to \neg Tfx))] \tag{5.87}$$

El sistema de Yablo estará formado por el siguiente conjunto de oraciones:

$$\{[\forall x(Nx \to (R_{>}x, c_i \to \neg Tfx))]/i \in \mathbb{N}\} \tag{5.88}$$

Abreviaremos la fórmula

$$(Nx \to (R_{>}x, c_i \to \neg Tfx)) \tag{5.89}$$

mediante $\gamma_i(x)$, por lo que la oración i-ésima del sistema de Yablo será $\forall x\gamma_i(x)$.

Estudiemos el sistema de ecuaciones de valores de verdad asociado. Si llamamos z_i a la incógnita que representa el valor de verdad de la oración i-ésima del sistema de Yablo, tendremos:

$$z_i = \mathcal{V}_{\mathfrak{M}}(\forall x\gamma_i(x)) = C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x))/e \in| \mathfrak{M} |\}) \tag{5.90}$$

Si, además, tenemos en cuenta que $e \in| \mathfrak{M} |$ equivale a

$$((e \in| \mathfrak{M} |-\mathbb{N}) \vee (e \in \mathbb{N} \wedge e \leq i)) \vee (e \in \mathbb{N} \wedge e > i) \tag{5.91}$$

nos quedará:

$$z_i = C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x))/((e \in| \mathfrak{M} |-\mathbb{N}) \vee (e \in \mathbb{N} \wedge e \leq i)) \vee (e \in \mathbb{N} \wedge e > i)\}) \tag{5.92}$$

Aplicando una propiedad elemental de la unión de conjuntos:

$$\begin{aligned} z_i = C_{\forall}(&\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x))/((e \in| \mathfrak{M} |-\mathbb{N}) \vee (e \in \mathbb{N} \wedge e \leq i))\} \cup \\ &\cup\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x))/(e \in \mathbb{N} \wedge e > i)\}) \end{aligned} \tag{5.93}$$

Cuando $e \in| \mathfrak{M} |-\mathbb{N}$, $\mathcal{V}_{\mathfrak{M},s[e/x]}(Nx) = \mathfrak{f}$ y, dada la interpretación del condicional:

$$\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x)) = \mathcal{V}_{\mathfrak{M},s[e/x]}((Nx \to (R_{>}x, c_i \to \neg Tfx))) = \mathfrak{v} \tag{5.94}$$

Cuando $(e \in \mathbb{N} \wedge e \leq i)$, $\mathcal{V}_{\mathfrak{M},s[e/x]}(R_{>}x, c_i) = \mathfrak{f}$ y, dada la interpretación del condicional:

$$\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x)) = \mathcal{V}_{\mathfrak{M},s[e/x]}((Nx \rightarrow (R_{>}x, c_i \rightarrow \neg Tfx))) = \mathfrak{v} \tag{5.95}$$

Así pues,

$$\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x))/((e \in |\, \mathfrak{M} \mid -\mathbb{N}) \vee (e \in \mathbb{N} \wedge e \leq i))\} = \{\mathfrak{v}\} \tag{5.96}$$

lo que nos permite pasar de (5.93) a:

$$z_i = C_{\forall}(\{\mathfrak{v}\} \cup \{\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x))/(e \in \mathbb{N} \wedge e > i)\}) \tag{5.97}$$

Si $(e \in \mathbb{N} \wedge e > i)$, entonces $\mathcal{V}_{\mathfrak{M},s[e/x]}(Nx) = \mathcal{V}_{\mathfrak{M},s[e/x]}(R_{>}x, c_i) = \mathfrak{v}$ y, dada la interpretación del condicional,

$$\mathcal{V}_{\mathfrak{M},s[e/x]}(\gamma_i(x)) = \mathcal{V}_{\mathfrak{M},s[e/x]}((Nx \rightarrow (R_{>}x, c_i \rightarrow \neg Tfx))) = \mathcal{V}_{\mathfrak{M},s[e/x]}(\neg Tfx) \tag{5.98}$$

De (5.97) y (5.98):

$$z_i = C_{\forall}(\{\mathfrak{v}\} \cup \{\mathcal{V}_{\mathfrak{M},s[e/x]}(\neg Tfx)/(e \in \mathbb{N} \wedge e > i)\}) \tag{5.99}$$

Los restantes pasos son sencillos:

$$\begin{aligned} & z_i = C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\neg Tfx)/(e \in \mathbb{N} \wedge e > i)\}) = \\ & = C_{\forall}(\{C_{\neg}(\mathcal{V}_{\mathfrak{M},s[e/x]}(Tfx))/(e \in \mathbb{N} \wedge e > i)\}) = \\ & = C_{\forall}(\{C_{\neg}(\mathcal{V}_{\mathfrak{M}}(Tfc_e))/(e \in \mathbb{N} \wedge e > i)\}) = \\ & = C_{\forall}(\{C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\forall x \gamma_e(x))))/(e \in \mathbb{N} \wedge e > i)\}) = \\ & = C_{\forall}(\{(C_{\neg} \circ C_T)(z_e))/(e \in \mathbb{N} \wedge e > i)\}) \end{aligned} \tag{5.100}$$

Este resultado significa que, con el supuesto que $e \in \mathbb{N}$, nuestro sistema de ecuaciones de valores de verdad es

$$\left.\begin{aligned} & z_1 = C_{\forall}(\{(C_{\neg} \circ C_T)(z_e))/e > 1\}) \\ & z_2 = C_{\forall}(\{(C_{\neg} \circ C_T)(z_e))/e > 2\}) \\ & \ldots \\ & z_i = C_{\forall}(\{(C_{\neg} \circ C_T)(z_e))/e > i\}) \\ & \ldots \end{aligned}\right\} \tag{5.101}$$

que, como sabemos por el razonamiento con el que hemos puesto de manifiesto la paradoja de Yablo, es un sistema sin solución. De acuerdo con nuestro planteamiento, concluimos que ninguna de las referencias aparentes a otras oraciones del sistema es real.

Como el modo en que una oración de la forma $\forall x(Nx \to (R_{>}x, c_i \to \neg Tfx))$ hace referencia aparente a otra oración del sistema es mediante el término fx con respecto a una asignación de variables particular,[25] diremos que, en el contexto de una oración de la forma $\forall x(Nx \to (R_{>}x, c_i \to \neg Tfx))$, la última aparición de la variable x no está en el dominio de ninguna asignación de variables de la forma $s[n/x]$, siendo n un número natural mayor que i. En consecuencia, si n es un número natural mayor que i,

$$\mathcal{V}_{\mathfrak{M},s[n/x]}(Tfx^{[\forall x(Nx\to(R_{>}x,c_i\to\neg Tfx))]}) = \mathfrak{i} \tag{5.102}$$

Para obtener la auténtica evaluación de $\forall x(Nx \to (R_{>}x, c_i \to \neg Tfx))$ es válido el desarrollo que hemos hecho de z_i —siempre que la última aparición de la variable x se especifique como en (5.102)— hasta llegar a

$$z_i = C_{\forall}(\{C_{\neg}(\mathcal{V}_{\mathfrak{M},s[e/x]}(Tfx^{[\forall x(Nx\to(R_{>}x,c_i\to\neg Tfx))]}))/(e \in \mathbb{N} \wedge e > i)\}) \tag{5.103}$$

—véase (5.100)—. A partir de aquí, basta tener en cuenta (5.102) para obtener como valor de verdad de $\forall x(Nx \to (R_{>}x, c_i \to \neg Tfx))$:

$$C_{\forall}(\{C_{\neg}(\mathfrak{i})\}) = C_{\forall}(\{\mathfrak{i}\}) = \mathfrak{i} \tag{5.104}$$

Informalmente, esto significa que hemos concluido que ninguna de las oraciones del sistema de Yablo es verdadera ni falsa. Pero, ¿podemos expresar el hecho de que ninguna de las oraciones del sistema de Yablo es verdadera mediante la oración formal $\forall x(Nx \to \neg Tfx)$?, es decir, ¿será verdadera dicha oración? No, porque igual que hemos hecho con la oración Y_1, debemos evaluarla conjuntamente con el resto de oraciones: al fin y al cabo esta oración se puede considerar como una oración Y_0 de un sistema de Yablo ampliado y, por simetría, su valor de verdad debe ser el mismo que el de Y_1. Veamos si es así al evaluarla conjuntamente con el resto de oraciones.

[25] Por ejemplo, la referencia habitual de fx con respecto a la asignación de variables $s[2/x]$ es la segunda oración del sistema de Yablo: $\forall x(Nx \to (R_{>}(x, c_2) \to \neg Tfx))$.

Si llamamos w_0 a la incógnita que representa el valor de verdad de la oración $\forall x(Nx \rightarrow \neg Tfx)$ y w_i a la incógnita que representa el valor de verdad de la oración i-ésima del sistema de Yablo, el sistema de ecuaciones de valores de verdad será:

$$\left.\begin{array}{l} w_o = C_\forall(\{(C_\neg \circ C_T)(w_n))/n \in \mathbb{N}\}) \\ w_1 = C_\forall(\{(C_\neg \circ C_T)(w_n))/n > 1\}) \\ ... \\ w_{i-1} = C_\forall(\{(C_\neg \circ C_T)(w_n))/n > i-1\}) \\ w_i = C_\forall(\{(C_\neg \circ C_T)(w_n))/n > i\}) \\ ... \end{array}\right\} \tag{5.105}$$

Este nuevo sistema y (5.101) son prácticamente iguales, pues basta con renombrar, para todo j, w_j como z_{j+1} para obtener:

$$\left.\begin{array}{l} z_1 = C_\forall(\{(C_\neg \circ C_T)(z_{n+1}))/n \in \mathbb{N}\}) \\ z_2 = C_\forall(\{(C_\neg \circ C_T)(z_{n+1}))/n > 1\}) \\ ... \\ z_i = C_\forall(\{(C_\neg \circ C_T)(z_{n+1}))/n > i-1\}) \\ ... \end{array}\right\} \tag{5.106}$$

Si llamamos e a $n+1$, nos queda exactamente el sistema (5.101). No hay pues más motivo para separar la evaluación de la oración $\forall x(Nx \rightarrow \neg Tfx)$ del resto de oraciones de Yablo que para separar la evaluación de Y_1 de las demás. Pero hemos visto que esto no es posible, porque por el mismo motivo tendríamos que separar la evaluación de $Y_2, Y_3, ...$ y no habría sistema de oraciones por el que comenzar la evaluación.

La auténtica evaluación de $\forall x(Nx \rightarrow \neg Tfx)$ será pues:

$$C_\forall(\{C_\neg(\mathcal{V}_{\mathfrak{M},s[n/x]}(Tfx^{[\forall x(Nx \rightarrow \neg Tfx)]}))/n \in \mathbb{N}\}) = C_\forall(\{C_\neg(\mathfrak{i})\}) = \mathfrak{i} \tag{5.107}$$

Sin embargo, uno de los requisitos exigibles a la solución de una paradoja en un lenguaje formal es que se pueda expresar dentro de ese lenguaje formal lo que en el metalenguaje afirmamos sobre el valor de verdad de las oraciones problemáticas. Surge entonces la cuestión: si $\forall x(Nx \rightarrow \neg Tfx)$ no la evaluamos como verdadera, ¿cómo podemos afirmar en el lenguaje formal que ninguna de las oraciones de Yablo es verdadera?

Observemos que para afirmar que la oración de Yablo Y_i no es verdadera disponemos de:

$$\neg T^{\ulcorner}\forall x(Nx \to (R_{>}x, c_i \to \neg Tfx))^{\urcorner} \tag{5.108}$$

Aparentemente esto nos llevaría a:

$$\forall y(Ny \to \neg T^{\ulcorner}\forall x(Nx \to (R_{>}x, y \to \neg Tfx))^{\urcorner}) \tag{5.109}$$

como una forma de decir que ninguna oración de Yablo es verdadera. Pero la última aparición de y forma parte de un término de cita por lo que no puede tomarse como una aparición de la variable y ligada por el cuantificador. Afortunadamente, el problema puede solventarse fácilmente.

Una forma de hacerlo es introducir en el lenguaje un operador monádico T cuya función de interpretación C_{T} sea idéntica a C_T. Entonces podríamos usar la oración

$$\forall y(Ny \to (\neg\, \mathsf{T}\, (\forall x(Nx \to (R_{>}x, y \to \neg Tfx))))) \tag{5.110}$$

para afirmar que ninguna oración de Yablo es verdadera. Obsérvese que al desaparecer el término de cita, desaparece el problema.

Finalizaremos comprobando que las oraciones (5.108) y (5.110) son verdaderas. Empecemos por la primera:

$$\begin{aligned}
&\mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\forall x(Nx \to (R_{>}x, c_i \to \neg Tfx))^{\urcorner}) = \\
&= (C_{\neg} \circ C_T)(\mathcal{V}_{\mathfrak{M}}[\forall x(Nx \to (R_{>}(x, c_i) \to \neg Tfx))]) = \\
&= (C_{\neg} \circ C_T)(\mathfrak{i}) = \mathfrak{v}
\end{aligned} \tag{5.111}$$

donde hemos tenido en cuenta que el valor de la i-ésima oración de Yablo es$\mathfrak{i}$.

En cuanto a (5.110):

$$\begin{aligned}
&\mathcal{V}_{\mathfrak{M}}(\forall y(Ny \to (\neg\, \mathsf{T}\, (\forall x(Nx \to (R_{>}x, y \to \neg Tfx)))))) = \\
&= C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[n/y]}(\neg\, \mathsf{T}\, (\forall x(Nx \to (R_{>}x, y \to \neg Tfx))))/n \in \mathbb{N}\}) = \\
&= C_{\forall}(\{(C_{\neg} \circ C_T)(\mathcal{V}_{\mathfrak{M},s[n/y]}(\forall x(Nx \to (R_{>}x, y \to \neg Tfx))))/n \in \mathbb{N})\})
\end{aligned} \tag{5.112}$$

donde

$$\begin{aligned}
&\mathcal{V}_{\mathfrak{M},s[n/y]}(\forall x(Nx \to (R_{>}x, y \to \neg Tfx)))) = \\
&= \mathcal{V}_{\mathfrak{M}}(\forall x(Nx \to (R_{>}x, c_n \to \neg Tfx)))) = \\
&= C_{\forall}(\{\mathcal{V}_{\mathfrak{M},s[e/x]}(\neg Tfx^{[\forall x(Nx \to (R_{>}x, c_n \to \neg Tfx))]})/e \in \mathbb{N} \wedge e > n\}) = \\
&= C_{\forall}(\{C_{\neg}(\mathcal{V}_{\mathfrak{M},s[e/x]}(Tfx^{[\forall x(Nx \to (R_{>}x, c_n \to \neg Tfx))]}))/e \in \mathbb{N} \wedge e > n\}) = \\
&= C_{\forall}(\{C_{\neg}(\mathfrak{i})\}) = C_{\forall}(\{\mathfrak{i}\}) = \mathfrak{i}
\end{aligned} \tag{5.113}$$

Aquí hemos tenido en cuenta (5.102) para pasar de la penúltima línea a la última. Finalmente, de (5.112) y (5.113) obtenemos:

$$\begin{aligned}
&\mathcal{V}_{\mathfrak{M}}(\forall y(Ny \to (\neg \top (\forall x(Nx \to (R_{>}x, y \to \neg Tfx)))))) = \\
&= C_{\forall}(\{(C_{\neg} \circ C_T)(\mathfrak{i})\}) = C_{\forall}(\{\mathfrak{v}\}) = \mathfrak{v}
\end{aligned} \tag{5.114}$$

Hemos confirmado que tanto (5.108) como (5.110) son verdaderas y por consiguiente, una vez más, hemos confirmado que nuestro lenguaje tiene capacidad suficiente para expresar cuál es el valor de verdad de las oraciones paradójicas.

5.5. Generalización a otras Paradojas

5.5.1. Introducción

¿Podemos encontrar una explicación común a las paradojas estudiadas hasta ahora y a otras con las que se observan importantes similitudes como la paradoja heterológica o las paradojas señaladas por Whitehead y Russell en *Principia Mathematica?*

El convencimiento de que esa explicación común existe porque hay una causa común es muy claro en Russell (la causa sería la violación del principio del círculo vicioso) y, más recientemente, en Priest (1994) quien muestra un esquema (*Qualified Russell's Schema*) al que se ajustan tanto las paradojas lógicas como las conjuntistas y defiende que las paradojas que comparten un mismo esquema deben compartir también una misma solución. Goldstein (2000, p. 64) señala otros autores, además de Priest, cuyo trabajo indica que al menos algunas de las paradojas semánticas y conjuntistas tienen un origen común.

En sentido contrario, se tiene la distinción de Ramsey (1925) en "contradicciones" lógicas y epistemológicas —que actualmente suelen denominarse paradojas conjuntistas y semánticas— o la réplica de Grattan-Guinness a Priest (1994):

Grattan-Guinness (1998) argumenta que al menos una paradoja escapa al esquema propuesto por Priest y, lo que es más importante, que el hecho de que varias paradojas se ajusten a un mismo esquema no tiene por qué conllevar que todas ellas tengan una explicación común.

Podemos matizar el argumento de Grattan-Guinness diciendo que las paradojas que comparten un mismo esquema tendrán una explicación común si la causa de la paradoja es ese esquema. Priest (1994) deja claro que su esquema conduce a contradicción pero este hecho es fácil de aceptar y no es en sí mismo paradójico. Lo paradójico es más bien que haya oraciones y definiciones aparentemente inocuas que se ajustan a ese esquema y, desde luego, no es evidente que esto ocurra por el mismo motivo en todos los casos.

Aunque creo que los planteamientos de Russell y Priest no son plenamente satisfactorios, no es objetivo de este trabajo discutirlos en profundidad. En los próximos apartados me limitaré a proponer unos criterios unificadores más relacionados con las propuestas que hemos hecho acerca de la paradoja del mentiroso.

5.5.2. Malas definiciones

La paradoja heterológica y las paradojas señaladas por Whitehead y Russell en *Principia Mathematica*[26] (con excepción de la del mentiroso) pueden entenderse como paradojas basadas en definiciones: definición del predicado heterológico, del conjunto de todos los conjuntos que no son miembros de sí mismos, del menor ordinal indefinible, el menor entero que no se puede nombrar, en inglés, con menos de diecinueve sílabas, etc. En un sentido general admiten un diagnóstico común: se basan en definiciones incorrectas, aunque la prueba de que son incorrectas será seguramente distinta para distintas paradojas.

Como es sabido, la definición de un objeto μ mediante una descripción definida tiene la forma:

$$\mu =_{def} \iota x\ \varphi(x) \tag{5.115}$$

donde $\varphi(x)$ es una función proposicional con argumento x. Para que una definición de la forma anterior sea correcta es necesario y suficiente que se cumpla:

$$\exists x\ \varphi(x) \tag{5.116}$$

[26] En este trabajo hay una sencilla descripción de las mismas a partir de la página 13.

y

$$\forall y \forall z\{[\varphi(y) \wedge \varphi(z)] \rightarrow y = z\} \tag{5.117}$$

Si la definición (5.115) es correcta, serán verdaderas las siguientes proposiciones:

$$\mu = [\iota x\ \varphi(x)] \tag{5.118}$$

$$\varphi(\mu) \tag{5.119}$$

$$\forall y[\varphi(y) \rightarrow y = \mu] \tag{5.120}$$

En cambio, si la definición no es correcta, la descripción definida"$\iota x\ \varphi(x)$" y, por tanto, el nombre "μ", son términos sin referencia.

A modo de ejemplo, consideremos ahora, en una lógica clásica, las paradojas de Russell y Weyl (esta última también llamada de Grelling o paradoja heterológica).

En el primer caso se trata de definir un conjunto. La forma general de definir un conjunto, C, es

$$C =_{def} \{x/\psi(x)\} \tag{5.121}$$

o, mediante una descripción definida (que hace uso de la relación de pertenencia, $\in$):

$$C =_{def} \iota y(\forall x\ (x \in y \leftrightarrow \psi(x))) \tag{5.122}$$

Si la definición es correcta, la aplicación de (5.119) nos dice que:

$$\forall x\ (x \in C \leftrightarrow \psi(x)) \tag{5.123}$$

Si los conjuntos están en el ámbito del cuantificador, por instanciación universal deducimos:

$$C \in C \leftrightarrow \psi(C) \tag{5.124}$$

Esta condición es falsa cuando $\psi(x)$ es $x \notin x$, porque entonces $\psi(C)$ será $C \notin C$. En tal caso la definición (5.122) no será, pues, correcta. Precisamente, este es el caso de la paradoja de Russell donde se pretende definir un conjunto,R, mediante

$$R =_{def} \iota y(\forall x\ (x \in y \leftrightarrow x \notin x)) \tag{5.125}$$

Así, la simple aplicación de criterios bien conocidos sobre la corrección de las definiciones nos deja claro que ni la definición (5.125) ni, por tanto, la descripción "el conjunto de todos los conjuntos que no son miembros de sí mismos", son correctas.

En el caso de la paradoja heterológica se trata de definir un predicado. La forma general de definir un predicado monádico, P, es

$$\forall x\ [Px \leftrightarrow_{def} \psi(x)] \tag{5.126}$$

o, mediante una descripción definida:

$$P =_{def} \iota B(\forall x\ Bx \leftrightarrow \psi(x)) \tag{5.127}$$

Si la definición anterior es correcta, la aplicación de (5.119) nos dice que:

$$\forall x\ (Px \leftrightarrow \psi(x)) \tag{5.128}$$

Si los nombres de predicado están en el ámbito del cuantificador, y $\ulcorner P \urcorner$ es el nombre del predicado P, podemos deducir:

$$P\ulcorner P\urcorner \leftrightarrow \psi(\ulcorner P\urcorner) \tag{5.129}$$

Para definir el predicado heterológico, H: 1/ admitiremos cuantificación sobre predicados; 2/ admitiremos términos de la forma $\ulcorner P \urcorner$, siendo P un predicado; 3/ definiremos $\mathcal{V}_{\mathfrak{M}}(P) =_{def} \mathfrak{I}_{\mathfrak{M}}(P)$, para cualquier predicado P; 4/ haremos uso de una función de denotación, δ, que se caracteriza porque dado un término, τ, con $\tau \in dom(\mathcal{V}_{\mathfrak{M}})$, se verifica $\mathcal{V}_{\mathfrak{M}}(\delta^{\ulcorner}\tau^{\urcorner}) = \mathcal{V}_{\mathfrak{M}}(\tau)$. Con estas herramientas, la definición del predicado H, usando la forma (5.126) es:

$$\forall x\ (Hx \leftrightarrow_{def} (\exists Q(\delta(x) = Q \wedge \neg Q(x)))) \tag{5.130}$$

Si la definición anterior fuese correcta, la aplicación de (5.129) a la definición anterior nos permitiría asertar:

$$H\ulcorner H\urcorner \leftrightarrow (\exists Q(\delta(\ulcorner H\urcorner) = Q \wedge \neg Q(\ulcorner H\urcorner))) \tag{5.131}$$

y, puesto que δ es la función denotación:

$$H^{\ulcorner}H^{\urcorner} \leftrightarrow (\exists Q(H = Q \wedge \neg Q(^{\ulcorner}H^{\urcorner}))) \tag{5.132}$$

Finalmente, dada la equivalencia entre $\exists Q(H = Q \wedge \neg Q(^{\ulcorner}H^{\urcorner})))$ y $\neg H^{\ulcorner}H^{\urcorner}$, obtendríamos la contradicción:

$$H^{\ulcorner}H^{\urcorner} \leftrightarrow \neg H^{\ulcorner}H^{\urcorner} \tag{5.133}$$

Nuestra conclusión, es pues, que la definición (5.130) no es correcta.

Aunque la formalización de la paradoja heterológica ha sido algo más compleja que la de la paradoja de Russell, de nuevo la simple aplicación de criterios bien conocidos sobre la corrección de las definiciones nos ha dejado claro que la definición del predicado heterológico es incorrecta.

5.5.3. Planteamientos unificadores

El análisis anterior sugiere un planteamiento unificador de aquellas paradojas basadas en una definición: clarificarla (idealmente usando un lenguaje formal) y comprobar que no cumple los requisitos exigibles a toda definición para que sea correcta.

Sin embargo, la paradoja del mentiroso tiene su origen en una oración sintácticamente correcta. A diferencia de lo que ocurre con las paradojas de Russell, de Weyl o de Berry, no hay una definición propia de la paradoja del mentiroso: lo característico de la paradoja del mentiroso es una oración —o, en algunas versiones, un sistema de oraciones—.

Parece razonable pensar que así como una definición sintácticamente correcta puede ser semánticamente incorrecta (es el caso de la definición del conjunto de la paradoja de Russell o la del predicado heterológico), una oración sintácticamente correcta puede ser semánticamente incorrecta (sería el caso de la oración del mentiroso). Y, así como el *definiens* de una definición semánticamente incorrecta de la forma $\mu =_{def} \iota x\, \varphi(x)$ es un término sin referencia (y, por consiguiente, también el *definiendum*), la situación análoga para una oración semánticamente incorrecta sería que dicha oración careciera de valor veritativo. Pero como sabemos, este planteamiento por sí solo no resuelve la paradoja del mentiroso reforzada.

Consiguientemente, es más difícil la resolución de las paradojas basadas en una oración (como la del mentiroso) que la de, al menos, algunas paradojas basadas en una definición (como las de Russell y Weyl). Porque mientras éstas se resuelven cuando se comprueba que el término definido no tiene referencia, las primeras no quedan resueltas por el solo hecho de afirmar que la oración no tiene valor de verdad. Es más, este supuesto no es siquiera posible si, consideramos con un tercer valor de verdad a toda oración que no sea verdadera ni falsa.

Ahora bien, hemos resuelto la paradoja del mentiroso basándonos en que la oración correspondiente contiene un nombre que no tiene referencia en su aparición en dicha oración. Por tanto, es sencillo ver una primera unificación en la solución a la paradoja del mentiroso y la solución a las paradojas basadas en definiciones: en todos los casos la solución se basa en que (una aparición de) un nombre que aparentemente tiene referencia, realmente no la tiene.

Es importante constatar que, en el uso de lenguajes formales para estudiar la paradoja del mentiroso, también hay una definición que no es correcta: se trata de la definición de la función de valuación $\mathcal{V}_{\mathfrak{M}}$ (tal como la hemos definido antes de introducir la noción de los contextos referencialmente anuladores). Obsérvese, por ejemplo, que si bien la definición 4.6 (p. 72) es correcta para un lenguaje interpretado de primer orden clásico, no lo es en cuanto enriquecemos su capacidad expresiva con términos de cita, un predicado de verdad y una función de diagonalización. La prueba de ello es que hay oraciones para las que no queda determinado su valor de verdad —en un lenguaje en que toda oración debe tenerlo— como es el caso de la oración del mentiroso reforzada, λ, o de la oración del veraz, θ. Como sabemos, para el valor de verdad de la primera, establece una condición imposible:

$$\mathcal{V}_{\mathfrak{M}}(\lambda) = (C_{\neg} \circ C_T)(\mathcal{V}_{\mathfrak{M}}(\lambda)) \tag{5.134}$$

y, para el de la segunda, solo establece una condición, que se cumple tanto si la oración es verdadera como si es falsa:

$$\mathcal{V}_{\mathfrak{M}}(\theta) = C_T(\mathcal{V}_{\mathfrak{M}}(\theta)) \tag{5.135}$$

Así pues ni $\mathcal{V}_{\mathfrak{M}}(\lambda)$ ni $\mathcal{V}_{\mathfrak{M}}(\theta)$ están bien definidas y, por ende, tampoco lo está $\mathcal{V}_{\mathfrak{M}}$. Algo similar podemos decir para las definiciones de $\mathcal{V}_{\mathfrak{M}}$ establecidas inicialmente para los otros tipos de lenguajes formales analizados. Además la formalización refleja lo que ocurre en lenguaje natural frente a oraciones como las del mentiroso

reforzada y del veraz: no parece haber forma coherente de establecer su valor veritativo.

Podría pensarse que si una paradoja como la de Russell queda resuelta al constatar que se basa en una definición incorrecta, la del mentiroso quedaría resuelta, en el lenguaje formal, al constatar que la función de valuación, $\mathcal{V}_{\mathfrak{M}}$, no se define correctamente. Pero esto no basta, pues no podemos considerar resuelta la paradoja del mentiroso reforzada mientras no sepamos determinar el valor de verdad (o, en su caso, su falta de valor de verdad) de la correspondiente oración. Necesitamos tener una función de valuación de las oraciones correctamente definida.

De todos modos, el planteamiento del problema de la paradoja del mentiroso (y similares), como un problema de determinación del valor de verdad de la oración asociada, es extensible a las otras paradojas autorreferenciales como la paradoja heterológica y las de *Principia Mathematica*. Porque en cada una de estas paradojas, basadas en una definición, se deriva una contradicción que puede verse como la afirmación de que una misma oración tiene que ser verdadera y falsa a la vez. Por ejemplo, si llamamos R al conjunto de todos los conjuntos que no se contienen a sí mismos, no parece haber un modo coherente de decidir si la oración $R \in R$ es verdadera o es falsa, ya que obtenemos la conclusión

$$R \in R \leftrightarrow \neg(R \in R) \tag{5.136}$$

la cual, asumiendo una lógica clásica, equivale a decir que $R \in R$ es verdadero si y solo si $R \in R$ no lo es. En el caso de la paradoja heterológica, la oración problemática es $H^{\ulcorner}H^{\urcorner}$ (formalización de "heterológico es heterológico") y no es difícil adivinar cuáles serían las oraciones problemáticas para las paradojas de Richard, Burali-Forti, Berry, etc.

El problema común es pues establecer un procedimiento que permita asignar coherentemente un valor de verdad a cada oración paradójica y cumpla los requisitos para una solución satisfactoria. En principio, la diferencia estriba en encontrar la causa de cada paradoja. En el caso de la paradoja del mentiroso y problemas afines (como el de la oración del veraz y, en general, el de los sistemas de oraciones con referencias semánticas aparentes entre ellas), hemos tenido que dedicar una parte importante de este trabajo a dicho objetivo. En las paradojas basadas en una definición (como la de Russell o la heterológica) el principal sospechoso de su causa es claro: la propia definición. Cuando esté justificado que la definición es

incorrecta, el término definido carecerá de referencia y la paradoja desaparecerá. Por ejemplo, aceptando que el supuesto conjunto R de la paradoja de Russell no está bien definido, el término "R" carecerá de referencia y la oración $R \in R$ no será verdadera ni falsa. La paradoja no surgirá, porque el *definiens* de una definición incorrecta no puede usarse como equivalente al *definiendum* (en el caso que nos ocupa, no podemos suponer que, dado un conjunto A, $A \in R$ equivale a $A \notin A$).

Hasta el momento hemos encontrado dos importantes aspectos comunes a la paradoja del mentiroso, a la heterológica y a las otras paradojas de *Principia Mathematica*: 1/ se trata de asignar justificadamente un valor de verdad a cada oración paradójica y 2/ esto se consigue justificando que un nombre que aparentemente tiene referencia realmente no la tiene.

Considero que un enfoque similar al utilizado para estudiar la paradoja del mentiroso será clarificador en el estudio de las paradojas de *Principia Mathematica* y otras. El uso de lenguajes formales disipa las ambigüedades —como las que introduce el concepto "definible" en las paradojas de Berry y Richard— y, tal vez, permita ver más aspectos comunes a la solución de todas estas paradojas. Aunque la búsqueda de tal solución queda fuera del ámbito del presente trabajo, no puedo dejar de reseñar un aspecto común a todas estas paradojas que ha sido clave en el diagnóstico y solución de la del mentiroso: en todas ellas interviene algún predicado o relación de desentrecomillado combinado con la autorreferencia.

Que en la paradoja heterológica y en las paradojas de *Principia Mathematica* hay autorreferencia es sobradamente conocido. Nos centraremos en mostrar que también interviene un predicado o relación de desentrecomillado.

- Paradoja heterológica. ¿Cómo podemos definir el predicado heterológico, H, en un lenguaje interpretado (mediante un modelo $\mathfrak{M}$) del tipo de los utilizados en este trabajo? Como se trata de un predicado sobre predicados monádicos necesitamos términos que citen predicados monádicos. Si P es un predicado monádico el término que lo cita será $\ulcorner P \urcorner$. La definición de H consistirá en que, dado cualquier predicado monádico P, y un término ν tal que $\mathfrak{M} \models (\nu = \ulcorner P \urcorner)$:

$$\mathcal{V}_{\mathfrak{M}}(H\nu) =_{def} \mathcal{V}_{\mathfrak{M}}(\neg P\ulcorner P \urcorner) \tag{5.137}$$

  En particular:

$$\mathcal{V}_{\mathfrak{M}}(H\ulcorner P \urcorner) =_{def} \mathcal{V}_{\mathfrak{M}}(\neg P\ulcorner P \urcorner) \tag{5.138}$$

donde queda patente que ha habido desentrecomillado de P en el sentido de que el valor de verdad de $H\ulcorner P\urcorner$ se puede expresar en función del valor de verdad de una oración en la que P aparece sin entrecomillar (aunque en este caso también aparece entrecomillado).

- Paradoja de la relación T.[27] Es muy similar a la anterior. Por definición de T debe cumplirse:

$$\mathcal{V}_{\mathfrak{M}}(T\ulcorner R\urcorner, \ulcorner S\urcorner) =_{def} \mathcal{V}_{\mathfrak{M}}(\neg R\ulcorner R\urcorner, \ulcorner S\urcorner) \tag{5.139}$$

donde se observa que la relación T produce un desentrecomillado de su primer argumento.

- Paradoja de Russell. Si suponemos que a cada conjunto corresponde una función proposicional que lo define, podremos formalizar los conjuntos mediante términos de la forma $\ulcorner\varphi(x)\urcorner$, siendo $\varphi(x)$ una fórmula con una única variable libre. Es decir, $\ulcorner\varphi(x)\urcorner$ es un término cuya interpretación es un conjunto. Si la fórmula $\varphi(x)$ representa la función proposicional $\Phi(z)$, entonces el término $\ulcorner\varphi(x)\urcorner$ representa el conjunto $\{z/\Phi(z)\}$. Más formalmente:

$$\mathcal{V}_{\mathfrak{M}}(\ulcorner\varphi(x)\urcorner) = \{z/\mathcal{V}_{\mathfrak{M},s[z/x]}(\varphi(x)) = \mathfrak{v}\} \tag{5.140}$$

La siguiente cuestión es: dado un símbolo de relación diádica, P, ¿qué requisito debe cumplirse para que P represente la relación de pertenencia? Si a es un término cuya referencia es un objeto A, ν es un término tal que $\mathfrak{M} \models (\nu = \ulcorner\varphi(x)\urcorner)$, y la fórmula $\varphi(x)$ representa la función proposicional $\Phi(z)$, entonces

$$Pa, \nu \tag{5.141}$$

representa la proposición $A \in \{z/\Phi(z)\}$, que a su vez, equivale a $\Phi(A)$. En definitiva, el requisito para que P represente la relación de pertenencia, es simplemente:

$$\mathcal{V}_{\mathfrak{M}}(Pa, \nu) = \mathcal{V}_{\mathfrak{M}}(\varphi(a)) \tag{5.142}$$

[27] Recordemos que consiste en lo siguiente: sea T la relación entre dos relaciones R y S que se da cuando R no tiene la relación R con S. ¿Tiene T la relación T con T? Cualquier posible respuesta conduce a una contradicción.

En particular,

$$\mathcal{V}_{\mathfrak{M}}(Pa, \ulcorner\varphi(x)\urcorner) = \mathcal{V}_{\mathfrak{M}}(\varphi(a)) \tag{5.143}$$

donde se pone de manifiesto que la relación P realiza una función de desentrecomillado de su segundo argumento.

- Paradojas de definibilidad. Se incluyen aquí las demás paradojas de *Principia Mathematica*, es decir, las paradojas del menor ordinal indefinible, la de Berry, la de Richard y la de Burali-Forti. En las tres primeras interviene explícitamente la noción de definibilidad, pero la de Burali-Forti también puede encuadrarse en este grupo. Recordemos que esta paradoja surge al considerar "el ordinal de la serie de todos los ordinales", luego debemos preguntarnos si la función proposicional "x es la serie de todos los ordinales" *define* o no un ordinal.
  Para las paradojas de definibilidad formalizaremos una relación diádica que indique si un objeto es definido por una función proposicional. Por tanto, admitiremos términos de cita de la forma $\ulcorner\varphi(x)\urcorner$. Nótese que ahora el término $\ulcorner\varphi(x)\urcorner$ no representa un conjunto, sino simplemente la fórmula $\varphi(x)$. Para que el símbolo de relación diádica B represente la relación de definibilidad, debe cumplirse para cualesquiera términos cerrados μ, ν:

$$\mathcal{V}_{\mathfrak{M}}(B\mu, \nu) =_{def} \begin{cases} \mathfrak{v} & si\ A \\ \mathfrak{f} & en\ otro\ caso \end{cases} \tag{5.144}$$

siendo A: "existe una fórmula con una única variable libre, $\varphi(x)$, tal que $\mathcal{V}_{\mathfrak{M}}(\nu) = \varphi(x)$ y se cumple $\mathcal{V}_{\mathfrak{M}}(\varphi(\mu)) = \mathfrak{v}$ y $\mathcal{V}_{\mathfrak{M}}(\forall y(\varphi(y) \to y = \mu)) = \mathfrak{v}$". En el caso particular de que ν sea un término de la forma $\ulcorner\varphi(x)\urcorner$, se tiene:

$$\mathcal{V}_{\mathfrak{M}}(B\mu, \ulcorner\varphi(x)\urcorner) =_{def} \begin{cases} \mathfrak{v} & si\ \mathcal{V}_{\mathfrak{M}}(\varphi(\mu)) = \mathfrak{v} = \mathcal{V}_{\mathfrak{M}}(\forall y(\varphi(y) \to y = \mu)) \\ \mathfrak{f} & en\ otro\ caso \end{cases} \tag{5.145}$$

En una lógica bivaluada, la expresión anterior puede reducirse a:

$$\mathcal{V}_{\mathfrak{M}}(B\mu, \ulcorner\varphi(x)\urcorner) =_{def} \mathcal{V}_{\mathfrak{M}}(\varphi(\mu) \wedge \forall y(\varphi(y) \to y = \mu)) \tag{5.146}$$

donde es evidente que el predicado B produce un desentrecomillado de su segundo argumento. En una lógica trivaluada, la expresión (5.145) se puede

reducir, con la ayuda del predicado de verdad (T) a:

$$\mathcal{V}_{\mathfrak{M}}(B\mu, \ulcorner\varphi(x)\urcorner) =_{def} \mathcal{V}_{\mathfrak{M}}(T\ulcorner\varphi(\mu) \wedge \forall y(\varphi(y) \rightarrow y = \mu)\urcorner) \tag{5.147}$$

y por tanto, a:

$$\mathcal{V}_{\mathfrak{M}}(B\mu, \ulcorner\varphi(x)\urcorner) =_{def} C_T(\mathcal{V}_{\mathfrak{M}}(\varphi(\mu) \wedge \forall y(\varphi(y) \rightarrow y = \mu))) \tag{5.148}$$

donde nuevamente puede observarse que B produce un desentrecomillado de su segundo argumento.

6

Conclusiones y futuros desarrollos

6.1. El problema

El estudio de las principales propuestas de solución de la paradoja del mentiroso y afines ha sido un primer paso para comprender el problema con más profundidad e identificar los errores más destacados en que incurren. Ello nos ha servido para establecer (apartado 3) una lista de requisitos que debería cumplir una solución satisfactoria, así como para corroborar la convicción, generalmente aceptada, de que ninguna de esas propuestas resuelve satisfactoriamente la paradoja. En mi opinión, el hecho de que la lista de requisitos sea bastante reducida y, sin embargo, sirva para invalidar las propuestas de solución, de naturaleza muy diferente, que hemos analizado, la hace especialmente interesante.

Desde un punto de vista bastante general, una paradoja es —como señalamos en el apartado 4.1.1—, un razonamiento en el que, a partir de supuestos aparentemente verdaderos, y principios aparentemente correctos, se deduce una conclusión aparentemente inaceptable. Las *auténticas* paradojas se caracterizan porque no es sencillo encontrar cuál (o cuáles) de las apariencias es engañosa. En este sentido, la paradoja del mentiroso es una *auténtica* paradoja pues no hay acuerdo en determinar cuál es la causa de la misma, cuál es la apariencia engañosa. Mientras las propuestas *inconsistentes* sostienen que la oración del mentiroso es verdadera y falsa a la vez y que esto —en contra de las apariencias— es aceptable, otras, como la de Skyrms (1970), afirman que la conclusión (la oración del mentiroso reforzada es verdadera y no verdadera a la vez) es inaceptable pero que el razonamiento que lleva a ella es —en contra de las apariencias— falaz. Por el contrario, la mayoría de propuestas no cuestionan que la conclusión sea realmente inaceptable ni que

haya errores deductivos en el razonamiento que lleva a ella; lo que cuestionan es algunos de los supuestos semánticos habitualmente aceptados que son sustituidos por otros nuevos (por ejemplo, el principio del círculo vicioso, en el caso de Russell, la distinción lenguaje/metalenguaje de Tarski, o la vaguedad del predicado *verdadero* de McGee).

Los intentos, como el de Skyrms, de solucionar una paradoja cambiando las reglas de deducción, siguen a mi juicio, un camino poco recomendable. Las reglas de deducción están muy bien establecidas y su cambio, además de ser difícil de justificar, provocará seguramente más problemas que beneficios. Tanto es así que algunos autores caracterizan una paradoja como un argumento válido que, a partir de proposiciones aparentemente verdaderas, llega a una conclusión inaceptable (la validez del argumento y la inaceptabilidad de la conclusión no se cuestionan). En esta línea se manifiesta Rescher (2001) para quien en la genuina paradoja la derivación es válida, lo cual la diferencia de la falacia, donde esa validez es solo aparente:

> With fallacies we have apparent rather than real derivation, but the reasoning of a genuine *paradox* must be cogent.[1]

Tras haber estudiado las propuestas *inconsistentes* y llegar a la conclusión de que no resuelven satisfactoriamente la paradoja del mentiroso, creemos sensato —como la mayoría de los lógicos— considerar inaceptable que de un argumento válido con premisas verdaderas se obtenga una conclusión contradictoria.

Con estas consideraciones podemos llegar a una nueva e interesante definición de paradoja:

> un conjunto inconsistente de proposiciones cada una de las cuales es aparentemente verdadera

Este es, con algunos matices, el planteamiento de Rescher (2001); un planteamiento que será muy útil para comprender diversas conclusiones del presente trabajo.

La primera observación importante es que para resolver una paradoja se necesita negar una o varias de las proposiciones que forman el conjunto inconsistente de modo que se restablezca la consistencia. Rescher (2001) analiza con este planteamiento más de ciento treinta paradojas y muestra que la lógica por sí sola no

[1] Rescher (2001, p. 6, nota al pie 8).

sirve para decidir qué proposiciones rechazar y cuáles mantener. Es preciso considerar criterios extra-lógicos que den prioridad a unas proposiciones sobre otras, a las más plausibles sobre las menos plausibles. Sin embargo, no siempre es sencillo establecer esas prioridades ni justificar por qué rechazamos ciertas proposiciones. Nuestra lista de requisitos nos ha servido de ayuda para resolver este problema.

Una vez que se plantea la resolución de una paradoja como un problema de priorización entre proposiciones más y menos plausibles dentro de un conjunto inconsistente de proposiciones, es claro que, si son razonables distintos criterios de priorización, la paradoja admitirá más de una solución razonable. A pesar de ello, la conclusión de nuestro estudio es que ninguna de las propuestas de solución de la paradoja del mentiroso analizadas es aceptable.

Frente a un problema de la naturaleza de la paradoja del mentiroso, cuya solución se muestra esquiva, se hace imprescindible caracterizarlo lo mejor posible antes de intentar abordarlo. Hemos partido de la excelente caracterización de Martin (1984, pp. 1 y 2), como un conjunto inconsistente de las dos siguientes proposiciones:

(S) Hay una oración que dice de sí misma únicamente que no es verdadera

(T) Una oración es verdadera ssi las cosas son como la oración dice que son

Este planteamiento nos conduce a la necesidad de negar (S) o negar (T), justificadamente, para resolver la paradoja. Bien es verdad que, como el mismo Martin (1984, p. 3) señala, hay un diagnóstico consistente en afirmar que el argumento de la incompatibilidad entre (S) y (T) es equivocado, pero creo que se trata de un diagnóstico demasiado artificioso, demasiado forzado y, por ende, poco satisfactorio. Sin embargo, la propia existencia de este diagnóstico es indicativa de la dificultad de justificar que (S) o (T) no sean verdaderas. Ante esta dificultad hemos creído necesario clarificar aún mejor el problema y situarlo en el contexto de otros problemas similares. Para ello hemos considerado que el mejor camino es el uso de lenguajes formales interpretados, teniendo siempre muy presente que no se trata de "resolver la paradoja dentro del lenguaje formal" pues esto ya lo han hecho otros a costa de reducir la capacidad expresiva de dicho lenguaje y *expulsar* la paradoja al metalenguaje.

Hemos comenzado usando lenguajes interpretados de primer orden clásicos a los que hemos añadido capacidad de cita y predicados de desentrecomillado —de

los que el predicado *verdadero* es un caso particular— dado que la oración del mentiroso reforzada se cita a sí misma (al menos aparentemente) y usa el predicado verdadero. De este modo podemos reflejar en el lenguaje el supuesto(T) de Martin. Por otra parte, hemos mostrado que, mediante el uso de una función de diagonalización u otros métodos, podemos conseguir que, un lenguaje interpretado de primer orden con capacidad de cita sea fuertemente autorreferencial, lo que permite reflejar en el lenguaje formal el supuesto (S) de Martin. Y la principal conclusión, al intentar formalizar la paradoja del mentiroso, es que si añadimos capacidad de cita a un lenguaje interpretado de primer orden clásico, el lenguaje resultante no puede tener un predicado de verdad y, al mismo tiempo, ser fuertemente autorreferencial.

Al tener en cuenta que una paradoja se caracteriza por un conjunto inconsistente de proposiciones (cada una de ellas aparentemente verdadera) es claro que, en un sistema formal consistente, será imposible que todas esas proposiciones sean verdaderas. Por eso, no es sorprendente que en él sea imposible tener a la vez un predicado de verdad que funcione según (T) y una oración que diga de sí misma que no es verdadera. Más bien al contrario, el hecho de que así ocurra indica que la formalización es adecuada.

La intuición mayoritaria en cuanto al estatus veritativo de la oración del mentiroso es que no es verdadera ni falsa. Una solución que tuviese esta conclusión no tendría cabida en un lenguaje formal donde cualquier oración deba ser verdadera o ser falsa. Resulta pues pertinente, en el estudio de la paradoja mediante lenguajes formales, utilizar lenguajes donde las oraciones puedan no ser verdaderas ni falsas. Hemos encontrado (apartado 4.1.3.1) que la mejor opción es usar lenguajes trivaluados en los que aparece un tercer valor de verdad, que llamamos *impropio* o *indefinido*.

Una primera conclusión, al intentar reflejar en el lenguaje formal trivaluado las propiedades del lenguaje natural que conducen a la paradoja del mentiroso reforzada, es que si dotamos al lenguaje de suficiente capacidad expresiva para poder decir en él, de una oración *indefinida*, que no es verdadera, el lenguaje no permite la existencia de una oración que diga de sí misma que no es verdadera.

Lo interesante es que podemos encuadrar esta conclusión dentro de un diagnóstico más general consistente en que, si dotamos de suficiente capacidad expresiva al lenguaje formal, no pueden coexistir el predicado de verdad y la autorreferencialidad fuerte, porque podríamos formar una oración —la paradoja del mentiroso

reforzada es solo un ejemplo— cuyo valor de verdad tendría que cumplir una condición imposible. Significativamente, este diagnóstico del problema es válido con independencia de numerosas cuestiones semánticas, como si las oraciones tienen dos, tres o más valores de verdad; si el predicado de verdad es vago o si la generalización a una lógica multivaluada de la caracterización tarskiana de ese predicado se debe realizar de una u otra forma.

Otro aspecto investigado mediante lenguajes formales adecuados es si el considerar que son los contenidos enunciativos de las oraciones y no éstas los auténticos portadores de verdad/falsedad, puede abrir un camino en la resolución de la paradoja del mentiroso. Porque una solución aparente consiste en afirmar que la oración del mentiroso reforzada carece de contenido enunciativo. Ahora bien, uno de los requisitos para que una solución sea satisfactoria es que no dé lugar a nuevas paradojas. Requisito que no se cumple en este caso, pues la oración "esta oración carece de contenido enunciativo o tiene un contenido enunciativo falso" constituye una versión de la paradoja más reforzada que no se resuelve afirmando que la oración carece de contenido enunciativo —entonces afirmar la oración sería afirmar algo verdadero, lo que supondría que la oración tiene un contenido verdadero— ni tampoco afirmando que tiene un contenido enunciativo —si ese contenido se supone verdadero la oración resulta tener un contenido enunciativo falso y viceversa—.

A los lenguajes formales interpretados que hemos usado para investigar la paradoja del mentiroso bajo la perspectiva de que los portadores de verdad son los contenidos enunciativos, los hemos denominado lenguajes interpretados enunciativos parciales. Se trata de lenguajes, originales de este trabajo, que permiten la existencia de oraciones con y sin contenido enunciativo y donde la interpretación de una oración es, si lo tiene, su contenido enunciativo, en vez de su valor de verdad como ocurre en los lenguajes formales extensionales. Además hay una función de extensionalización que, entre otras cosas, asocia a cada contenido enunciativo su valor de verdad.

Al pretender formalizar la oración paradójica "esta oración carece de contenido enunciativo o tiene un contenido enunciativo falso", en un lenguaje interpretado enunciativo parcial, la principal conclusión obtenida es que hay una incompatibilidad entre la autorreferencialidad fuerte y la existencia de predicados de desentrecomillado. Se trata de una conclusión casi idéntica a la que habíamos obtenido en el estudio de la paradoja con los anteriores tipos de lenguajes formales. La única

diferencia es que en vez de hablar de predicado de verdad, hablamos de predicados de desentrecomillado. Puesto que el predicado de verdad es un caso particular de los predicados de desentrecomillado, la conclusión obtenida corrobora y generaliza las conclusiones que obtuvimos en aquellos lenguajes en que la interpretación de una oración era su valor de verdad. El considerar los contenidos enunciativos como portadores de verdad no altera, pues, en lo esencial el diagnóstico de la paradoja del mentiroso.

El proceso de caracterización de la paradoja del mentiroso culmina con importantes generalizaciones. La primera consiste en el uso de lenguajes interpretados enunciativos trivaluados, de los que los lenguajes formales utilizados con anterioridad son un caso particular. La segunda muestra que hay un problema de términos paradójicos similar al de las oraciones paradójicas como la del mentiroso: el papel del predicado de desentrecomillado lo hace una función de desentrecomillado (de la que la función *denotación* sería un ejemplo) y la incompatibilidad aparecería entre la existencia de funciones de desentrecomillado y la autorreferencialidad de términos. La tercera generalización viene de la mano de la función de tarskificación que permite una forma diferente de autorreferencia, lo que nos lleva a definir la noción de lenguaje autorreferencial, una noción más general que la de lenguaje fuertemente autorreferencial.

Con estas generalizaciones, obtenemos la conclusión de que, en un lenguaje interpretado enunciativo trivaluado, la simple autorreferencialidad es incompatible con la existencia de predicados de desentrecomillado. Esto supone un mayor refinamiento en la caracterización del problema. Pero al mismo tiempo, el problema va más allá de la oración del mentiroso en sus distintas versiones. Por una parte, como ya hemos dicho, hay también términos paradójicos (en cuyo estudio no profundiza este trabajo) y, por otra, en el caso de las oraciones, el problema en que se enmarca la paradoja del mentiroso es el siguiente: si en un lenguaje autorreferencial, hubiese un predicado de desentrecomillado, D; a partir de una fórmula de la forma $\rho(Dx)$, se podría construir una oración, ψ_ρ, que debería cumplir $\mathcal{V}_{\mathfrak{M}}(\psi_\rho) = (C_\rho \circ C_D)(\mathcal{V}_{\mathfrak{M}}(\psi_\rho))$. En el caso de la oración del mentiroso, y otras, la función $(C_\rho \circ C_D)$ no tiene ningún punto fijo, por lo que la igualdad anterior es imposible. Pero en otros casos, como la oración del veraz, la función $(C_\rho \circ C_D)$ tiene varios puntos fijos y ello supone también un problema a la hora de determinar $\mathcal{V}_{\mathfrak{M}}(\psi_\rho)$. Hay casos en los que la función $(C_\rho \circ C_D)$ tiene un único punto fijo —por ejemplo, cuando ψ_ρ es la formalización de la oración "esta oración carece de

contenido enunciativo (completo)"—; pero incluso en ellos no hay acuerdo sobre el auténtico valor de verdad de la oración. En este sentido hemos distinguido dos planteamientos que hemos dado en llamar G y MK: en el primero aquellas oraciones a las que se puede asignar un único valor de verdad coherentemente, tienen ese valor de verdad; en el segundo, se exige que la oración tenga un contenido enunciativo completo para que pueda ser verdadera o falsa. Con el planteamiento G, la oración "esta oración carece de contenido enunciativo (completo)" es falsa. La opción MK requiere establecer un criterio para determinar, si dada una oración, tiene o no un contenido enunciativo. Para Mackie la oración "esta oración carece de contenido enunciativo (completo)" no tiene contenido por lo que no es verdadera ni falsa. Pero esta solución ha de enfrentarse a diversos problemas. Uno de ellos aparece en el ejemplo que nos ocupa: si la oración carece de contenido, afirmar la oración es afirmar una verdad, lo cual contradice que la oración carezca de contenido.

En resumen, el proceso de caracterización de la paradoja del mentiroso (y afines) que hemos seguido nos ha ido aclarando qué aspectos son *responsables* de la misma y cuáles no.

La paradoja no se debe: 1/ a que tomemos un número insuficiente o excesivo de valores de verdad; 2/ a que se considere que los portadores de verdad son las oraciones o que se considere que son sus contenidos enunciativos; 3/ a que el predicado de verdad sea vago y no lo tengamos en cuenta; 4/ a la forma de generalizar a una lógica multivaluada la condición de adecuación material de Tarski.

Todo indica que el problema que plantea la paradoja es el de la incompatibilidad entre cierta capacidad referencial del lenguaje (la capacidad de autorreferencia en el caso de la versión habitual de la paradoja del mentiroso) y la existencia de predicados de desentrecomillado.

El haber distinguido los aspectos responsables de la paradoja de los que no lo son, nos permite que la búsqueda de una solución a la misma esté bien orientada. Y, si nuestras conclusiones son acertadas, debemos investigar qué es lo que falla en nuestros conceptos habituales sobre referencia o sobre predicados de desentrecomillado, para que se produzca la paradoja del mentiroso y similares.

6.2. La propuesta básica

Desde luego, descartamos soluciones como las de Russell o Tarski, donde la referencia se restringe mucho más de lo necesario, o visiones del concepto de verdad como las de Gupta y Belnap, Barwise y Etchemendy, McGee, Priest, etc. que, por distintos motivos —analizados en el apartado 2.3—, no servían para resolver el problema de la paradoja.

Nuestra investigación acerca de una mínima modificación del funcionamiento de los predicados de desentrecomillado de modo que se resuelva el problema planteado por la paradoja nos ha llevado a la conclusión de que inevitablemente traería consigo una restricción inaceptable a la capacidad expresiva del lenguaje formal. Como en la propuesta de Kripke, la paradoja desaparecería en el lenguaje formal y reaparecería en el metalenguaje.

En cuanto a la autorreferencia hemos llegado a dos conclusiones importantes: 1/ el uso de nuestros lenguajes formales permite distinguir la autorreferencia inocua de la *peligrosa*; 2/ en el caso de la referencia *peligrosa* está justificada su restricción. El siguiente ejemplo nos muestra un caso particular de esta justificación (una explicación más general y detallada se tiene en los apartados 5.1.3 y 5.1.4): si A es el nombre de una oración verdadera, "A no es verdadera" será una oración falsa, y si A es el nombre de una oración no verdadera, "A no es verdadera" será una oración verdadera, luego A no podrá ser el nombre de la oración "A no es verdadera".

Si, en el razonamiento anterior cambiamos A por "esta oración", llegamos a la conclusión de que la expresión "esta oración", en su aparición dentro de la oración "esta oración no es verdadera" no puede tener como referencia dicha oración. Ahora bien, como tampoco tiene sentido que tenga como referencia otra oración distinta, debemos concluir que la expresión "esta oración" —en su aparición dentro de la oración "esta oración no es verdadera"— carece de referencia. En cambio, en casos de autorreferencia inocua como la que ocurre en la oración "esta oración tiene cinco palabras" la expresión "esta oración" se refiere a la oración que la contiene, la cual en este caso resulta ser verdadera.

El estudio de la autorreferencia y la justificación de cierta restricción a las referencias posibles de un nombre (o de una expresión referencial), cristaliza, en el apartado 5.1.5, en una de las principales tesis de este trabajo:

> Una oración puede constituir un contexto anulador de la referencia aparente o habitual de un nombre que forma parte de ella. Estos contextos son de naturaleza semántica.

Por tanto, esta propuesta, aunque tiene una dimensión contextual, no es del tipo de las de Barwise y Etchemendy (1987) o Simmons (1993) que consideran que la explicación de la paradoja del mentiroso radica en que el significado del predicado "verdadero" es sensible al contexto de emisión. Además, la contextualidad que introducimos produce una pérdida de composicionalidad mínima en la interpretación de las oraciones en el lenguaje formal trivaluado: basta tener en cuenta la proposición 4.8 (p. 124) para saber que el valor de verdad de la oración $R\,t_1, t_2, ..., t_n$ cuando alguno de los términos $t_1, t_2, ..., t_n$ carece de referencia, es $\mathfrak{i}$ (cualquiera que sea la interpretación del símbolo de relación R). Trasladado al lenguaje natural, diremos que, una vez justificado que la aparición de la expresión "esta oración" dentro de la oración "esta oración no es verdadera" carece de referencia, dicha oración no es verdadera ni falsa. Queda así resuelta la paradoja. La réplica de que al afirmar la oración se afirma una verdad no es correcta porque el sujeto de la oración está en un contexto que anula su referencia: afirmar "esta oración no es verdadera" es similar a afirmar "la oración de la página 1000 de este trabajo no es verdadera" pues en ambos casos el sujeto es una expresión referencial sin referencia (en un caso por motivos semánticos y en el otro por motivos empíricos). Tampoco perdemos capacidad expresiva —ni en lenguaje natural ni en lenguaje formal— en el sentido de que no podamos afirmar que la oración "esta oración no es verdadera" tiene o no tiene un determinado valor de verdad. Por ejemplo, en lenguaje natural, podemos construir la siguiente oración verdadera

> la oración "esta oración no es verdadera" no es verdadera

En lenguaje formal, si formalizamos la oración "esta oración no es verdadera" mediante $\neg T\tau$, la oración mediante la que podemos expresar que $\neg T\tau$ no es verdadera será:

$$\neg T^{\ulcorner}\neg T\tau^{\urcorner}$$

Tanto en el caso informal como en el formal, hemos hecho uso de una propiedad importante que garantiza que no perdamos capacidad expresiva:

> los términos de cita nunca están en un contexto referencialmente anulador

En el lenguaje formal, tampoco está en un contexto referencialmente anulador la aparición de ningún término que no vaya inmediatamente precedido de un predicado (o una función) de desentrecomillado.

Las conclusiones descritas hasta el momento constituyen una propuesta de solución a la paradoja del mentiroso que, en mi opinión, teniendo en cuenta la lista de requisitos del apartado 3 (p. 62), se puede considerar satisfactoria. Comprobemos que la propuesta verifica estos requisitos:

1. *Determinar qué principio o principios normalmente aceptados deben ser revisados para que las paradojas objeto de estudio desaparezcan. Los nuevos principios deben tener una justificación independiente de las paradojas o, al menos, no deben resultar demasiado artificiosos ni acarrear nuevas consecuencias inadmisibles.*

   Revisamos el principio de que la referencia de un nombre pueda ser arbitraria, el cual queda modificado por la introducción del principio de los contextos referencialmente anuladores: nombres que normalmente tienen referencia, dejan de tenerla en determinadas apariciones dentro de ciertas oraciones. Gran parte del trabajo está dedicado a la justificación de este principio. De un modo simplificado esta justificación se resume en: a) una caracterización de la paradoja que permite aislar las verdaderas raíces de la misma (autorreferencia y desentrecomillado) y descartar causas que son solo aparentes; b) un análisis que muestra que la modificación de los predicados de desentrecomillado (de los que el predicado *verdadero* es un ejemplo) no conduce a una solución satisfactoria; c) si no introducimos modificaciones al predicado *verdadero*, la oración "A no es verdadera" será una oración falsa cuando A sea el nombre de una oración verdadera y será una oración verdadera cuando A sea el nombre de una oración falsa: por tanto, queda justificado que A no puede ser el nombre de la oración "A no es verdadera". Creo que esta justificación no es nada artificiosa. Tampoco veo cómo el principio de que hay contextos que impiden que un nombre tenga su referencia habitual pueda acarrear consecuencias inadmisibles.

   En cambio, se puede alegar que la justificación del principio de los contextos referencialmente anuladores no es independiente de las paradojas, pero como ya se señaló en el apartado 3, no se trata de una condición imprescindible para una solución satisfactoria. A veces, precisamente lo interesante de una paradoja es que nos enseña que algunos principios que creíamos verdaderos

no lo son realmente y la única causa de ello podría ser la propia paradoja. Entonces, la solución a la misma solo puede tener una justificación dependiente de la paradoja. En el caso del principio de que la referencia de un nombre pueda ser arbitraria, lo aceptamos porque habitualmente los nombres se refieren a objetos extralingüísticos o a objetos lingüísticos de modo no problemático. Pero la paradoja del mentiroso (y afines) constituye una situación excepcional en el uso del lenguaje que, según nuestras conclusiones, muestra la necesidad de limitar aquel principio.

2. *Solucionar claramente la paradoja del mentiroso reforzada sin limitar la capacidad expresiva del lenguaje.*
   Como hemos visto, la paradoja del mentiroso reforzada queda resuelta en cuanto aceptamos que la expresión referencial que aparece en la oración está en un contexto anulador de su referencia. Siguiendo el ejemplo informal, si la expresión "esta oración" en su aparición en la oración "esta oración no es verdadera" carece de referencia, la oración no es verdadera ni falsa. También hemos comprobado que no queda limitada la capacidad expresiva del lenguaje (el uso de términos de cita lo garantiza).

3. *Los principios revisados no deben ser demasiado restrictivos, es decir, no deben evitar situaciones que no eran problemáticas.*
   Así es, puesto que no se anula la referencia de ningún nombre en las oraciones no problemáticas.

4. *Debe haber un criterio claro para determinar, suponiendo conocidos los hechos empíricos, si las oraciones autorreferenciales como las del mentiroso, del mentiroso reforzada, etc. son verdaderas, falsas o ni verdaderas ni falsas.*
   El valor de verdad de las oraciones queda determinado sin necesidad de introducir criterios nuevos. Solo merece señalarse que, en el lenguaje formal, el valor de verdad de la oración $R\,t_1, t_2, ..., t_n$ cuando alguno de los términos $t_1, t_2, ..., t_n$ carece de referencia, es $\mathfrak{i}$ (cualquiera que sea la interpretación del símbolo de relación R).

5. *Debe evitar introducir conceptos poco acertados, como el concepto de verdad* débil*, o conceptos que permitan generar nuevas paradojas.*
   No parece verosímil que la existencia de contextos referencialmente anuladores permita generar nuevas paradojas. Al menos, no vemos cómo. En

cualquier caso, una cualidad de esta propuesta es que produce una alteración mínima de principios semánticos comúnmente aceptados: no se introduce ninguna novedad en cuanto al predicado de verdad y, en cuanto a la referencia de los nombres, simplemente se resalta que, por sencillas razones semánticas, no siempre se puede elegir arbitrariamente.

Otra virtud de la propuesta, que se recomendaba, es que contiene una formulación precisa mediante lenguajes formales y, en todo caso, las conclusiones en lenguaje formal son trasladables al lenguaje natural. Finalmente la solución será mejor cuanto mayor variedad de paradojas resuelva. Por esta razón, en el apartado 5.1.6 y siguientes, hemos generalizado el planteamiento de los contextos referencialmente anuladores a otras oraciones y sistemas de oraciones emparentados con la paradoja del mentiroso.

6.3. Generalización de la propuesta

Hemos visto que, según las reglas de evaluación, si los nombres conservasen su referencia habitual en una oración autorreferencial de la forma $\rho(D\tau)$, como la del mentiroso o la del veraz, el valor de verdad, x, de dicha oración habría de cumplir una condición de la forma $x = (C_\rho \circ C_D)(x)$. En el caso de la oración del mentiroso su función $(C_\rho \circ C_D)$ no tiene punto fijo, pero en el caso de la oración del veraz tiene varios puntos fijos y también hay casos en que la función$(C_\rho \circ C_D)$ tiene un único punto fijo. Al intentar generalizar la solución dada a la oración del mentiroso a los casos en que la función $(C_\rho \circ C_D)$ tiene uno o más puntos fijos encontramos que, en principio, hay varias formas razonables de hacerlo. Tras una generalización razonable de los planteamientos G y MK vemos que, aunque ambos dan como resultado una misma evaluación de la oración del mentiroso, pueden diverger en la evaluación de oraciones donde $(C_\rho \circ C_D)$ tiene uno o más puntos fijos. Estamos ante un problema de generalizar la evaluación de oraciones a casos que admiten varias soluciones sin caer en contradicción. Y, como en la ampliación de conceptos matemáticos lo sensato es recurrir a criterios como la sencillez, la conservación del mayor número de propiedades posible, etc. para elegir una de las alternativas, pero puede haber varias soluciones aceptables. Por ejemplo, tan aceptable es suponer que el resto de la división entera de -8 entre 3 es -2 (y el cociente sería -2) que suponer que es 1 (y el cociente sería -3). Tenemos pues varias

formas de definir la división para enteros de modo compatible con la división de números naturales. Análogamente, tenemos varias formas de establecer criterios de evaluación de oraciones problemáticas de modo compatible con la evaluación de las oraciones no problemáticas. Por ejemplo, la oración "esta oración carece de contenido enunciativo (completo)" es falsa según el planteamiento G y no es verdadera ni falsa según el planteamiento MK.

En nuestro caso, a falta de un estudio más profundo de la opción MK, hay claros indicios de que la opción G (planteamiento puramente extensional) es más sencilla y menos problemática. Por eso es la que utilizamos en el resto del trabajo.

Muchas paradojas, e incluso diversas versiones de la paradoja del mentiroso, ocurren al intentar evaluar, no una sola oración, sino un conjunto de oraciones. En el apartado 5.2 generalizamos el planteamiento de los contextos referencialmente anuladores —con el que hemos encontrado una solución para la evaluación de la oración del mentiroso y un criterio razonable para evaluar otras oraciones como la del veraz— a sistemas finitos de oraciones no cuantificadas. El resultado es un método que permite evaluar coherentemente todas las oraciones de estos sistemas, gracias al uso de ecuaciones de valores de verdad (una por cada oración) y conceptos como el de referencia semántica aparente.

También hemos podido extender el planteamiento de los contextos referencialmente anuladores a oraciones cuantificadas y resolver paradojas basadas en ellas como la de Epiménides.

Más difícil es encontrar un método general para evaluar sistemas infinitos de oraciones, pero hemos mostrado cómo resolver uno de los más conocidos (el que constituye la paradoja de Yablo) lo que nos enseña aspectos importantes sobre la evaluación de tales sistemas.

El trabajo termina analizando la cuestión de si existe una explicación común a las paradojas como las del mentiroso y afines en que el problema es asignar coherentemente un valor de verdad a una oración o conjunto de oraciones y otras paradojas emparentadas como la paradoja heterológica o las clásicas paradojas de *Principia Mathematica*. Hemos comprobado que: a) también estas paradojas se pueden plantear mediante oraciones a las que aparentemente no se puede asignar ningún valor de verdad sin caer en contradicción; b) tanto en unas como en otras, la solución surge al justificar que un nombre que aparentemente tiene referencia realmente no la tiene; c) la combinación de autorreferencia y predicados de desentrecomillado, que explica la paradoja del mentiroso, también se produce en

la paradoja heterológica y en las paradojas de *Principia Mathematica*. Todo ello sugiere que podría haber una solución unificada. En cualquier caso, un enfoque similar al utilizado para estudiar la paradoja del mentiroso sería clarificador en el estudio de las paradojas de *Principia Mathematica* y similares.

Finalmente, dentro de cierta perspectiva unificadora, considero interesante resaltar que así como en la definición de un conjunto (u otras definiciones) no es aceptable cualquier propiedad, este trabajo pone de manifiesto que tampoco un nombre puede tener siempre cualquier referencia. Hace tiempo que se asimiló lo primero (era necesario para formalizar la teoría de conjuntos). Creo que la paradoja del mentiroso y afines nos enseña la necesidad de asimilar lo segundo.

6.4. Comparación con otras propuestas

Considero clarificador realizar un breve análisis de cómo la propuesta de los contextos referencialmente anuladores supera los principales problemas de otras propuestas estudiadas en el apartado 2 con las que se pueden encontrar importantes aspectos en común.

- Propuestas de Russell y Tarski. Aspecto destacado en común con nuestra propuesta: restringen la autorreferencia.
  Sin embargo, mientras que, como vimos, las propuestas de Russell y Tarski son demasiado restrictivas, nuestra propuesta distingue la autorreferencia inocua de la peligrosa y solo restringe la autorreferencia cuando es preciso para evaluar coherentemente las oraciones. Las propuestas de Russell y Tarski admiten importantes críticas por su condición de jerárquicas (jerarquía de los predicados *verdadero* y *falso*, jerarquía de lenguajes, etc.). Nuestra propuesta no es jerárquica.

- Propuesta de Kripke. Aspectos en común: hay oraciones que no son verdaderas ni falsas, el predicado *verdadero* no está jerarquizado, la lógica trivaluada fuerte de Kleene tiene un papel relevante, una misma oración puede ser paradójica o no en función de los hechos.
  Los principales fallos de la propuesta de Kripke son la limitación de la capacidad expresiva del lenguaje objeto que exige usar el metalenguaje para hacer asertos relevantes sobre las oraciones paradójicas y el hecho de que las paradojas reaparecen en el metalenguaje. Ninguno de estos problemas

aparece en nuestra propuesta.

En cuanto a la capacidad expresiva de nuestros lenguajes formales hemos justificado que no queda limitada en los aspectos relevantes a las paradojas objeto de estudio. Tomemos dos ejemplos relevantes de oraciones que no pueden expresarse en los lenguajes interpretados de Kripke. Una es la oración verdadera "la oración del mentiroso no es verdadera". Si llamamos λ a la oración del mentiroso en lenguaje objeto, la oración formal que podría corresponderle es $\neg T^{\ulcorner}\lambda^{\urcorner}$ (o $\sim T(^{\ulcorner}\lambda^{\urcorner})$ si nos queremos ajustar más a la notación usada por Kripke). Pero con el planteamiento de Kripke, la oración $\neg T^{\ulcorner}\lambda^{\urcorner}$ no es fundamentada y por tanto no es verdadera ni falsa, luego no puede interpretarse como una oración verdadera del lenguaje natural. En cambio, con nuestro planteamiento, la oración $\neg T^{\ulcorner}\lambda^{\urcorner}$ es verdadera, ya que su evaluación es:

$$\mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\lambda^{\urcorner}) = C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\lambda))) = C_{\neg}(C_T(\mathfrak{i})) = C_{\neg}(\mathfrak{f}) = \mathfrak{v} \tag{6.1}$$

Ahora no hay problema en considerar que $\neg T^{\ulcorner}\lambda^{\urcorner}$ es una buena formalización de "la oración del mentiroso no es verdadera". El otro ejemplo relevante es la oración del mentiroso reforzada ("esta oración no es verdadera") que, como vimos en el apartado 2.2, no es expresable en lenguaje objeto de Kripke. En cambio, sí es expresable en nuestros lenguajes formales una vez establecida la existencia de contextos referencialmente anuladores. Por ejemplo, usando la función de diagonalización, d, la oración del mentiroso reforzado es $\neg Td^{\ulcorner}\neg Tdx^{\urcorner}$, donde la referencia habitual del término $d^{\ulcorner}\neg Tdx^{\urcorner}$ es la oración $\neg Td^{\ulcorner}\neg Tdx^{\urcorner}$, pero, en el contexto de dicha oración el término carece de referencia, es decir:

$$\begin{array}{l} \mathcal{V}^h_{\mathfrak{M}}(d^{\ulcorner}\neg Tdx^{\urcorner}) = \neg Td^{\ulcorner}\neg Tdx^{\urcorner} \\ d^{\ulcorner}\neg Tdx^{\urcorner[\neg Td^{\ulcorner}\neg Tdx^{\urcorner}]} \notin dom(\mathcal{V}_{\mathfrak{M}}) \end{array} \tag{6.2}$$

Esto nos permite evaluar sin problemas la oración. Por una parte, las reglas de evaluación establecen que $\mathcal{V}_{\mathfrak{M}}(P\tau) = \mathfrak{i}$ cuando P es un predicado y τ un término que (en su aparición en la oración $P\tau$) carece de referencia. En nuestro caso resulta $\mathcal{V}_{\mathfrak{M}}(Td^{\ulcorner}\neg Tdx^{\urcorner}) = \mathfrak{i}$ y, finalmente,

$$\mathcal{V}_{\mathfrak{M}}(\neg Td^{\ulcorner}\neg Tdx^{\urcorner}) = C_{\neg}(\mathcal{V}_{\mathfrak{M}}(Td^{\ulcorner}\neg Tdx^{\urcorner})) = C_{\neg}(\mathfrak{i}) = \mathfrak{i} \tag{6.3}$$

Tampoco tenemos problemas en expresar en lenguaje formal que la oración del mentiroso reforzada no es verdadera. No se puede hacer con la propia oración del mentiroso reforzada (que no es verdadera ni falsa) pero sí, usando un término que la cita, como ocurre en:

$$\neg T^{\ulcorner}\neg Td^{\ulcorner}\neg Tdx^{\urcorner\urcorner} \tag{6.4}$$

Esta oración que afirma que la del mentiroso reforzada no es verdadera, es una oración verdadera, en consonancia con la afirmación en lenguaje natural. En efecto:

$$\mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\neg Td^{\ulcorner}\neg Tdx^{\urcorner\urcorner}) = C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\neg Td^{\ulcorner}\neg Tdx^{\urcorner}))) \tag{6.5}$$

y, teniendo en cuenta (6.3) y (6.5),

$$\mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner}\neg Td^{\ulcorner}\neg Tdx^{\urcorner\urcorner}) = C_{\neg}(C_T(\mathfrak{i})) = C_{\neg}(\mathfrak{f}) = \mathfrak{v} \tag{6.6}$$

En la propuesta de Kripke las paradojas reaparecen en el metalenguaje. En la nuestra no. De hecho, la solución de los contextos referencialmente anuladores es aplicable a las oraciones paradójicas expresadas en lenguaje natural. Por ejemplo, en el caso paradigmático de la oración "esta oración no es verdadera", la expresión "esta oración" pierde su referencia en el contexto de dicha oración. Estamos pues ante una oración cuyo sujeto carece de referencia y que, por ello, no es verdadera ni falsa (no tiene contenido enunciativo completo).

- Propuesta de van Fraassen. Como en la propuesta de Kripke, en la de van Fraassen hay oraciones que no son verdaderas ni falsas y la oración del mentiroso reforzada no es formalizable. Van Fraassen establece una jerarquía de *tipos-valor* y acaba reconociendo que la no veracidad de algunas oraciones no se puede expresar en el lenguaje formal. Por consiguiente, como en la propuesta de Kripke, hay una limitación de la capacidad expresiva del lenguaje formal que permite regenerar la paradoja del mentiroso reforzado en el metalenguaje. Y, como acabamos de ver, estos inconvenientes no aparecen en nuestra propuesta.

- Propuesta de Martin. Como en las dos anteriores hay oraciones que no son verdaderas ni falsas, entre las que se encuentra la del mentiroso. Como en la propuesta de Kripke, la oración verdadera "la oración del mentiroso no es verdadera" no es formalizable, y, como reconoce el propio Martin, la negación exclusiva no puede representarse en el lenguaje formal. Son muestras de una limitación de la capacidad expresiva del lenguaje formal que impide que la solución formal de Martin a las paradojas sea trasladable al lenguaje natural. Para mostrar que nuestra propuesta no tiene estos inconvenientes, nuevamente podemos remitirnos a los comentarios hechos en relación a la de Kripke.

- Propuesta de Gupta y Belnap y propuesta de Skyrms. Aunque el planteamiento inicial sea diferente, ambas propuestas tienen problemas similares a los señalados para las de Martin, van Fraassen y Kripke.

- Propuestas basadas en ecuaciones. En su versión más general, nuestra propuesta es, en buena medida, una propuesta basada en ecuaciones y puede considerarse una mejora de las propuestas de Hansson y Lan Wen (analizadas en el apartado 2.3.6). La mejora consiste, esencialmente, en que las ecuaciones de valores de verdad —o, en un planteamiento intensional, de contenidos enunciativos— asociadas a las oraciones son solo una herramienta para determinar qué apariciones de nombres pierden su referencia habitual. Una vez determinado, se pueden evaluar las oraciones sin problema. En cambio, en las propuestas de Hansson y Lan Wen no se contempla la posibilidad de que las expresiones referenciales puedan perder su referencia en el contexto de ciertas oraciones y no consiguen resolver satisfactoriamente la paradoja del mentiroso reforzada.

No compararemos nuestra propuesta con otras, como la de McGee o las *inconsistentes*, con cuyo planteamiento no comparte aspectos relevantes. Tampoco lo hace con las propuestas basadas en un concepto de verdad sensible al contexto de emisión de la oración. Los contextos de emisión de las oraciones poco o nada tienen que ver con los contextos referencialmente anuladores de nuestra propuesta. Además ella no establece un *nuevo* concepto de verdad, simplemente *refina* el funcionamiento de la referencia de las expresiones denotativas.

6.5. Futuros desarrollos

El problema que plantean las paradojas puede verse como el de establecer coherentemente el valor de verdad (o su falta de valor de verdad) de algunas oraciones enunciativas. Surge tanto en lenguaje natural como en lenguajes formalizados con suficiente capacidad expresiva. Con los conceptos semánticos usuales no hay modo coherente de asignar un valor de verdad a oraciones como la del mentiroso reforzada. La propuesta basada en los contextos referencialmente anuladores resuelve la paradoja del mentiroso y similares introduciendo una mínima y justificada alteración en el funcionamiento de la referencia de los nombres. Sin embargo, el problema va más allá de la evaluación de oraciones paradójicas, porque hay oraciones, no estrictamente paradójicas, como la del veraz, para las que las reglas de evaluación usuales no definen su valor de verdad.

Tenemos pues un problema de generalizar la evaluación de oraciones a casos que admiten varias soluciones sin caer en contradicción. Y, como en la ampliación de conceptos matemáticos, lo sensato es recurrir a criterios como la sencillez, la conservación del mayor número de propiedades posible, etc. para elegir una de las alternativas, pero puede haber varias soluciones aceptables. A este respecto merece señalarse la existencia de diversas axiomatizaciones de la teoría de conjuntos (Zermelo-Fraenkel con y sin axioma de elección, von Neumann-Bernays-Gödel, Kripke-Platek, etc.): todas ellas evitan paradojas como la de Russell pero no hay una considerada indiscutiblemente la mejor.

En este trabajo hemos considerado razonables dos enfoques, G y MK, en relación al valor de verdad de las oraciones enunciativas. El enfoque G, de carácter extensional (ver apartado 5.1.6) es el que hemos desarrollado por considerarlo menos problemático. Con él hemos podido establecer un método de evaluación de sistemas finitos de oraciones no cuantificadas. También nos ha servido para evaluar oraciones cuantificadas paradójicas y lo hemos podido generalizar para evaluar algunos sistemas infinitos de oraciones como el de la paradoja de Yablo. Sin embargo, no hemos establecido un método con la suficiente generalidad como para evaluar cualquier sistema de oraciones (finito o infinito de oraciones cuantificadas o no). Desde luego, sería interesante investigar la posibilidad de conseguirlo. La dificultad está en encontrar criterios adecuados para que la generalización del método de evaluación tenga buenas propiedades. Así que, como en la axiomatización

de la teoría de conjuntos, muy posiblemente se encontrarían varias alternativas razonables.

El enfoque MK, de carácter intensional, también merece ser estudiado pues, a pesar de sus dificultades, se basa en un análisis más fino de las oraciones al tener en cuenta su eventual contenido enunciativo y no solo su valor de verdad. Debido a ello, este enfoque resuelve la evaluación de la oración del veraz de un modo más natural: dado que no puede reducirse a una oración sin predicados de desentrecomillado (véase apartado 5.1.6.2), dicha oración carece de contenido enunciativo y por tanto no es verdadera ni falsa. En cambio, con el planteamiento extensional, para concluir que la oración del veraz no es verdadera ni falsa, hay que añadir la aplicación de un principio de simetría.

Incluso, como ya hemos señalado, la opción G podría considerarse un caso particular de la MK: aquel en que se considere que toda oración a la que se puede asignar coherentemente un único valor de verdad clásico (𝔱 o 𝔣) tiene un contenido enunciativo. Por consiguiente, merece investigarse si otras variantes del planteamiento MK pueden estar mejor justificadas que la opción G.

Aunque hemos encontrado importantes similitudes en la solución de la paradoja del mentiroso y otras paradojas como las de *Principia Mathematica* queda por ver si el estudio de todas esas paradojas mediante el uso de lenguajes formales nos llevaría a una solución común o por el contrario cada paradoja esconde problemas diferentes. Como hemos señalado más arriba, la cuestión no está zanjada a pesar de las ideas de Russell y de que Priest (1994) ha encontrado una estructura común a las más conocidas paradojas de autorreferencia. Estas paradojas también comparten el hecho de que en su formalización aparece la combinación de autorreferencia y desentrecomillado. Por tanto, considero interesante estudiar si una variante o una generalización de nuestra propuesta de solución de la paradoja del mentiroso y afines podría servir para establecer una solución unificada al conjunto más amplio de paradojas formado por las de *Principia Mathematica* y similares.

Varias paradojas, y en particular la del mentiroso, han jugado un papel importante en la demostración de algunos de los más conocidos teoremas sobre lenguajes y sistemas formales. Tal es el caso del teorema de Tarski sobre la indefinibilidad de la verdad en el lenguaje formal de la aritmética o el primer teorema de incompletitud de Gödel. Ambos se pueden obtener de un modo bastante directo a partir del teorema de punto fijo según el cual, en un sistema formal (sobre un lenguaje de predicados de primer orden) con suficiente riqueza expresiva dada una fórmula

$\psi(x)$ con una única variable libre, siempre hay una oración σ tal que:

$$\Phi \vdash (\sigma \leftrightarrow \psi(\ulcorner \sigma \urcorner)) \tag{6.7}$$

(Aquí Φ es el conjunto de axiomas del sistema formal y $\ulcorner \sigma \urcorner$ es un término cuya referencia es la oración σ). Si hubiera una fórmula, $T(x)$, que representara el predicado de verdad en Φ, debería verificar el esquema T de Tarski, es decir, para toda oración φ:

$$\Phi \vdash (\varphi \leftrightarrow T(\ulcorner \varphi \urcorner)) \tag{6.8}$$

Mas, un caso particular de (6.7) sería el que corresponde a la fórmula $\neg T(x)$, es decir, existirá una oración —llamémosla λ— tal que:

$$\Phi \vdash (\lambda \leftrightarrow \neg T(\ulcorner \lambda \urcorner)) \tag{6.9}$$

y un caso particular de (6.8) sería:

$$\Phi \vdash (\lambda \leftrightarrow T(\ulcorner \lambda \urcorner)) \tag{6.10}$$

Por último, de (6.9) y (6.10) se deduce:

$$\Phi \vdash (\neg T(\ulcorner \lambda \urcorner) \leftrightarrow T(\ulcorner \lambda \urcorner)) \tag{6.11}$$

lo cual solo es posible si Φ es inconsistente.

El proceso que acabamos de ver se puede resumir así: si en un sistema formal en que se cumple el teorema de punto fijo hubiese una fórmula que representase el predicado *verdadero*, existiría una oración formal, λ, equivalente a $\neg T(\ulcorner \lambda \urcorner)$, es decir, una oración del mentiroso reforzada. Y el poder formalizar esta oración paradójica en un sistema formal en el cual se verifique el esquema de Tarski va ligado a su inconsistencia.

El primer teorema de incompletitud de Gödel se obtiene —a partir del teorema de punto fijo— de forma similar al de Tarski, cambiando el predicado *verdadero* por el predicado demostrable. En vez de formalizar una oración, λ, que "dice" de sí misma que no es verdadera, formalizamos una oración, γ, que "dice" de sí misma que no es demostrable, es decir, si llamamos B al predicado *demostrable*,

$$\Phi \vdash (\gamma \leftrightarrow \neg B(\ulcorner \gamma \urcorner)) \tag{6.12}$$

Pero ahora la conclusión no es que el sistema sea inconsistente, sino que, si es consistente, la oración γ ha de ser verdadera y, por tanto, indemostrable dentro del sistema.

El teorema de Löb se basa en una demostración formal análoga a la que produce informalmente la paradoja de Löb: véase Boolos y Jeffrey (1989, pp. 186-7). También Boolos demuestra una versión del teorema de incompletitud de Gödel[2] explotando la paradoja de Berry, en lugar de la paradoja del mentiroso. E incluso el propio Gödel (1965, p. 9, nota al pie 14), señala que toda antinomia epistemológica puede usarse para una prueba similar de la indecidibilidad.

A diferencia de lo que ocurre en la lógica de predicados clásica, nuestra propuesta para superar la paradoja del mentiroso y afines, consiste en una lógica de predicados trivaluada y con contextos anuladores de la referencia de ciertos términos. Y, gracias a ello, se consigue que el lenguaje no pierda riqueza expresiva y, al mismo tiempo pueda contener el predicado verdadero (y predicados de desentrecomillado, en general). Así pues, en nuestros lenguajes interpretados trivaluados con capacidad de cita y contextos referencialmente anuladores no se cumple el correspondiente teorema de Tarski. Esto es solo un ejemplo de un amplio campo de estudio que abarcaría tanto las propiedades semánticas de este tipo de lenguajes interpretados como la posibilidad de establecer sistemas deductivos para ellos y, en su caso, estudiar sus propiedades. Y en este campo de estudio la formalización de las paradojas podrá seguir cumpliendo un papel relevante.

[2] "No hay algoritmo cuya salida contenga todas las sentencias verdaderas de la aritmética y ninguna falsa".

Apéndice A

Correspondencias, funciones y relaciones binarias: nomenclatura utilizada

Este apéndice no pretende revisar todos los conceptos básicos conjuntistas, los cuales, generalmente, se suponen conocidos. Su objetivo es destacar algunas propiedades por su interés para el presente trabajo y, sobretodo, determinar la nomenclatura a usar en aquellos casos en que, por ser comunes varias, resulta necesario para evitar interpretaciones erróneas o ambiguas.

A.1. Correspondencias

Una *correspondencia*, f, está definida por una terna de conjuntos no vacíos (A, B, G), donde G es un subconjunto del producto cartesiano $A \times B$.

A es el conjunto *inicial*, B es el conjunto *final* y G es el *grafo* de la correspondencia: $A = ini(f)$; $B = fin(f)$ y $G = graf(f)$.

Decir que una correspondencia, f, está definida entre los conjuntos A y B significa que A es el conjunto *inicial* y B es el conjunto *final* de la correspondencia. Esto a veces se expresa mediante:

$$f : A \to B$$

Cuando $(a, b) \in G$ se dice que la correspondencia asocia al elemento a, del conjunto inicial, el elemento b, del conjunto final.

El *dominio* de la correspondencia es el subconjunto de A formado por los elementos a los que la correspondencia asocia al menos un elemento de B:[1]

$$dom(f) = \{x/\ \exists y\ (x, y) \in G\}$$

El *rango* de la correspondencia es el subconjunto de B formado por los elementos a los que la correspondencia asocia al menos un elemento de A:

$$ran(f) = \{x/\ \exists y\ (y, x) \in G\}$$

A.2. Funciones

Una *función* es una correspondencia que asocia a cada elemento de su dominio uno y solo un elemento del conjunto final.

Si f es una función y $a \in dom(f)$, el elemento asociado a a por f se designa mediante $f(a)$.

Una función es *total* cuando su dominio coincide con su conjunto inicial. En caso contrario, la función es *parcial*.

Sean: f, una función definida entre A y B; g, una función definida entre B y C; G, el siguiente conjunto:

$$G = \{(x, y) \in A \times C / x \in dom(f) \wedge f(x) \in dom(g) \wedge y = g(f(x))\}$$

Si G no es vacío, se llama *función compuesta* de f y g a la función $(g \circ f)$ definida entre A y C cuyo grafo es G.[2] Así pues,

$$dom((g \circ f)) = \{x / x \in dom(f) \wedge f(x) \in dom(g)\}$$

y, cuando $x \in dom((g \circ f))$,

$$(g \circ f)(x) = g(f(x))$$

[1] No debe confundirse el dominio de una correspondencia con el dominio o universo de un modelo. Para este último, véase por ejemplo, la definición 4.4 en la p. 72.

[2] Puesto que el grafo de una correspondencia no puede ser el conjunto vacío, cuando G resulta vacío la función compuesta de f y g no está definida.

Cuando g es una función total la función compuesta $(g \circ f)$ estará definida y su dominio será:

$$dom((g \circ f)) = dom(f)$$

Si además f es una función total, $(g \circ f)$ será una función total.

Se dice que a es un *punto fijo* de la función h, cuando se cumple $h(a) = a$. Es claro que $a \in dom(h)$ y que $a \in ran(h)$.

Si f es una función definida por la terna (A, B, G), y $A' \subseteq A$ llamamos *restricción* de f *al conjunto inicial* A' a la función definida por la terna (A', B, G_i) con

$$G_i = \{(x, y)/\ x \in A' \wedge (x, y) \in G\}$$

Análogamente, si $B' \subseteq B$, la *restricción* de f *al conjunto final* B' es la función definida por la terna (A, B', G_f) con

$$G_f = \{(x, y)/\ (x, y) \in G \wedge y \in B'\}$$

Finalmente, la *restricción* de f *al conjunto inicial* A' *y al conjunto final* B' es la función definida por la terna (A', B', G_{if}) con

$$G_{if} = \{(x, y)/\ x \in A' \wedge (x, y) \in G \wedge y \in B'\}$$

Obsérvese que $G_i = G \cap (A' \times B)$, $G_f = G \cap (A \times B')$ y $G_{if} = G \cap (A' \times B')$.

Proposición A.1. *Dado un conjunto A, sean f y g dos funciones definidas entre A y A. Supongamos que las funciones compuestas $(g \circ f)$ y $(f \circ g)$ están definidas. Llamemos B al conjunto de puntos fijos de $(g \circ f)$ y C al conjunto de puntos fijos de $(f \circ g)$. Entonces el número de elementos de B es el mismo que el de C.*

Demostración. Primero comprobemos que si B no es vacío, C tampoco lo será. En efecto, tomemos $b \in B$ y tendremos $g(f(b)) = b$ —por ser b un punto fijo de $(g \circ f)$—, pero entonces $f(g(f(b))) = f(b)$, lo que indica que $f(b)$ es un punto fijo de $(f \circ g)$, es decir, $f(b) \in C$. Es obvio que, de modo perfectamente simétrico se puede probar que si C no es vacío, B tampoco lo será. Así pues, o B y C son ambos vacíos o ninguno de ellos lo es. En el primer caso el número de elementos de B y C es el mismo (cero). En el segundo caso, vamos a probar que hay una función biyectiva entre B y C lo que es suficiente para afirmar que ambos conjuntos tienen el mismo número de elementos.

Puesto que las funciones $(g \circ f)$ y $(f \circ g)$ están definidas entre A y A, sus puntos fijos son elementos de A. Así pues $B \subseteq A$ y $C \subseteq A$, lo que hace posible definir la función h como *la restricción de f al conjunto inicial B y al conjunto final C*. Entonces:

h es suprayectiva porque, dado cualquier $c \in C$, se tiene $g(c) \in B$ y $h(g(c)) = c$. En efecto:

$1^o/\ \ f(g(c)) = c$ —por ser c un punto fijo de $(f \circ g)$—, luego $g(f(g(c))) = g(c)$, lo que indica que $g(c)$ es un punto fijo de $(g \circ f)$, es decir, $g(c) \in B$.

$2^o/$ Como $f(g(c)) = c$ y $c \in C$, resulta $f(g(c)) \in C$, lo que, junto a $g(c) \in B$ y a la definición de h, nos permite afirmar que $h(g(c)) = f(g(c))$ y, por ende, $h(g(c)) = c$.

h es inyectiva, es decir, dados $b, b' \in B$, $h(b) = h(b') \Rightarrow b = b'$. En efecto, $h(b) = h(b') \Rightarrow f(b) = f(b') \Rightarrow g(f(b)) = g(f(b')) \Rightarrow b = b'$.

Como conclusión, h es biyectiva. ■

Por simetría, es obvio que la restricción de g al al conjunto inicial C y al conjunto final B será una función biyectiva. Además, no es difícil ver que esa función es la inversa de h.

A.3. Relaciones binarias

Una *relación binaria*, R, definida sobre un conjunto no vacío, A, es un par (A, G) donde G es un subconjunto del producto cartesiano $A \times A$. El conjunto G es el grafo de la relación binaria.

Para indicar que $(a, b) \in G$ cuando G es el grafo de una relación binaria R, es costumbre escribir aRb.

Dados dos conjuntos A y C, si $C \subseteq A$ y R es una relación binaria definida sobre A, diremos que C es un conjunto *R-cerrado* si y solo si

$$\forall (x, y) \in A \times A \quad ((x \in C) \wedge (xRy)) \rightarrow (y \in C)$$

Destacamos algunas propiedades que puede cumplir (o no) una *relación binaria*, R, definida sobre un conjunto no vacío, A:

- Reflexiva: $\forall x \in A \quad xRx$
- Irreflexiva: $\forall x \in A \quad \neg(xRx)$

- Simétrica: $\forall(x,y) \in A \times A \quad (xRy) \rightarrow (yRx)$
- Asimétrica: $\forall(x,y) \in A \times A \quad (xRy) \rightarrow \neg(yRx)$
- Antisimétrica: $\forall(x,y) \in A \times A \quad (xRy) \wedge (yRx) \rightarrow x = y$
- Transitiva: $\forall(x,y,z) \in A \times A \times A \quad (xRy) \wedge (yRz) \rightarrow (xRz)$

Una relación binaria es de *equivalencia* si cumple las propiedades reflexiva, simétrica y transitiva. Si $\sim$ es una relación de equivalencia definida sobre el conjunto A, y $a \in A$, la clase de equivalencia de a es el conjunto

$$[a]_{\sim} = \{x \in A / aRx\}$$

Es sabido y fácil de comprobar que cuando $(a,b) \in A \times A$: $1/$ si $a \sim b$ entonces $[a]_{\sim} = [b]_{\sim}$; $2/$ si $\neg(a \sim b)$ entonces $[a]_{\sim} \cap [b]_{\sim} = \emptyset$. Por tanto una relación de equivalencia, $\sim$, definida sobre un conjunto A establece una partición del conjunto A en clases de equivalencia. El conjunto cuyos elementos son las clases de equivalencia se denomina *conjunto cociente* de A respecto a la relación $\sim$ y se denota mediante $A/\sim$.

Una relación binaria es de *orden estricto* o *irreflexivo* si cumple las propiedades irreflexiva, y transitiva. Se deduce fácilmente que una relación de este tipo siempre cumple la propiedad asimétrica.

Una relación binaria es de *orden reflexivo* si cumple las propiedades reflexiva, antisimétrica y transitiva.

Cuando simplemente se dice que una relación binaria es de *orden* se entiende que es de orden reflexivo.

Se puede comprobar que dada una relación de orden irreflexivo, $\prec$, definida sobre un conjunto A, la relación $\preceq$ definida sobre A mediante:

$$\forall(x,y) \in A \times A \quad (x \preceq y) \leftrightarrow ((x \prec y) \vee x = y)$$

es una relación de orden reflexivo.

En sentido inverso, dada una relación de orden reflexivo, $\preceq$, definida sobre un conjunto A, podemos definir la relación $\prec$ sobre A mediante:

$$\forall(x,y) \in A \times A \quad (x \prec y) \leftrightarrow ((x \preceq y) \wedge \neg(x = y))$$

y comprobar que $\prec$ es una relación de orden estricto.

Por consiguiente, tiene sentido decir que tanto una relación de orden reflexivo como una de orden irreflexivo definidas sobre un conjunto A establecen una *ordenación* del conjunto.

Apéndice B

Ejemplo de aplicación del procedimiento de evaluación de sistemas finitos de oraciones no cuantificadas

Supongamos que deseamos conocer el valor veritativo de lo escrito en cada una de las páginas 1 a 6 de un cuaderno. Lo escrito en cada página es únicamente:

Página 1: "Lo escrito en la página 2 es verdadero y lo escrito en la página 1 es falso".

Página 2: "Lo escrito en la página 1 no es verdadero".

Página 3: "No es verdadero 'lo escrito en la página 1 es falso"'.

Página 4: "El río Amazonas es caudaloso" y lo escrito en la página 8 es falso"

Página 5: "Lo escrito en la página 6 es verdadero y lo escrito en la página 1 es verdadero".

Página 6: "O es verdadero lo escrito en la página 4 o es falso lo escrito en la página 5 o es falso lo escrito en la página 7".

Podemos formalizar este sistema de oraciones representando el término resultante de sustituir i por un número de página en "lo escrito en la página i" mediante el término formal resultante de sustituir i por un número de página en τ_i, representando los predicados *verdadero*, *falso* y *caudaloso* mediante los símbolos T, F y L, respectivamente y representando *El río Amazonas* mediante el símbolo de constante a. Llamando φ_i a la oración escrita en la página i, la referencia habi-

tual de cada término τ_i será la oración φ_i $(1 \leq i \leq 6)$ y el sistema de oraciones formalizado será:

$$\left.\begin{array}{ll} (\varphi_1) & (T\tau_2 \wedge F\tau_1) \\ (\varphi_2) & \neg T\tau_1 \\ (\varphi_3) & \neg T\ulcorner F\tau_1 \urcorner \\ (\varphi_4) & (La \wedge F\tau_8) \\ (\varphi_5) & (T\tau_6 \wedge T\tau_1) \\ (\varphi_6) & ((T\tau_4 \vee F\tau_5) \vee F\tau_7) \end{array}\right\} \tag{B.1}$$

Supongamos que usamos un modelo, $\mathfrak{M}$, en el que: 1/ la interpretación de las conectivas lógicas es la de la lógica trivaluada fuerte de Kleene; 2/ la interpretación de los símbolos de predicado de desentrecomillado T y F viene dada por C_T y C_F, respectivamente, siendo $C_T(\mathfrak{f}) = C_T(\mathfrak{i}) = \mathfrak{f}$, $C_T(\mathfrak{v}) = \mathfrak{v}$, $C_F(\mathfrak{v}) = C_F(\mathfrak{i}) = \mathfrak{f}$, $C_F(\mathfrak{f}) = \mathfrak{v}$; 3/ $\mathfrak{I}_{\mathfrak{M}}(a)$ es el río Amazonas e $\mathfrak{I}_{\mathfrak{M}}(L)$ es una función que asocia al río Amazonas el valor de verdad $\mathfrak{v}$.

Procedamos a aplicar a (B.1) el método de evaluación para sistemas finitos de oraciones no cuantificadas.

1. *Añadir las oraciones necesarias para formar un conjunto de oraciones S-cerrado, Φ.*

   Si hay una oración enunciativa escrita en la página 7 o en la página 8, las oraciones de (B.1) no forman un conjunto *S-cerrado* porque $\varphi_6 S \varphi_7$ o $\varphi_4 S \varphi_8$ y ni φ_7 ni φ_8 forman parte del conjunto de oraciones de (B.1). Supongamos que en la página 8 no hay escrito nada y en la página 7 hay escrito lo siguiente: "lo escrito en la página 6 es falso". Entonces, para conseguir un conjunto de oraciones S-cerrado, tenemos que añadir al sistema de oraciones la formalización de lo escrito en la página 7, es decir, la oración

$$(\varphi_7) \quad F\tau_6 \tag{B.2}$$

resultando el sistema de oraciones

$$\left.\begin{array}{ll} (\varphi_1) & (T\tau_2 \wedge F\tau_1) \\ (\varphi_2) & \neg T\tau_1 \\ (\varphi_3) & \neg T\ulcorner F\tau_1 \urcorner \\ (\varphi_4) & (La \wedge F\tau_8) \\ (\varphi_5) & (T\tau_6 \wedge T\tau_1) \\ (\varphi_6) & ((T\tau_4 \vee F\tau_5) \vee F\tau_7) \\ (\varphi_7) & F\tau_6 \end{array}\right\} \tag{B.3}$$

y, por supuesto,

$$\Phi = \{\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5, \varphi_6, \varphi_7\} \tag{B.4}$$

Además, τ_8 no tiene referencia habitual y la referencia habitual de τ_7 es la oración φ_7.

2. *Establecer el orden en que se recorrerán los elementos de* $\Phi/\!\sim_S$ *con el siguiente criterio: empezar por los vértices (clases de equivalencia establecidas por la relación* $\sim_S$ *en el conjunto de oraciones* Φ*) que no tienen sucesor en el grafo* $G(\prec_S, \Phi/\!\sim_S)$ *y continuar en sentido inverso al que establece la relación* $\prec_S$*.*
Para obtener $\Phi/\!\sim_S$, primero elaboramos el grafo $G(S, \Phi)$. Teniendo en cuenta la definición 5.2 (p. 176) y el conjunto de oraciones Φ, se obtiene:

$$\begin{aligned} G(S, \Phi) = \{&(\varphi_1, \varphi_1), (\varphi_1, \varphi_2), (\varphi_2, \varphi_1), (\varphi_3, \varphi_1), (\varphi_5, \varphi_1), \\ &(\varphi_5, \varphi_6), (\varphi_6, \varphi_4), (\varphi_6, \varphi_5), (\varphi_6, \varphi_7), (\varphi_7, \varphi_6)\} \end{aligned} \tag{B.5}$$

Es fácil comprobar que las clases de equivalencia establecidas por la relación $\sim_S$ (dos oraciones están relacionadas cuando son iguales o cuando en el grafo $G(S, \Phi)$ hay un camino de una a otra y viceversa) son $\{\varphi_1, \varphi_2\}$, $\{\varphi_3\}$, $\{\varphi_4\}$ y $\{\varphi_5, \varphi_6, \varphi_7\}$, es decir,

$$\Phi/\!\sim_S = \{\{\varphi_1, \varphi_2\}, \{\varphi_3\}, \{\varphi_4\}, \{\varphi_5, \varphi_6, \varphi_7\}\} \tag{B.6}$$

En este conjunto establecemos la relación $\prec_S$ (dos clases de equivalencia están relacionadas cuando no son iguales y hay un arco en $G(S, \Phi)$ entre un elemento de la primera clase de equivalencia y otro de la segunda). Así

podemos ver que $\{\varphi_3\} \prec_S \{\varphi_1, \varphi_2\}$, porque existe en $G(S, \Phi)$ el arco (φ_3, φ_1) entre un elemento de $\{\varphi_3\}$ y un elemento de $\{\varphi_1, \varphi_2\}$. Análogamente, es sencillo comprobar que el resto de pares relacionados es: $\{\varphi_5, \varphi_6, \varphi_7\} \prec_S \{\varphi_1, \varphi_2\}$ y $\{\varphi_5, \varphi_6, \varphi_7\} \prec_S \{\varphi_4\}$. En resumen, el grafo $G(\prec_S, \Phi/\sim_S)$ es el conjunto:

$$\{(\{\varphi_3\}, \{\varphi_1, \varphi_2\}), (\{\varphi_5, \varphi_6, \varphi_7\}, \{\varphi_1, \varphi_2\}), (\{\varphi_5, \varphi_6, \varphi_7\}, \{\varphi_4\})\} \tag{B.7}$$

por lo que debemos recorrer $\{\varphi_1, \varphi_2\}$ antes que $\{\varphi_3\}$ y antes que $\{\varphi_5, \varphi_6, \varphi_7\}$ y también $\{\varphi_4\}$ antes que $\{\varphi_5, \varphi_6, \varphi_7\}$.

3. *Recorrer las clases de equivalencia de* $\Phi/\sim_S$ *en el orden establecido para evaluar sus oraciones. Con cada clase de equivalencia,* $[u]_{\sim_S}$ *hay que hacer lo siguiente:*

   a) Hallar el sistema de ecuaciones de valores de verdad asociado:

      1) *Al valor de verdad de cada oración,* φ_i*,* $\varphi_i \in [u]_{\sim_S}$*, se asignará una incógnita,* x_i *y una ecuación de valores de verdad de la forma* $x_i = ...$

      2) *Sea* $A_i = \{\psi/\psi \in [u]_{\sim_S} \wedge \varphi_i S \psi\}$*. Para obtener el segundo miembro de la ecuación de valores de verdad:*

         a′ Téngase en cuenta que la aparición de un término en φ_i *pierde su referencia cuando se da en una oración (igual a* φ_i *o a una parte propia de* φ_i*) que ya ha sido evaluada (una oración del sistema o una parte propia de una oración ya evaluada) con el resultado de que la aparición del término perdía su referencia.*

         b′ Supóngase que ningún otro término de φ_i *pierde su referencia y aplíquese a la oración la composicionalidad y la definición de los predicados de desentrecomillado hasta que su valor de verdad quede expresado exclusivamente en función de los valores de verdad de oraciones ya evaluadas y de los valores de verdad de oraciones de* A_i*.*

   b) Aplicar el principio de simetría para resolver las referencias semánticas aparentes (es decir, decidir si son reales o no): si el sistema de ecuaciones de valores de verdad tiene una única solución, todas las referencias

semánticas aparentes a oraciones de $[u]_{\sim_S}$ *(contenidas en oraciones de* $[u]_{\sim_S}$*) son reales; en caso contrario, de esas referencias, solo son reales las realizadas mediante términos de cita.*

c) Evaluar las oraciones de $[u]_{\sim_S}$.

Realizamos el recorrido de las clases de equivalencia respetando el orden contrario a $\prec_S$:

- Paso 3a aplicado a $\{\varphi_1, \varphi_2\}$. Teniendo en cuenta la composicionalidad y que T y F son predicados de desentrecomillado, el sistema de ecuaciones asociado es

$$\left.\begin{array}{l} x_1 = C_\wedge(C_T(x_2), C_F(x_1)) \\ x_2 = C_\neg(C_T(x_1)) \end{array}\right\} \tag{B.8}$$

- Paso 3b aplicado a $\{\varphi_1, \varphi_2\}$. Puesto que el sistema anterior no tiene solución las referencias semánticas aparentes de oraciones de $\{\varphi_1, \varphi_2\}$ contenidas en oraciones de $\{\varphi_1, \varphi_2\}$ que no se realicen mediante términos de cita no son reales. Ello equivale a decir que las apariciones de τ_1 y τ_2 en las oraciones de $\{\varphi_1, \varphi_2\}$ carecen de referencia.
- Paso 3c aplicado a $\{\varphi_1, \varphi_2\}$.

$$\left.\begin{array}{l} \mathcal{V}_{\mathfrak{M}}(\varphi_1) = \mathcal{V}_{\mathfrak{M}}(T\tau_2 \wedge F\tau_1) = C_\wedge(\mathcal{V}_{\mathfrak{M}}(T\tau_2), \mathcal{V}_{\mathfrak{M}}(F\tau_1)) \\ \mathcal{V}_{\mathfrak{M}}(\varphi_2) = \mathcal{V}_{\mathfrak{M}}(\neg T\tau_1) = C_\neg(\mathcal{V}_{\mathfrak{M}}(T\tau_1)) \end{array}\right\} \tag{B.9}$$

donde, dado que τ_1 y τ_2 carecen de referencia en φ_1 y en φ_2—y, por tanto, en cualquiera de sus partes propias—,

$$\mathcal{V}_{\mathfrak{M}}(T\tau_2) = \mathfrak{i}; \quad \mathcal{V}_{\mathfrak{M}}(F\tau_1) = \mathfrak{i}; \quad \mathcal{V}_{\mathfrak{M}}(T\tau_1) = \mathfrak{i} \tag{B.10}$$

Sustituyendo (B.10) en (B.9) obtenemos:

$$\left.\begin{array}{l} \mathcal{V}_{\mathfrak{M}}(\varphi_1) = C_\wedge(\mathfrak{i}, \mathfrak{i}) = \mathfrak{i} \\ \mathcal{V}_{\mathfrak{M}}(\varphi_2) = C_\neg(\mathfrak{i}) = \mathfrak{i} \end{array}\right\} \tag{B.11}$$

- Paso 3a aplicado a $\{\varphi_3\}$. La aparición del término τ_1 en φ_3 se da en la oración $F\tau_1$ que es una parte propia de φ_3 y que ya ha sido evaluada (para evaluar φ_1) con el resultado de que τ_1 carece de referencia en la

oración $F\tau_1$. Puesto que

$$\mathcal{V}_{\mathfrak{M}}(\varphi_3) = \mathcal{V}_{\mathfrak{M}}(\neg T^{\ulcorner} F\tau_1 {}^{\urcorner}) = C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(F\tau_1))) \tag{B.12}$$

y $\mathcal{V}_{\mathfrak{M}}(F\tau_1) = \mathfrak{i}$, la ecuación de valores de verdad asociada a $\{\varphi_3\}$ es:

$$x_3 = C_{\neg}(C_T(\mathfrak{i})) \tag{B.13}$$

- Pasos 3b y 3c aplicados a $\{\varphi_3\}$. Evidentemente, la ecuación anterior tiene una única solución que es el valor de verdad de φ_3:

$$\mathcal{V}_{\mathfrak{M}}(\varphi_3) = C_{\neg}(C_T(\mathfrak{i})) = C_{\neg}(\mathfrak{f}) = \mathfrak{v} \tag{B.14}$$

- Pasos 3a, 3b y 3c aplicados a $\{\varphi_4\}$. En φ_4 no hay ninguna referencia semántica aparente (recuérdese que τ_8 es un término sin referencia habitual), por tanto la ecuación asociada nos permite calcular el valor de verdad de la oración independientemente de los valores veritativos de otras oraciones:

$$\mathcal{V}_{\mathfrak{M}}(\varphi_4) = x_4 = C_{\wedge}(\mathcal{V}_{\mathfrak{M}}(La), \mathcal{V}_{\mathfrak{M}}(F\tau_8)) \tag{B.15}$$

donde

$$\mathcal{V}_{\mathfrak{M}}(La) = \mathfrak{I}_{\mathfrak{M}}(L)(\mathcal{V}_{\mathfrak{M}}(a)) = \mathfrak{I}_{\mathfrak{M}}(L)(\mathfrak{I}_{\mathfrak{M}}(a)) = \mathfrak{v} \tag{B.16}$$

y

$$\mathcal{V}_{\mathfrak{M}}(F\tau_8) = \mathfrak{i} \tag{B.17}$$

por lo cual

$$\mathcal{V}_{\mathfrak{M}}(\varphi_4) = C_{\wedge}(\mathfrak{v}, \mathfrak{i}) = \mathfrak{i} \tag{B.18}$$

- Paso 3a aplicado a $\{\varphi_5, \varphi_6, \varphi_7\}$. La aparición del término τ_1 en φ_5 se da en la oración $T\tau_1$ que es una parte propia de φ_5 y que ya ha sido evaluada (para evaluar φ_2) con el resultado de que τ_1 carece de referencia en la oración $T\tau_1$. Para hallar el sistema de ecuaciones asociado hay que suponer que el resto de términos conserva su referencia habitual y tener en cuenta la composicionalidad y que T y F son predicados de desentrecomillado, hasta expresar el valor de verdad de cada oración

en función del valor de verdad de oraciones ya evaluadas y del valor de verdad de oraciones de $\{\varphi_5, \varphi_6, \varphi_7\}$. El resultado es:

$$\left.\begin{array}{l} x_5 = C_\wedge(C_T(x_6), \mathcal{V}_\mathfrak{M}(T\tau_1)) \\ x_6 = C_\vee(C_\vee(C_T(\mathcal{V}_\mathfrak{M}(\varphi_4)), C_F(x_5)), C_F(x_7)) \\ x_7 = C_F(x_6) \end{array}\right\} \tag{B.19}$$

Como sabemos que $\mathcal{V}_\mathfrak{M}(T\tau_1) = \mathfrak{i}$ y $\mathcal{V}_\mathfrak{M}(\varphi_4) = \mathfrak{i}$, tendremos

$$x_5 = C_\wedge(C_T(x_6), \mathfrak{i}) \tag{B.20}$$

y

$$\begin{array}{l} C_\vee(C_T(\mathcal{V}_\mathfrak{M}(\varphi_4)), C_F(x_5)) = \\ = C_\vee(C_T(\mathfrak{i}), C_F(x_5)) = C_\vee(\mathfrak{f}, C_F(x_5)) = C_F(x_5) \end{array} \tag{B.21}$$

Lo que nos permite dejar el sistema de ecuaciones en:

$$\left.\begin{array}{l} x_5 = C_\wedge(C_T(x_6), \mathfrak{i}) \\ x_6 = C_\vee(C_F(x_5), C_F(x_7)) \\ x_7 = C_F(x_6) \end{array}\right\} \tag{B.22}$$

cuya única solución es

$$x_5 = \mathfrak{i}, \quad x_6 = \mathfrak{v}, \quad x_7 = \mathfrak{f} \tag{B.23}$$

- Pasos 3b y 3c aplicados a $\{\varphi_5, \varphi_6, \varphi_7\}$. Puesto que el sistema de ecuaciones tiene una única solución las apariciones de los términos τ_5, τ_6 y τ_7 en las oraciones de $\{\varphi_5, \varphi_6, \varphi_7\}$ tienen su referencia habitual y el valor de verdad de las oraciones es, según (B.23):

$$\mathcal{V}_\mathfrak{M}(\varphi_5) = \mathfrak{i}, \quad \mathcal{V}_\mathfrak{M}(\varphi_6) = \mathfrak{v}, \quad \mathcal{V}_\mathfrak{M}(\varphi_7) = \mathfrak{f} \tag{B.24}$$

Una vez que ha terminado la evaluación formal del sistema de oraciones, podemos trasladar al lenguaje natural nuestras conclusiones.

Sabemos que al término formal τ_1 corresponde la expresión informal "lo escrito en la página 1", al término formal τ_2, la expresión informal "lo escrito en la página 2", etc. También sabemos que F y T formalizan los predicados *verdadero* y *falso*, respectivamente. Por tanto, a las oraciones formales $F\tau_1$, $T\tau_1$ y $T\tau_2$ corresponden

las oraciones "lo escrito en la página 1 es falso", "lo escrito en la página 1 es verdadero" y "lo escrito en la página 2 es verdadero", respectivamente. Ahora bien, una de las conclusiones de la evaluación formal es que las apariciones de los términos τ_1 y τ_2 en las oraciones $F\tau_1$, $T\tau_1$ y $T\tau_2$ carecen de referencia. El reflejo de esta conclusión en lenguaje natural es, pues, que las expresiones "lo escrito en la página 1" y "lo escrito en la página 2" carecen de referencia en su aparición en las oraciones "lo escrito en la página 1 es falso", "lo escrito en la página 1 es verdadero" y "lo escrito en la página 2 es verdadero". Por esa razón estas tres oraciones no son verdaderas ni falsas.

La evaluación de las oraciones ha sido:

$$\begin{array}{l} \mathcal{V}_{\mathfrak{M}}(\varphi_1) = \mathcal{V}_{\mathfrak{M}}(\varphi_2) = \mathcal{V}_{\mathfrak{M}}(\varphi_4) = \mathcal{V}_{\mathfrak{M}}(\varphi_5) = \mathfrak{i} \\ \mathcal{V}_{\mathfrak{M}}(\varphi_3) = \mathcal{V}_{\mathfrak{M}}(\varphi_6) = \mathfrak{v}; \quad \mathcal{V}_{\mathfrak{M}}(\varphi_7) = \mathfrak{f} \end{array} \tag{B.25}$$

Al fijarnos en $\mathcal{V}_{\mathfrak{M}}(\varphi_1) = \mathfrak{i}$, podría parecer que la oración "lo escrito en la página 1 es verdadero" es falsa, pero no es así, puesto que, como acabamos de ver, en esta oración, la expresión "lo escrito en la página 1" carece de referencia y la oración no es verdadera ni falsa.

Tanto en el lenguaje formal como en el informal, podemos construir oraciones verdaderas (y también falsas) acerca del valor de verdad de cada una de las siete oraciones del sistema. Por ejemplo, podemos expresar que φ_2 no es verdadera ni falsa mediante la oración formal

$$\neg T\ulcorner\neg T\tau_1\urcorner \wedge \neg F\ulcorner\neg T\tau_1\urcorner \tag{B.26}$$

dado que los términos de cita no pierden su referencia habitual; pero no mediante

$$\neg T\tau_2 \wedge \neg F\tau_2 \tag{B.27}$$

porque el término τ_2 no tiene referencia en la oración $T\tau_2$.

La traducción a lenguaje natural de (B.26) será: "no es verdadero 'no es verdadero lo escrito en la página 1' y no es falso 'no es verdadero lo escrito en la página 1"'.

Cabe preguntarse si se puede expresar que φ_2 no es verdadera ni falsa mediante la oración formal

$$\neg T\ulcorner\neg T\tau_1\urcorner \wedge \neg F\tau_2 \tag{B.28}$$

Puesto que esta oración no es la referencia habitual de ningún término de las oraciones de (B.3), aunque la añadiésemos al sistema, su evaluación será posterior a la evaluación de todas esas oraciones y, dado que $F\tau_2$ no ha sido evaluada como parte de (B.3), el término τ_2 no pierde la referencia en la oración $F\tau_2$. Así pues la respuesta a la pregunta es afirmativa y la comprobación de que (B.28) es verdadera es:

$$\begin{aligned}&\mathcal{V}_{\mathfrak{M}}(\neg T\ulcorner\neg T\tau_1\urcorner \wedge \neg F\tau_2) = C_{\wedge}(C_{\neg}(C_T(\mathcal{V}_{\mathfrak{M}}(\neg T\tau_1))), C_{\neg}(C_F(\mathcal{V}_{\mathfrak{M}}(\varphi_2)))) = \\ &= C_{\wedge}(C_{\neg}(C_T(\mathfrak{i})), C_{\neg}(C_F(\mathfrak{i}))) = C_{\wedge}(C_{\neg}(\mathfrak{f}), C_{\neg}(\mathfrak{f})) = C_{\wedge}(\mathfrak{v}, \mathfrak{v}) = \mathfrak{v}\end{aligned} \tag{B.29}$$

Bibliografía

Austin, J. L. (1950). Truth. *Proceedings of the Aristotelian Society*. Supp. vol. xxiv.

Barwise, J. y Etchemendy, J. (1987). *The Liar: An Essay on Truth and Circularity*. Oxford University Press, Nueva York.

Barwise, J. y Moss, L. (1996). *Vicious Circles: on the Mathematics of Non-Wellfounded Phenomena*. CSLI Publications, Stanford (CA).

Bealer, G. (1998). Propositions. *Mind*, 107(425): 1–32. Reimpreso en Jacquette (2002), pp. 120-139.

Bencivenga, E. (1980). *Una Logica dei Termini Singolari*. Boringhieri, Turín.

Beth, E. W. (1965a). *The Foundations of Mathematics*. North-Holland, Amsterdam, 2ª. ed.

Beth, E. W. (1965b). The Paradoxes of Logic and Set Theory and their Solution. En Beth (1965a), pp. 479–518.

Beth, E. W. (1975). Las paradojas de la lógica. *Cuadernos Teorema*, (4). Traducción al castellano de Beth (1965b).

Bhave, S. (1992). The Liar Paradox and Many-Valued Logic. *The Philosophical Quarterly*, 42(169): 465–479.

Blackburn, S. y Simmons, K. (editores) (1999). *Truth*. Oxford University Press, Oxford.

Boolos, G. S. (1998). *Logic, Logic, and Logic*. Harvard University Press, Cambridge (MA), Londres.

Boolos, G. S. y Jeffrey, R. C. (1989). *Computability and Logic*. Cambridge University Press, Cambridge, 3ª. ed.

Borga, M. y Palladino, M. (1997). *Oltre il mito della crisi: fondamenti e filosofia della matematica nel XX secolo*. Editrice la Scuola, Brescia.

Burge, T. (1979). Semantical Paradox. *Journal of Philosophy*, 76: 169–198. Reimpreso en Martin (1984), pp. 83-117.

Davis, M. (editor) (1965). *The Undecidable*. Raven Press, Nueva York.

Ebbinghaus, H.-D., Flum, J. y Thomas, W. (1996). *Mathematical Logic*. Undergraduate Texts in Mathematics. Springer-Verlag, Nueva York, Berlin, Heidelberg, 2ª. ed.

Falletta, N. (1983). *The Paradoxicon. A Collection of Contradictory Challenges, Problematical Puzzles and Impossible Illustrations*. Turnstone Press, Wellingborough, Northamptonshire. Las citas se refieren a la segunda impresión (1986) de la primera edición inglesa (1985).

Feferman, S. (1984). Toward Useful Type-Free Theories, I. *Journal of Symbolic Logic*, 49: 75–111. Reimpreso en Martin (1984), pp. 237-287.

Ferreirós, J. (1999). *Labyrinth of Thought. A History of Set Theory and Its Role in Modern Mathematics*. Birkhäuser, Boston (MA).

Field, H. (2002). Saving the Truth Schema from Paradox. *Journal of Philosophical Logic*, 31: 1–27.

Fitch, F. B. (1970). Comments and a Suggestion. En Martin (1970b), pp. 75–77.

Frege, G. (1984). Sobre sentido y referencia. En *Estudios sobre semántica*. Ariel, Barcelona. Traducción de C.U. Moulines.

Gamut, L. (1991). *Logic, Language and Meaning, vol. I: Introduction to Logic, vol. II: Intensional Logic and Logical Grammar*. University of Chicago Press, Chicago, Londres.

Garciadiego, A. R. (1992a). *Bertrand Russell and the origins of the set-theoretic 'paradoxes'*. Birkhäuser, Basel.

Garciadiego, A. R. (1992b). *Bertrand Russell y los orígenes de las "paradojas"de la teoría de conjuntos*. Alianza Universidad, Madrid. Versión castellana de Garciadiego (1992a).

García Suárez, A. (1997). *Modos de significar. Una introducción temática a la filosofía del lenguaje*. Tecnos, Madrid.

Gardner, M. (1986). *¡aja! Paradojas. Paradojas que hacen pensar*. Labor, Barcelona, 3ª. ed.

Givant, S. R. y McKenzie, R. N. (editores) (1986). *Alfred Tarski Collected Papers. Volume 4, 1958-1979*. Birkhäuser, Basel, Boston, Stuttgart.

Globe, L. (editor) (2001). *The Blackwell Guide to Philosophical Logic*. Malden (MA), Oxford.

Goldstein, L. (2000). A Unified Solution to Some Paradoxes. *Proceedings of the Aristotelian Society*, 100: 53–74.

Grattan-Guinness, I. (1998). Structural Similarity or Structuralism? Comments on Priest's Analysis of the Paradoxes of Self-Reference. *Mind*, 107(428): 823–834.

Groeneveld, W. (1994). Dynamic Semantics and Circular Propositions. *Journal of Philosophical Logic*, 23(3): 267–306.

Gupta, A. (1982). Truth and Paradox. *Journal of Philosophical Logic*, 11: 1–60. Reimpreso en Martin (1984), pp. 175-235.

Gupta, A. (2001). Truth. En Globe (2001), pp. 90–114.

Gupta, A. y Belnap, N. (1993). *The Revision Theory of Truth*. MIT Press, Cambridge (MA), Londres.

Gödel, K. (1965). On Formally Undecidable Propositions of the Principia Mathematica and Related Systems. I. En Davis (1965), pp. 4–38.

Hansson, B. (1978). Paradoxes in a Semantic Perspective. En Hintikka y cols. (1978), pp. 371–385.

Hintikka, J., Niiniluoto, I. y Saarinen, E. (editores) (1978). *Essays on Mathematical and Philosophical Logic*. Reidel Publishing Company, Dordrecht, Boston, Londres.

Hofstadter, D. R. (1998). *Gödel, Escher, Bach. Un Eterno y Grácil Bucle*. Tusquets, Barcelona, 8ª. ed.

Jacquette, D. (editor) (2001). *Philosophy of Mathematics. An Anthology*. Blackwell, Oxford.

Jacquette, D. (editor) (2002). *Philosophy of Logic. An Anthology*. Blackwell, Oxford.

Janssen, T. (1997). Compositionality. En van Benthem y ter Meulen (1997), cap. 7, pp. 417–473.

King, P. (1994). Reconciling Austinian and Russellian Accounts of the Liar Paradox. *Journal of Philosophical Logic*, 23(5): 451–494.

Kripke, S. (1975). Outline of a Theory of Truth. *Journal of Philosophy*, 72: 690–716. Reimpreso en Martin (1984), pp. 53-81.

Little, C. (2003). Parametrized Quotation and Self-Reference. *Electronic Notes in Theoretical Computer Science*, 74. `http://www.elsevier.nl/locate/entcs/volume74.html`.

Lorenzo, J. (1998). *La matemática: de sus fundamentos y crisis*. Tecnos, Madrid.

Mackie, J. (1973). *Truth, Probability and Paradox*. Oxford University Press, Oxford.

Malinowski, G. (1993). *Many-Valued Logics*. Oxford Logic Guides, vol. 25. Oxford University Press, Oxford.

Martin, R. L. (1970a). A Category Solution to the Liar. En Martin (1970b).

Martin, R. L. (editor) (1970b). *The Paradox of the Liar*. Ridgeview Publishing Company, Atascadero (CA).

Martin, R. L. (editor) (1984). *Recent Essays on Truth and the Liar Paradox*. Oxford University Press, Nueva York.

McDonald, B. (2000). On Meaningfulness and Truth. *Journal of Philosophical Logic*, 29: 433–482.

McGee, V. (1989). Applying Kripke's Theory of Truth. *Journal of Philosophy*, 86(10): 530–539.

McGee, V. (1990). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*. Hackett Publishing Company, Indianapolis.

McGee, V. (1998). Semantic Paradoxes and Theories of Truth. En Craig, E. (editor): *Routledge Encyclopedia of Philosophy*, vol. 8, pp. 642–648. Routledge, Londres y Nueva York.

Mellor, D. H. (editor) (1990). *Philosophical Papers. F. P. Ramsey*. Cambridge University Press, Cambridge.

Mills, A. (1995). Unsettled Problems with Vague Truth. *Canadian Journal of Philosophy*, 25(1): 103–117.

Parsons, C. (1974). The Liar Paradox. *Journal of Philosophical Logic*, 3: 381–412. Reimpreso en Martin (1984), pp. 9-45.

Parsons, C. (1984). Assertion, Denial, and the Liar Paradox. *Journal of Philosophical Logic*, 13: 137–152.

Picardi, E. (2001). *Teorías del significado*. Alianza, Madrid.

Priest, G. (1984). Logic of Paradox Revisited. *Journal of Philosophical Logic*, 13: 153–179.

Priest, G. (1994). The Structure of the Paradoxes of Self-Reference. *Mind*, 103: 25–34.

Priest, G. (1998). The Import of Inclosure: Some Comments on Grattan-Guinness. *Mind*, 107(428): 835–840.

Priest, G. (2000). Truth and Contradiction. *The Philosophical Quarterly*, 50(200): 305–319.

Priest, G. (2001). *An Introduction to Non-Classical Logic*. Cambridge University Press, Nueva York.

Priest, G. y Restall, G. (1998). Truth and Diagonalisation using Naïve Comprehension. `http://www.phil.mq.edu.au/staff/grestall/files/trdiagcomp.pdf`.

Putnam, H. (2000a). Paradox Revisited I: Truth. En Sher y Tieszen (2000), pp. 3–15.

Putnam, H. (2000b). Paradox Revisited II: Sets - A Case of All or None? En Sher y Tieszen (2000), pp. 16–26.

Quine, W. v. O. (1961). *From a Logical Point of View*. Harvard University Press, Cambridge (MA), 2ª. ed.

Quine, W. v. O. (1962). *Desde un punto de vista lógico*. Ariel, Barcelona. Traducción al castellano de Quine (1961) por Manuel Sacristán.

Quine, W. v. O. (1976). *TheWays of Paradox, and Other Essays*. Harvard University Press, Cambridge (MA), Londres.

Ramsey, F. P. (1925). The Foundations of Mathematics. En Mellor (1990), cap. 8.

Rescher, N. (2001). *Paradoxes. Their Roots, Range, and Resolution*. Open Court, Chicago y La Salle (Illinois).

Russell, B. (1903). *The Principles of Mathematics*. Cambridge University Press, Cambridge.

Salmon, N. y Soames, S. (1988). *Propositions and Attitudes*. Oxford Readings in Philosophy. Oxford University Press.

Sher, G. y Tieszen, R. (editores) (2000). *Between Logic and Intuition. Essays in Honor of Charles Parsons*. Cambridge University Press, Cambridge.

Simmons, K. (1993). *Universality and the Liar : An Essay on Truth and the Diagonal Argument*. Cambridge University Press, Cambridge.

Simmons, K. (1999). Deflationary Truth and the Liar. *Journal of Philosophical Logic*, 28: 455–488.

Skyrms, B. (1970). Notes on Quantification and Self-Reference. En Martin (1970b), pp. 67–74.

Smullyan, R. (1996). *Diagonalization and Self-Reference*. Clarendon Press, Oxford.

Sorensen (1998). Yablo's Paradox and Kindred Infinite Liars. *Mind*, 107: 137–155.

Tarski, A. (1933). Pojecie prawdy w jezykach nauk dedukcyjnych. (Sobre el concepto de verdad en los lenguajes de las ciencias deductivas). Varsovia.

Tarski, A. (1969). Truth and Proof. *Scientific American*, 220: 63–77. Reimpreso en Givant y McKenzie (1986), pp. 401-423.

Tarski, A. (1983a). The Concept of Truth in Formalized Languages. En Tarski (1983b), cap. VIII, pp. 152–278. Versión inglesa de Tarski (1933).

Tarski, A. (1983b). *Logic, Semantics, Methamatematics: Papers from 1923 to 1938*. Hackett Publishing Company, Indianapolis, 2ª. ed.

Tarski, A. (1999). La concepción semántica de la verdad y los fundamentos de la semántica. Traducción de Paloma García Abad. *A Parte Rei*, (6). `http://serbal.pntic.mec.es/%7ecmunoz11/tarski.pdf`.

van Benthem, J. (1988). *A Manual of Intensional Logic*. CSLI Publications, Stanford (CA).

van Benthem, J. y ter Meulen, A. (editores) (1997). *Handbook of Logic and Language*. Elsevier, Amsterdam.

van Fraassen, B. C. (1966). Singular Terms, Truth-Value Gaps, and Free Logic. *Journal of Philosophy*, 63(17): 481–495.

van Fraassen, B. C. (1968). Presupposition, Implication, and Self-Reference. *Journal of Philosophy*, 65(5): 136–152.

van Fraassen, B. C. (1970). Truth and Paradoxical Consequences. En Martin (1970b), pp. 13–23.

van Heijenoort, J. (1967a). *From Frege to Gödel. A source book in mathematical logic, 1879-1931*. Harvard University Press, Cambridge (MA).

van Heijenoort, J. (1967b). Logical Paradoxes. En Edwards, P. (editor): *The Encyclopedia of Philosophy*, vol. 5, pp. 45–51. Macmillan Publishing Co. et al., Nueva York, Londres.

Wen, L. (2001). Semantic Paradoxes as Equations. *Mathematical Intelligencer*, 23(1): 43–48.

Whitehead, A. N. y Russell, B. (1927). *Principia Mathematica*. Cambridge University Press, Cambridge, 2ª. ed.

Yablo, S. (1993). Paradox without self-reference. *Analysis*, 53(4): 251–252.