# Zirkularitätsverbot und zirkuläre Definitionen

Magisterarbeit
zur Erlangung des akademischen Grades
eines Magister Artium
der Universität Hamburg

vorgelegt von

Özgür Lütfü Özçep
aus Hamburg

Hamburg 2000

# Inhaltsverzeichnis

# 1 Einleitung

In dieser Arbeit möchte ich eine Definitionskonzeption vorstellen und besprechen, welche die Philosophen Anil Gupta und Nuel Belnap in ihrem 1993 erschienenen Buch „The Revision Theory of Truth“[1] entwickeln. Die in diesem Buch vorgestellte Revisionstheorie, in der eine Theorie der Definition und der Wahrheit inbegriffen sind, geht auf die Arbeiten [20], [21] von Hans Herzberger, [15] von Anil Gupta und [4] von Nuel Belnap zurück. Guptas und Belnaps Definitionskonzeption bricht radikal mit dem Zirkularitätsverbot der klassischen Definitionstheorie. Das Innovative an Guptas und Belnaps Vorgehen besteht darin, daß es im Gegensatz zu anderen Versuchen, sich argumentativ gegen das Zirkularitätsverbot zu wenden, nicht etwa danach strebt, möglichst viele zirkuläre Definitionen als harmlos zirkulär auszuweisen und somit gegen das Zirkularitätsverbot zu legitimieren, sondern darin, daß es sämtliche zirkuläre Definitionen für formal korrekt erklärt und diesen einen semantischen Sinn abzugewinnen versucht. Die Folge eines derart radikalen Bruchs mit der klassischen Definitionstheorie ist eine neue Konzeption der Bedeutung und die Postulierung einer bisher fast unbemerkt gebliebenen Gattung von Begriffen: Neben vage und/oder partiell definierte Begriffe treten nun zirkuläre Begriffe.

Da diese Konzeption recht neuartig ist und den Intuitionen - zumindest auf den ersten Blick - gehörig widersprechen dürfte, werde ich die Definitionstheorie, auf die ich unter dem Kürzel ‘(RTD)’ für ‘**R**evions**t**heorie der **D**efinition’ Bezug nehmen werde, recht ausführlich darstellen. Dieser Darstellung geht ein Abschnitt über das Zirkularitätsverbot und die klassische Definitionstheorie voran. Der Definitionsbegriff, welcher in diesem Abschnitt wie auch über die gesamte Arbeit hinweg zugrunde gelegt wird, ist sehr schwach. Der Grund für die Beschränkung liegt darin, daß Gupta und Belnap bei der Entwicklung ihres Definitionskonzeptes von dem in der Logik verwandten schwachen Begriff der Definitionen ausgehen, wonach Definitionen nicht mehr sein sollen als die Angabe von Anwendungsbedingungen für den definierten Ausdruck. Genauer muß man sagen, daß Gupta und Belnap für ihre Definitionen, die sie betrachten wollen, einen schwachen Bewertungsstandard für ausreichend halten: Sie fordern, daß ihre Definitionen nicht mit einem Standard höher als der der intensionalen Adäquatheit bewertet werden dürfe. Für ihre Zwecke, glauben Gupta und Belnap, ist die intensionale Adäquatheit ausreichend, für andere Zwecke – das gestehen sie ein – mag man stärkere Standards anlegen wollen.

Daher wäre es verkehrt, Guptas und Belnaps Konzeption nun mit einem Begriff der Definition zu beurteilen, der stärker ist als der, von dem sie ausgehen.

Für die klassische Definitionstheorie in der Logik sind zwei Kriterien von emi-

[1] Auf das Buch will ich mit dem Kürzel ‘RTT’ Bezug nehmen; bei Zitaten wird statt des Kürzels die diesem Buch in der Literaturliste zugewiesene Nummer in eckigen Klammern aufgeführt; auf die im Buch RTT entwickelte Revisionstheorie der Wahrheit werde ich unter dem Kürzel ‘(RTW)’ Bezug nehmen.

nenter Wichtigkeit: Das Kriterium der Eliminierbarkeit (EK) und das Kriterium der Nicht-Kreativität (NK). Während (NK) im ersten Abschnitt lediglich erwähnt wird, wird (EK) genauer untersucht: Insbesondere wird der Frage nachgegangen, was sich aus (EK) für das Zirkularitätsverbot folgern läßt.

Obwohl RTT auch eine neuartige Definitionstheorie vorstellt, ist es in der Hauptsache ein Buch zur Wahrheitstheorie. Ausgehend vom Lügnerparadoxon und in Auseinandersetzung mit verschiedenen Zugängen zum Lügnerparadoxon sowie den darauf aufbauenden Wahrheitstheorien entwickeln Gupta und Belnap ihre Revisionstheorie der Wahrheit (RTW). (RTW), so glauben sie, kann als die beste Lösung des Lügnerparadoxons angesehen werden – sofern man in einem angemessenen Sinne von der Lösung eines Paradoxons reden kann. Ihre Erklärung läuft darauf hinaus zu sagen, daß Wahrheit ein zirkulärer Begriff ist, für den das im Lügnerparadoxon aufgedeckte Verhalten im Rahmen der Theorie zirkulärer Definitionen erklärt werden kann. Wie das gesagt werden kann, ohne in einen Widerspruch zu geraten, werde ich im Abschnitt zur Wahrheit zu beschreiben versuchen.

Über andere, die Lügnerparadoxie eher weniger betreffende Aspekte der Wahrheit sagt (RTW) kaum etwas aus, insofern ist (RTW) keineswegs als eine vollständige Theorie der Wahrheit anzusehen.

Der Vergleich mit anderen Wahrheitstheorien geschieht im großen und ganzen dann auch nur bezüglich solcher Aspekte, die das Lügnerparadoxon (direkt oder indirekt) betreffen. Insbesondere der Auseinandersetzung mit der auf Kripkes berühmten Aufsatz „Outline of a Theory of Truth“ [25] zurückgehenden Fixpunkttheorie wird mit einem ganzen Kapitel viel Platz eingeräumt.

Integraler und der Theorie ihren Namen gebender Bestandteil der (RTW) ist die bereits erwähnte eigenständige Theorie zirkulärer Definitionen (RTD). In meiner Arbeit wird es mir hauptsächlich um (RTD) gehen und nur als Anwendungsbeispiel soll die Wahrheit als ein nach Gupta und Belnap durch (RTD) angemessen charakterisierbarer Begriff behandelt werden: Meine Arbeit ist folglich keine zur Theorie der Wahrheit. Allerdings ist dieses Vorhaben, die Wahrheit nur als *ein* Anwendungsbeispiel von (RTD) zu behandeln, durch die Tatsache eingeschränkt, daß die Anwendung von (RTD) im Grunde nur auf das Wahrheitsprädikat erfolgt und alle anderen Anwendungen entweder sich auf das Wahrheitsprädikat zurückführen lassen oder einfach nicht mehr als Spekulationen sind. Der Grund hierfür ist, daß die (RTD) erst durch die Betrachtung und Verallgemeinerung der Revisionsregel für das Wahrheitsprädikats entstanden ist. Das macht die Priorität des Wahrheitsbegriffes verständlich.

Die Grundintuitionen der (RTD) sind leicht darstellbar. (RTD) ist aber mehr als eine Ansammlung von intuitiven Thesen zur Zirkularität: Im Rahmen der mathematischen Logik werden Syntax und Semantik zirkulärer Definitionen geklärt, verschiedene logische Systeme betrachtet und diese mit Methoden tief aus der Trickkiste des (mathematischen) Logikers untersucht. Diesen technischen Ausbau von (RTD), der in Guptas und Belnaps Buch RTT in den Kapiteln 5 und

6 seinen Platz hat, werde ich anreißend in kurzen technischen Abschnitten behandeln, welche aufgrund der ihnen vorangestellten kurzen Zusammenfassungen getrost übersprungen werden können; die Grundideen der (RTD) werden auch aus der informellen Behandlung heraus verständlich werden – und die kurze Zusammenfassung sowie der Hinweis auf die strenge und korrekte Fundierung im Rahmen der mathematischen Logik sollte genügen. Nur soweit bestimmte technische Ergebnisse einen philosophisch relevanten Aspekt besitzen, werde ich sie in die nichttechnischen Kapitel einarbeiten.

Zum Ende der Einleitung noch einige Anmerkung zur Notation und Verwendung bestimmter Termini.
*Verwendung von Anführungsstrichen:* Einfache Anführungsstriche will ich verwenden, um auf die innerhalb der Anführungsstriche stehende Zeichenkette Bezug zu nehmen. Damit ist dann der folgende Satz wahr:

> ‘Frege’ ist eine aus fünf Buchstaben bestehende Zeichenkette.

Doppelte Anführungsstriche haben mehrere Funktionen:
1.) Sie werden für Zitate benutzt, wenn diese im Haupttext angeführt und nicht eingerückt sind.
2.) Sie dienen dazu, um den Leser darauf aufmerksam zu machen, daß der Ausdruck zwischen den Anführungsstrichen nicht in wörtlicher Bedeutung zu lesen ist.
3.) Mit Belnap und Gupta werden doppelte Anführungsstriche auch verwendet, um auf Begriffe Bezug zu nehmen. Damit läßt sich z.B. der folgende Satz formulieren:
Der Ausdruck ‘Wahrheit’ drückt den Begriff „Wahrheit“ aus. Neben diesen Anführungszeichen verwende ich gelegentlich auch die von Quine eingeführten „Corner Quotes“ (‘$\ulcorner$’ und ‘$\urcorner$’), um ein Mittel zur selektiven Anführung von Zeichen zur Hand zu haben.
*Definitionen:* Belnap und Gupta verwenden durchweg das Zeichen ‘$=_{Df}$’ für Definitionen. Ich werde ebenfalls dieses Zeichen[2] verwenden, wenn Definitionen einer quantorenlogischen Sprache als Beispiel angeführt werden. Manchmal wird es nötig sein, in der deutschen Sprache als der Metasprache, in der ich Belnaps und Guptas Entwicklung der Theorie zirkulärer Definitionen nachvollziehen werde, bestimmte Begriffe zu definieren. Für diese Definitionen verwende ich dann das Zeichen ‘$\Leftrightarrow_{Df}$’ oder einfach ‘: *gdw*’ bzw. ‘*gdw*’ als Abkürzung für ‘genau dann, wenn’.

Statt von Satzschemata werde ich mit Belnap und Gupta einfach von Sätzen reden; damit ist z.B. ‘$(\exists x)(Fx)$’ ein Satz einer quantorenlogischen Sprache mit einem einstelligen Prädikatbuchstaben ‘$F$’.

[2]Gegen die ausschließliche Verwendung des Zeichens ‘$=_{Df}$’ für Definitionen spricht, daß der größte Teil von Definitionen bestimmte Äquivalenzen ausdrücken und nicht Identitäten. Definitionen durch Identität, welche eh nur für Namen und Funktionssymbole in Frage kommen – beispielsweise die Definition der Tangensfunktion durch $\tan x = \frac{\sin x}{\cos x}$ – bilden die Ausnahme.

# 2 Definition und Zirkularitätsverbot

## 2.1 Der Begriff der Definition

Unter den Begriff „Definition“ fallen derart viele verschiedene Dinge, daß ein nach Vollständigkeit strebender Versuch, diese Dinge systematisch zu erfassen, als ein fast hoffnungsloses Projekt erscheinen muß. Der Grund dafür ist nicht nur, daß es etwa sehr viele verschiedene Verwendungsweisen des Ausdrucks in der Umgangssprache geben würde, die nur geringfügig mit dem philosophischen oder logischen Begriff der Definition zusammenhängen[3], sondern auch, daß es in der Philosophiegeschichte sehr viele Versuche gegeben hat, den Ausdruck ‘Definition’ zu erläutern und eine dieser Erläuterung angemessene Theorie der Definition zu entwickeln. Die folgende Auflistung aus Walter Dubislavs Buch [10] gibt einen groben Überblick darüber, was alles in der Philosophiegeschichte unter dem Titel ‘Definitionen’ fungierte.

> Die wichtigsten über die Definition aufgestellten Lehren sind die folgenden:
>
> A. Eine Definition besteht in der Hauptsache aus einer Wesensbestimmung (Sacherklärung).
>
> B. Eine Definition besteht in der Hauptsache aus einer Begriffsbestimmung (Begriffskonstruktion bzw. -zergliederung).
>
> C. Eine Definition besteht in der Hauptsache aus einer Feststellung (nicht Festsetzung) der Bedeutung, die ein Zeichen besitzt, bzw. der Verwendung, die es findet.
>
> D. Eine Definition besteht in der Hauptsache aus einer Festsetzung (nicht Feststellung) über die Bedeutung eines (neu einzuführenden) Zeichens bzw. über die Verwendung, die es finden soll. ([10], S.2)

Neben diesen Definitionstypen gibt es wesentlich andere Definitionssorten, bei denen nicht ein sprachlich geäußerter oder schriftlich fixierter Satz ausreichend ist, um erfolgreich einen Definitionsakt zu vollziehen: So ist etwa bei ostensiven Definitionen ein deiktischer Akt zur Definition nötig, bei dem auf einen Gegenstand F hingewiesen und gesagt wird: „Das da ist ein F“.

Ein klassischer Vertreter der Lehre unter (A) ist Aristoteles, der in der Topik I, 5 schreibt: „Definition ist eine Rede, die das Wesen anzeigt.“([3], S.5) Auf ihn geht das berühmte Zirkularitätsverbot zurück, welches bei fast allen Lehren unter (A)-(D) in irgendeiner Weise involviert ist. Die mit dem Zirkularitätsverbot verbundene Grundintuition darüber, was keine korrekte Definition sein kann, mag für alle Lehren aus (A)-(D) aus grundsätzlich ähnlichen Überlegungen entwachsen

[3] Ich denke an Verwendungsweisen wie die in den Äußerungen ‘Ich habe da eine Sache gesehen – ich kann es nicht definieren – ...’ oder ‘Das ist ein undefinierbares Gefühl’. Hier ist ‘definieren’ wohl zu lesen als ‘beschreiben’ oder ‘in Worte fassen’.

sein, muß es aber nicht. Wenn man sich also für das Zirkularitätsverbot interessiert und dessen Plausibilität zu ergründen sucht, dann sollte man sich dessen bewußt sein, daß man eigentlich für jede der verschiedenen Lehren der Definitionen (oder genauer: für jede der verschiedenen Definitionstypen) von neuem die Frage stellen muß: Was bedeutet es, daß eine Definition gemäß der Lehre XY zirkulär ist; und ist das Zirkularitätsverbot für Definitionen gemäß XY plausibel? Mit dieser Unterscheidung verbaut man sich nicht im voraus die mögliche Einsicht, daß für einen Definitionstyp das Zirkularitätsverbot durchaus gerechtfertigt ist, für einen anderen Definitionstyp aber nicht. Neben dieser Unterscheidung muß man aber noch eine weitere machen, die den Inhalt des Zirkularitätsverbots für einen ganz konkreten Definitionstyp betrifft: Es ist nicht von vornherein auszuschließen, daß man für Definitionen desselben Typs verschiedenen Formen von Zirkularität ausmachen kann und daß die eine Form der Zirkularität fatal, die andere hingegen nicht fatal ist; in diesem Falle wäre genau zu prüfen, welche Form von Zirkularität das Zirkularitätsverbot de facto ausschließt. Es wäre also zu prüfen, ob das Zirkularitätsverbot tatsächlich in dem Sinne adäquat ist, daß es genau die Formen von Zirkularität in einer Definition ausschließt, die in einem vernünftigen Sinne als fatal bezeichnet werden können. Die Bedingungen dafür, wann die Zirkularität in einer Definition fatal ist, müßten dann ebenfalls benannt werden.

Ich werde mich in diesem und dem folgenden Abschnitt auf die Theorie der Definition beschränken, die in der Logik für bestimmte künstliche Sprachen entwickelt wird, und das Zirkularitätsverbot im Rahmen dieser Theorie betrachten.

## 2.2 Wider die Zirkularität

Das Zirkularitätsverbot ist eine klassische Regel der Definitionstheorie, die – sehr grob gesagt – es verbietet, daß bei der Definition eines sprachlichen Ausdrucks F bzw. des Dinges (Wesen oder Begriff), das durch F ausgedrückt wird, mit Hilfe anderer Ausdrücke offen oder versteckt der Ausdruck F selbst verwendet wird.[4] Wenn wir unter 'Definition' für diesen Moment einfach einen Satz verstehen, der die Bedeutung eines Ausdrucks festlegt, dann ist der folgende Satz ein Beispiel für eine Definition, die das Zirkularitätsverbot verletzt:

$(\forall x)$(x ist ein Bruder $\Leftrightarrow_{Df}$ x ist ein Bruder $\vee$ x ist mit sich identisch)

Hier soll die Bedeutung des Ausdrucks 'ist ein Bruder' auf der Basis des Ausdrucks, der links vom '$\Leftrightarrow_{Df}$' steht, definiert werden. Nun kommt aber in diesem Ausdruck wieder 'ist Bruder' vor. Das scheint fatal zu sein. Wieso? Eine Antwort hierauf gibt vielleicht das folgende Zitat von Humberstone. Humberstone bespricht in seinem Aufsatz ([22]) zwei Formen von Zirkularität: die inferentielle

[4]Diese Formulierung bedarf natürlich einer Präzisierung. Für das angegebene Beispiel ist diese Formulierung jedoch hinreichend.

und analytische Zirkularität. Das Ergebnis seines Aufsatzes ist, daß die inferentielle Zirkularität nicht immer fatale Folgen zeitigt. Die analytische Zirkularität hingegen hält auch er für fatal. Zwar beziehen sich seine Aussagen hier auf Analysen und nicht Definitionen, aber die Einwände gegen analytische Zirkularität, die er vorstellt, bilden den Rahmen, in dem fast alle Begründungen gegen zirkuläre Definitionen oder Analysen zu finden sind:

> The general form of an account of the application of a concept K we are concerned with says that the concept K applies if and only if certain conditions $C_1,\ldots,C_n$ obtain. [...] Now, in the first place, we may be thinking of such an account as a putative *analysis* of the concept, in which case the analysis is circular if that concept is (overtly or covertly) employed in specifying the conditions $C_1,\ldots,C_n$. We will call this *analytical* circularity.
>
> [...] The mereological metaphor of constitution is of apiece with the idea that analytical circularity is a fatal flaw even in this partial and unidirectional form of would-be analysis: a whole cannot be composed of parts at least one of which has the original whole as a proper part. But, rather than *via* any such speculative involvement in the mereology of concepts, the more usual way of explaining why circularity is a flaw in a putative analysis adverts to the role analysis is supposed to play for thinkers: if a concept is being explained, the explanation should not be one intelligible only to those already possessing the concept [...] Analytical circularity is a fault, then, when and because it obstructs the transfer of understanding an account of the application conditions of a concept may be designed to effect: from understanding of the terms in which the account is couched to understanding the concept being analysed.

Humberstone spricht sich gegen das Teil-Ganzes-Modell der Analyse aus. Tatsächlich müßte man eine wohlfundierte, mereologisch einwandfreie Konzeption von Begriffen oder Bedeutungen haben, um mit einer Teil-Ganzes-Argumentation gegen zirkuläre Analysen/Definitionen angehen zu können. Was den zweiten Einwand gegen zirkuläre Analysen betrifft, so glaube ich, daß die erste Formulierung der Problematik eher auf Definitionen im Sinne von 'Angabe des Sinns eines Ausdrucks' zutrifft. Zirkularität einer Analyse ist doch nicht deswegen fatal, weil wir nicht mit seiner Hilfe jemanden dazu bringen können, diesen Begriff zu erwerben. Müssen also tatsächlich Analysen auch für solche Leute verständlich sein, die nicht über den Begriff, der analysiert werden soll, verfügen? Genauer: Müssen solche Leute, vorausgesetzt sie besitzen bestimmte kognitive Fähigkeiten, mit Hilfe einer Analyse in den Stand versetzt werden können, den analysierten Begriff zu erwerben? Zweck einer Analyse ist doch, den analysierten Begriff zu erhellen, indem man ihn in Beziehung zu anderen Begriffen setzt. Jemand, der über den analysierten Begriff bereits verfügt, kann von einer Analyse profitieren. Der Punkt ist doch, daß eine (echt) zirkuläre Analyse auch für den, der über den analysierten Begriff verfügt, nicht viel Erhellendes über die Anwendungsbedingungen des analysierten Begriffs C wird sagen können. Dieser jemand wird nicht

auf den Ausdruck A (das Analysans) schauen können, mit dem analysiert wird, und sagen können, er wisse jetzt, daß etwas genau dann ein C ist, wenn es ein A ist. Er wird nichts erfahren über die Beziehung des Begriffs C und der Begriffe, die ausgedrückt werden durch Terme, welche in A vorkommen.

Ob dieses Argument in jedem Falle fruchtet, wird im Abschnitt 'Begründung des Zirkularitätsverbots' untersucht.

Mindestens seit Aristoteles von Definitionen nicht mehr wegzudenken findet sich das Zirkularitätsverbot nun in nacharistotelischen Logikbüchern (vor 1879) in einem Kanon von Definitionsregeln wieder, die bei Aristoteles nur als beiläufige Erwägungen, verstreut in Abschnitten der Analytica Posteriora und der Topik, vorkommen. Den Anfang der um diesen Regelkanon aufbauenden traditionellen Definitionslehre markiert die Schule von Port Royal. Auch wenn sich letztlich diese Regeln nicht als adäquat in dem Sinne erwiesen, daß sich trotz Befolgung aller Regeln „schlechte" – etwa zu Inkonsistenzen – führende Definitionen aufstellen ließen, so hat sich doch das Zirkularitätsverbot über die Revision dieser Regeln in der an die mathematische Logik orientierten Standardtheorie der Definition hindurch gerettet. Üblicherweise werden als die traditionellen an Aristoteles orientierten Regeln der Definition die vier folgenden aufgeführt:[5]

1. Eine Definition muß das Wesen dessen wiedergeben, das definiert werden soll.
2. Eine Definition darf nicht zirkulär sein.
3. Eine Definition darf nicht negativ (formuliert) sein, wenn es positiv (formuliert) sein kann.
4. Eine Definition darf nicht in figurativer oder obskurer Sprache ausgedückt werden.

Suppes in [41] (S.151-152) versucht zu motivieren, daß diese vier Regeln nicht hinreichend sind, um den formalen Begriff einer Definition im Rahmen von Theorien wie der der Arithmetik zu erhellen. Die folgende Definition des Pseudooperators * erfülle alle vier Regeln, doch läßt sich aus ihr bei Hinzunahme zur Arithmetik ein Widerspruch ableiten.

$$(\forall x)(\forall y)(\forall z)(x * y = z \quad \Leftrightarrow_{df} \quad x < z \quad \& \quad y < z)$$

Nach der Definition gilt nun sowohl $1{*}2 = 3$, da ja $1 < 3$ und $2 < 3$, als auch $1{*}2 = 4$ , da $1 < 4$ und $2 < 4$; damit würde aber $3 = 4$ folgen, was im Widerspruch zur Standard-Arithmetik steht, in der sich $3 \neq 4$ beweisen läßt. Etwas problematisch an Suppes Beispiel ist der Punkt, die obige Definition würde tatsächlich das Wesen der Operation * wiedergeben. Die erste Regel scheint doch nur für den Fall formuliert zu sein, daß es bereits etwas gibt, dessen Wesen definiert werden soll. Ist das bei der Definition der Operation * auch der Fall? Die Definition von '*' ist eine

[5]Siehe z.B. [41], S. 151.

rein stipulative, festsetzende Definition; hier liegt kein Gegenstand vor, höchstens vielleicht das Zeichen '*'. Aber was sollte es heißen, das Wesen dieses Zeichens zu definieren – und haben wir das überhaupt beabsichtigt? Ich sehe nicht, was Suppes als Gründe dafür anführen kann, daß die vier traditionellen Regeln zu verwerfen sind. Der einfache Punkt muß hier doch sein, daß diese Regeln nichts über den Typ von Definitionen sagen, bei denen es darum geht, die Bedeutung eines Ausdrucks festzusetzen. Und für diesen Definitionstyp interessiert sich der Mathematiker und Logiker. Daß sich bei einer schludrigen Übertragung dieser Regeln auf stipulative Definitionen Probleme wie das oben dargestellte ergeben, ist nur verständlich.

# 3 Eliminierbarkeit und Kreativität in (STDL)

In diesem Abschnitt möchte ich jene Theorie der Definition besprechen, die in den meisten Logikbüchern, welche Definitionen überhaupt einen Kapitel einräumen, als die klassische, nach-aristotelische Theorie der Definition dargestellt wird. Konkret denke ich an die Darstellung im Buch „Introduction to Logic" von Patrick Suppes ([41]) und die u.a. hierauf verweisende Darstellung im Artikel „On rigorous Definitions" von Nuel Belnap ([5]). Ich werde im folgenden auf diese Theorie der Definition unter dem Kürzel '(STDL)' (für '**S**tandard-**T**heorie der **D**efinition in der **L**ogik') Bezug nehmen. Diese Theorie ist eine den „Bedürfnissen" und der Praxis des Logikers entwachsene und ihr angepaßte Theorie der Definition, von der sich der Philosoph aufgrund der Verwendung des Wortes 'Definition' in der Philosophiegeschichte nicht allzuviel versprechen darf. (STDL) bildet nichtsdestotrotz eine Verständigungsbasis, in der gewisse Minimalintuitionen zum Begriff der Definition zugrunde gelegt werden. Und diese wird vermutlich auch der meist „höhere" Ansprüche an Definitionen stellende Philosoph teilen.

Unter „Definition" soll in diesem Zusammenhang eine satzförmige sprachliche Entität verstanden werden, die die Bedeutung eines Ausdrucks erläutert. Diese Erklärung läßt einen gewissen Spielraum, der durch den ambigen Ausdruck 'Bedeutung' bedingt ist.

Man kann sich die (STDL) aus zwei Teilen bestehend denken. Der eine Teil hat zum Thema Kriterien für (gute) Definitionen, der andere Teil hat zum Thema Regeln für die Aufstellung von Definitionen. Dabei ist der zweite Teil dem ersten in folgendem Sinne untergeordnet: Der Anspruch ist, mit den im zweiten Teil gegebenen Regeln die Konstruktion solcher und nur solcher Definitionen zu gestatten, welche den im ersten Teil angegebenen Kriterien genügen. In dieser Zweigliedrigkeit ähnelt das Projekt der (STDL) dem der klassischen Logik;[6] so stößt man hier zum einen auf die Explikation des Begriffs der (semantischen) Folgerung, zum anderen auf den Begriff der Ableitbarkeit. Und auch darin ähneln

[6]Darauf weist Belnap hin ([5], S.118), der diesen sehr schönen Vergleich vermutlich bei Tarski aufgelesen hat.

sich die Projekte der (STDL) und der klassischen Logik, daß in dieser ebenfalls der eine Teil dem anderen untergeordnet zu sein scheint: Die in einem Kalkül ableitbaren Argumentschemata sollen gerade die und nur die sein, die gemäß des im ersten Teil explizierten Begriffs der semantischen Folgerung korrekt sind.

Wiewohl die Regeln zur Aufstellung von Definitionen nur solche den Kriterien genügende Definition aufzustellen gestatten sollen, so ist mit ihnen doch ein von den Kriterien unabhängiges Projekt verbunden. Auf der einen Seite steht das unter den Regeln vorkommende Zirkularitätsverbot, auf der anderen Seite aber der durch das Eliminierbarkeitskriterium implizierte Ausschluß einer zirkulären Definition. Da ist die Frage naheliegend, ob es nicht eine für das Zirkularitätsverbot im wesentlichen andere Motivation gibt als für das Eliminierbarkeitskriterium. Wenn das nicht der Fall ist, so müßte es eine gute Begründung für das Eliminierbarkeitskriterium geben, die auch das Zirkularitätsverbot fundieren würde. Die mögliche Beobachtung, daß aus der Eliminierbarkeitsforderung zwar das Zirkularitätsverbot folgt, aber nicht aus dem Zirkularitätsverbot die Eliminierbarkeitsforderung, läßt noch keine Schlüsse über die Frage der Unabhängigkeit oder Abhängigkeit der Rechtfertigung des einen durch den anderen zu. Man kann eine Version des Zirkularitätsverbots über das Eliminiebarkeiskriterium rechtfertigen, da jenes aus diesem folgt – man muß es aber nicht.

Für die folgende Diskussion will ich die übliche Redeweise von 'Definiendum' und 'Definiens' zugrunde legen: Unter dem *Definiendum* einer Definition ist derjenige in der Definition vorkommende Ausdruck zu verstehen, dessen Bedeutung durch die Definition angegeben wird. Derjenige Ausdruck in einer Definition, der weder das Definiendum noch das Definitionszeichen darstellt, ist das *Definiens*. Das Definiens ist also der Ausdruck, mit dessen Hilfe das Definiendum seine Bedeutung erhält – wie das vonstattengeht, ist noch nicht gesagt. Diese Redeweise erhält erst dann die nötige Präzision – und wurde vermutlich auch lediglich für diesen im folgenden genannten Fall konzipiert – wenn Definitionen in Form eines (mit Allquantoren abgeschlossenen) Bikonditionals oder in Form einer Identitätsaussage dargestellt werden (mit Hilfe von dem Bikonditional „ähnlich" fungierenden Zeichen '$\Leftrightarrow_{Df}$' bzw. bei der Identität '$=_{Df}$'): Dann läßt sich vereinbaren, daß der Ausdruck links (abzüglich der eventuell vorhandenen Quantoren) vom Bikonditional bzw. vom Gleichheitszeichen das Definiendum ist und der Ausdruck rechts davon das Definiens. Bei Sätzen S, die ein Zeichen O enthalten und dessen Bedeutung (im Rahmen einer vorgegebenen Theorie) festlegen, ohne doch von der Gestalt eines Bikonditionals oder einer Identitätsaussage zu sein, ist zwar die Rede vom 'Definiendum' noch möglich, aber die Rede vom 'Definiens', seinem Pendant, erweist sich als nicht besonders glücklich: Was sollte das Definiens sein? Der Rest von dem Satz S, der übrigbleibt, wenn man das Definiendum O rausnimmt? Dann wären – was nicht besonders problematisch erscheinen mag – das Definiens und das Definiendum sprachliche Entitäten aus eventuell *unterschiedlichen* logisch-grammatischen Kategorien. Ist O z.B. ein Funktionssymbol, dann wäre das Definiendum ein Satz-bildender Funktionsoperator - oder etwas, was als

Input offene Terme hat.

## 3.1 Eliminierbarkeit und Nicht-Kreativität - klassisch

Das Kriterium der Nicht-Kreativität[7] verlangt grob gesagt, daß mit der aufgestellten Definition nur solche Aussagen ableitbar (folgerbar) sein dürfen, die auch ohne die Definition ableitbar (folgerbar) sind. D.h. lax gesprochen, daß mit der Einführung der Definition nicht neue Aussagen eingekauft werden dürfen.

Das Kriterium der Eliminierbarkeit für Definitionen fordert grob gesagt, daß jedes Vorkommnis eines neu definierten Ausdrucks in allen Kontexten (eines bestimmten Typs) durch bereits zur Verfügung stehende Ausdrücke[8] ersetzt werden kann, so daß zwischen dem (sprachlichen) Gebilde, in dem sich das Vorkommnis des Definiendums befindet, und demjenigen (sprachlichen) Gebilde, das die ersetzenden Ausdrücke enthält, eine gewisse Gleichwertigkeit[9] besteht.

Beide Kriterien sollen nach der (STDL) jeweils notwendige Bedingungen und zusammen (i.e. konjunktiv verknüpft) eine hinreichende Bedingung für eine gute Definition darstellen. Man kann einsehen, daß mit diesen Kriterien gewisse Sätze als nicht-korrekte Definitionen ausgeschlossen werden können, welche ansonsten nicht gewünschte Konsequenzen hätten. Aber wie werden sie motiviert? Da für das Zirkularitätsverbot hauptsächlich das Eliminierbarkeitskriterium wichtig ist, werde ich ausschließlich dieses behandeln. Lediglich bei der exakten Definition des Kriteriums der Nicht-Kreativität mache ich eine Ausnahme.

### 3.1.1 Eliminierbarkeit: Motivation

Eine ursprüngliche Fassung der Eliminierbarkeit findet man bereits bei Blaise Pascal. Dubislav bespricht Pascals Definitionstheorie im Kapitel über Definitionstheorien, denen zufolge eine Definition eine Festsetzung über die Bedeutung ist, die man einem neu einzuführenden Terminus zu geben beabsichtigt, bzw. über die Verwendung, die er finden soll.[10] Dubislav referiert, daß Pascal in seiner

[7] Für dieses Kriterium ist im Englischen auch der Terminus 'criterion of conservativeness' gebräuchlich.

[8] Die Wendung „bereits zur Verfügung stehende Ausdrücke" ist natürlich noch auszubuchstabieren. Man denke zunächst an Ausdrücke, die als einer Sprachgemeinschaft bekannt und als von ihr verstanden vorausgesetzt werden dürfen.

[9] Man denke konkret an intensionale Äquivalenz im üblichen Sinne, so z.B für Prädikate formuliert: Zwei Prädikate F und G sind genau dann intensional äquivalent, wenn folgendes erfüllt ist: Es gilt notwendigerweise, daß $(\forall x)(Fx \leftrightarrow Gx)$.

[10] Man beachte Dubislavs umsichtige Unterscheidung zwischen Bedeutung und Verwendung eines Terminus. Ich vermute, daß sein „bzw." nicht als explikatives zu verstehen ist; es sollen tatsächlich zwei verschiedene Dinge gemeint sein. Aller Wahrscheinlichkeit fand er diese Unterscheidung aber derart offensichtlich, daß er es nicht für nötig befunden hat - zumindest zu Anfang seines Werkes, wo er die Kategorisierung der Definitionen vornimmt - diese zu motivieren.

Abhandlung „De L'Art der Persuader“ neben zwei anderen diese Regel für die Aufstellung von Definitionen gegeben habe:

> Er [sc.: Pascal] fordert weiterhin, daß man in den benutzten Ausdrücken, wenn definierte Zeichen in ihnen enthalten seien, diese in Gedanken durch die ihnen per definitionem zugeordneten ersetze, um durch Fortsetzung dieses Verfahrens schließlich feststellen zu können, ob die ursprünglichen Ausdrücke auch restlos durch allerdings ungefüge Kombinationen durch sich selbst verständlicher Zeichen ersetzbar seien. ([10], S.22)

Diese „Regel“[11], wie Dubislav es nennt, enthält im Kern einen Eliminierbarkeitsgedanken: In den „benutzten Ausdrückenëiner Sprache[12], die definierte Ausdrücke enthalten, müsse überprüft werden, ob diese definierten Ausdrücke durch andere, ausgezeichnete Ausdrücke - ausgezeichnet dadurch, daß sie aus sich selbst heraus verständlich seien - ersetzbar sind. Leider wird nicht gesagt, wann so ein in Gedanken ablaufendes Prüfungsprogramm zu enden hat und anhand welcher Kriterien man am Ende dieses Verfahrens die Ersetzbarkeit oder Nicht-Ersetzbarkeit entscheidet. Die Verwendung des Wortes „Verfahren“ suggeriert natürlich ein mechanisches Verfahren, auch wenn dieses Verfahren in Gedanken stattzufinden hat: Vielleicht hat Pascal an ein rein mechanisches Prüfungsverfahren gedacht, welches auch eine Maschine übernehmen könnte. Ist tatsächlich im technischen Sinne entscheidbar, ob ein definierter Ausdruck durch andere aus sich selbst verständliche Ausdrücke[13] ersetzbar ist? Für natürliche Sprachen kann man sich kein mechanisches Verfahren vorstellen, das nach endlich vielen Schritten in der Antwort ja oder nein auf die Frage, ob eine vorgelegte Definition eliminierbar ist, terminiert - aber für künstliche logische Sprachen erster (oder höherer) Stufe? Leider gilt das auch nicht für die Sprache erster Stufe. Wie man seit Churchs Unentscheidbarkeitssatz weiß, gibt es kein allgemeines mechanisches Verfahren, welches für jedes Satzschema bzw. Argumentschema der (engeren) Quantorenlogik nach endlich vielen Schritten die Frage beantwortet, ob es allgemeingültig ist oder nicht bzw. ob es korrekt ist oder nicht. Daher wird es auch kein Verfahren geben, mit dem man entscheiden kann, ob ein Zeichen/Ausdruck mit Hilfe anderer Ausdrücke einer Theorie definierbar ist - und damit auch nicht, ob ein Ausdruck einer Definition eliminierbar ist. Denn dazu müßte man feststellen können, ob ein bestimmtes Bikonditional aus der Theorie folgt oder nicht. Anders mag es sich bei der monadischen Quantorenlogik verhalten, die, wie man weiß, entscheidbar ist. Aber auch hier müßte man erst überhaupt einen Satz finden, der als Kandidat für eine Definition in Frage käme. Davon gibt es aber unendlich viele.

[11] Im Vergleich zu den Regeln, die unten angegeben werden, ist diese sogenannte „Regel“ nicht besonders konstruktiv: Hier liegt eher die Formulierung einer Bedingung vor, die die Definitionen erfüllen müssen, und nicht die Angabe von Richtlinien, die – wenn man sie befolgt – zur Aufstellung guter Definitionen verhelfen.

[12] Meint das in allen Ausdrücken einer Sprache?

[13] Welche sind das überhaupt?

### 3.1.2 Begründung der Eliminierbarkeit

Die Suche nach einer einleuchtenden Begründung für das Eliminierbarkeitskriterium wird zumindest im Falle von Suppes Buch vergebliche Müh sein. Auf S. 153 seines Buches [41] wird man lediglich ein Beispiel für einen eliminierbaren Ausdruck aufgeführt finden: Das Subtraktionssymbol '–', welches man in der Sprache der Arithmetik (mit dem Additionssymbol '+', einem Nullelementsymbol '0', einem Multiplikationssymbol '*' und eventuell einem zweistelligen Prädikatbuchstaben '<' für eine Ordnungsrelation) über

$$x - y = z \quad \Leftrightarrow_{df} \quad x = y + z$$

definiert, läßt sich mit Hilfe dieser Definition z.B. aus

$$\neg(y = 0) \ \rightarrow \ \neg(x - y = x)$$

eliminieren, wobei man dann das arithmetisch äquivalente[14] Schema

$$\neg(y = 0) \ \rightarrow \ \neg(x = y + x)$$

erhält. Dem Beispiel folgt die Bemerkung:

> It seems reasonable to require that any definition introducing a new symbol may be used to eliminate all subsequent meaningful occurrences of the new symbols. To be eliminable is a characteristic property of a defined symbol, as opposed to a primitive symbol. ([41], S.153-154)

Warum es '"reasonable" erscheint, erfährt der Leser nicht. Außerdem wird sich der Leser, der über die Logik erster Stufe hinausschaut, zweierlei fragen: Aus welchen Kontexten soll die Eliminierbarkeit gewährleistet sein? Und wann ist ein Vorkommnis eines Zeichens ein sinvolles? Die erste Frage stelle ich zurück. Was die zweite Frage betrifft, so könnten prinzipiell vier Dinge gemeint sein:

- Ein Zeichen A einer durch die Definition festgelegten logisch-grammatischen Kategorie C wird in einem bestimmten Zeichenkontext nicht gemäß der diese Kategorie auszeichnenden Gebrauchsvorschrift verwendet. Beispielsweise ist das oben definierte zweistellige Operationszeichen '+' im Kontext der folgenden Zeichenkette nicht korrekt verwendet, da es wie ein einstelliges Funktionssymbol gebraucht zu sein scheint: $(\forall x)(+5 = x)$

- Ein Zeichen A einer durch die Definition festgelegten logisch-grammatischen Kategorie C wird zwar syntaktisch korrekt benutzt, jedoch werden andere Zeichen, die sich in demselben Zeichenkontext wie das Zeichen A befinden,

[14] Es sind dabei zwei Schemata A und B arithmetisch äquivalent gdw A $\leftrightarrow$ B folgt (oder ist ableitbar) aus den Axiomen der Arithmetik. Allgemein sagt man, zwei Formeln A und B sind bzgl. eine Theorie T äquivalent gdw A$\leftrightarrow$B folgt (ist ableitbar) aus T

syntaktisch falsch verwendet. Beispielsweise findet sich '+' in der Zeichenfolge '$(\forall x)(\forall y)(x + y = *y)$' in einer syntaktisch mangelhaften Umgebung: Das zweistellige Operationssymbol '$*$' für die Multiplikation wird hier augenscheinlich als einstelliges verwandt.

- Der Zeichenkontext, in dem sich ein Zeichen A befindet, ist sinnlos aufgrund anderer denn syntaktischer Fehler bestimmter Zeichenkomponenten verschieden von A. Beispielsweise kommt das Zeichen '+' in der folgenden Zeichenfolge in einem derartigen Kontext vor: Das Blau da hoppelt über die Zahl zwei und 2 + 2 ist dasselbe wie 4.

- Das Zeichen A ist der Grund für die Sinnlosigkeit der Zeichenkette, in der es vorkommt – und das nicht aufgrund syntaktisch inkorrekter Verwendung von A. Beispielsweise ist das Zeichen '+' in der Zeichenkette 'Der Eiffelturm + 42 = 49' zwar syntaktisch korrekt verwendet worden, da links und rechts von '+' singuläre Terme stehen. Allerdings können gemäß der üblichen intendierten Bedeutung von '+' nur Zahlen und nicht auch physikalische Gegenstände addiert werden. Die Einschränkung auf die „*übliche* intendierte Bedeutung von '+' " trägt der wohlbekannten Tatsache Rechnung, daß Frege hier eine andere Position einnehmen würde. Für ihn gibt es keine partiellen oder bedingten Definitionen: Funktionen und Prädikate sind für den Bereich *aller* Gegenstände zu erklären, insbesondere ist also auch für das Paar <Eiffelturm, 42> anzugeben, welchen „Wert" + diesem zuordnet.

Bei der Aufführung des Eliminierbarkeitskriteriums im Rahmen von Sprachen, bei denen defekte Kontexte nicht ausgeschlossen sind, muß man sich also überlegen, welche defekten Kontexte als diejenigen auszuschließen sind, aus denen die Eliminierbarkeit des neu definierten Zeichens nicht gewährleistet sein muß.

Wenig Erhellendes findet man auch bei Essler zur Begründung der Eliminierbarkeit:

> Um die Adäquatheit einer solchen Definition des Definitionsbegriffs überprüfen zu können, benötigt man Kriterien, unter deren Zuhilfenahme man ihre Korrektheit nachweisen kann; nach solchen Kriterien muß nun also gesucht werden. Dabei überlegt man sich zweckmäßigerweise zunächst, was man mit einer derartigen Festsetzung bezweckt. Durch sie soll ausgedrückt werden, daß ein in der Theorie oder in der Alltagssprache schon zur Verfügung stehender Ausdruck oder aber ein neu in ihr einzuführender die gleiche Intension hat wie ein Komplex von anderen Begriffen der Theorie bzw. der Alltagssprache. Es wird damit angezeigt, daß dieser Ausdruck zwar vielleicht von großem *praktischem* Wert ist, da er die Formulierung von Aussagen und insbesondere von Gesetzen wesentlich erleichtert und uns damit zu einer besseren Übersicht über das Ausgesagte verhilft, daß er jedoch vom *theoretischen* Standpunkt aus überflüssig ist, da er in allen Verwendungen

> durch den Komplex aus jenen Begriffen ersetzt werden kann; eine Definition ist also eine Regel zur Ersetzung des definierten Ausdrucks durch andere. ([11], S. 79-80)

Hier fließt die Eliminierbarkeit gleich in den Definitionsbegriff ein, ohne daß eine ernsthafte Begründung gegeben wird.

Die erste Vorverständigung über den Begriff der Definition vorausgesetzt, kann sich Belnap ein (wenn auch sehr elementares) Bild über die Motivation der beiden Kriterien machen. Nach diesem Bild resultiert die Eliminierbarkeitsforderung aus der Vorstellung, Definitionen erläuterten die *gesamte* Bedeutung („all the meaning"), die ein Ausdruck besitzt oder besitzen soll, und die Nicht-Kreativitätsforderung aus der Vorstellung, Definitionen dürften nicht mehr tun, als die Bedeutung eines Ausdrucks angeben.[15]

Schauen wir uns Belnaps Überleung, wie eine Begründung von (EK) lauten könnte, etwas genauer an. In einer ersten Annäherung läßt sich Belnaps Argument so rekonstruieren[16]: Definitionen erläutern die Bedeutung von Ausdrücken. Daraus folge, daß Definitionen die gesamte Bedeutung von Ausdrücken erläutern. Für den folgenden Schritt stellt Belnap kommentarlos die These auf, daß die Bedeutung von Wörtern deren Gebrauch ist. Ich weiß nicht, wie mit dieser lax hingeworfenen Gleichsetzung umzugehen ist: Sollte mit ihr auf eine bestimmte Bedeutungskonzeption Bezug genommen worden sein oder nur eine zwar im einzelnen nicht korrekte, aber im groben und ganzen doch hinreichende Charakterisierung des Begriffs der Bedeutung gegeben worden sein? Wenn ersteres der Fall ist, dann stellt sich die Frage, ob das Argument wesentlich von dieser Bedeutungskonzeption abhängt. Ich bin mir nicht ganz darüber im Klaren, ob es eine Begründung für das Eliminierbarkeitskriterium gibt, die nicht auf einer konkreten speziellen Bedeutungskonzeption basiert; eine Begründung, die konzeptionsübergreifend ist, die also nicht eine bestimmte Antwort auf die Frage „Was ist die Bedeutung von Worten?" bedarf, wäre das Ideal.[17]

Wenn also die Bedeutung eine Ausdrucks dessen Gebrauch ist und die gesamte Bedeutung des Ausdrucks definiert werden soll, dann heißt das, daß alle seine Verwendungen erläutert werden sollen, das heiße aber, daß der Gebrauch eines Ausdrucks in allen Kontexten zu erläutern ist. Eine von der Tradition gemachte Einschränkung ist, den Gebrauch eines Ausdrucks in allen deklarativen Satzkontexten zu erläutern. Belnap weist zurecht darauf hin, daß das nicht zu einer adäquaten (Definitions-)Theorie für eine Sprache hinreichen kann. Die Erläuterung der Bedeutung eines Ausdrucks in einem Satzkontext B nun sei gleichwertig mit der Erläuterung der Bedeutung des beinhaltenden Satzes B. Auffällig ist,

[15] [5], S.119

[16] [5], S.119-120

[17] Nach Belnap und Gupta gibt es so eine Begründung naürlich nicht und kann es sie nicht geben. Bei ihrer Bedeutungskonzeption verletzten nämlich wesentlich zirkuläre Definitionen die Eliminierbarkeit.

daß Belnap hier wieder von der Bedeutung eines Ausdrucks und nicht seinem Gebrauch spricht. Allerdings ist es jetzt die Bedeutung eines Ausdrucks *in einem bestimmten Kontext* und nicht einfach die Bedeutung allein. Die Erläuterung der Bedeutung eines Satzes B sei gleichwertig mit der Erläuterung der Rolle von B in Folgerungen/Ableitungen. Hiermit ist nicht etwa gesagt, daß die Bedeutung eines Satzes irgendein Konglomerat aus all den Sätzen ist, die aus ihr folgen und/oder aus denen sie folgt, sondern daß man zur Klärung der Bedeutung eines Satzes nachzuschauen hat, wie der Satz logisch mit den anderen Sätzen zusammenhängt. Hier nun kämen zwei wichtige Überlegungen ins Spiel, die man als Forderung nach Nichtzirkularität und als das Verbot einer inferentiellen Bereicherung ansehen könne: „First, explanations quite generally are more prized if they are given in terms that are previously understood.“ ([5], S.120) Hier geht also in einer bestimmten Form das Verbot der Zirkularität ein. Leider macht sich Belnap nicht die Mühe, dieses Verbot zu motivieren. Fast scheint es, als ob er hier auch nur eine soziologische Beobachtung aufführen wollte, wonach Erläuterungen (von der Masse(?)) eher gepriesen würden, wenn sie nicht den zu erläuternden Ausdruck enthielten. Den zweiten Punkt formuliert Belnap folgendermaßen: „Second, in favorable situations we can hope to explain the inferential role of a new sentence by identifying it with that of an old sentence.“([5], S.120) Insgesamt erhalten wir also, daß es zu jedem Satz B, der den zu definierenden Ausdruck enthält, in der zugrunde gelegten Sprache einen Satz B' geben muß, der a) nicht den definierten Term enthält und b) dieselben logischen Verknüpfungen wie B aufweist. Ein Problem mit dieser Formulierung ist ihre Allgemeinheit. Hier wird gefordert, daß aus allen Satzkontexten Eliminierbarkeit gewährleistet sein muß. Wir haben bei der Besprechung eines Zitats von Suppes gesehen, daß schon bei der Frage, was denn die nichtdefekten Kontexte sind, aus denen Eliminierbarkeit gewährleistet sein soll, keine absolut eindeutige Anwort zu erwarten ist. Es muß eine Spezifizierung der Kontexte gegeben werden, aus denen Eliminierbarkeit gewährleistet sein soll. Denn z.B. aus Anführungskontexten würden man nicht unbedingt die Eliminierbarkeit verlangen wollen.

Wie diese Erwägungen zum Eliminierbarkeitskriterium exakt dargestellt werden können, wird im folgenden Abschnitt gezeigt.

### 3.1.3 Eine einheitliche Formulierung der Eliminierbarkeit und Nicht-Kreativität im Rahmen der Logik

Mit der (STDL) sind gewisse Einschränkungen verbunden, die im folgenden aufgelistet sind. Es sind dies von Belnap gemachte Vorschläge für den Rahmen, in dem die beiden Definitionskriterien und auch die zugehörigen Regeln formuliert werden sollen.[18] Bei Suppes sind diese Einschränkungen nicht in allen Details explizit aufgeführt: Aber auch er beschränkt sich bei der Formulierung der Eli-

[18] [5], S.126-127

minierbarkeit und Nicht-Kreativität auf Sprachen, die elementaren Sprachen der klassischen Quantorenlogik ähneln.

1. Die zugrunde liegende Sprache hat man sich als angewandte Sprache erster Stufe vorzustellen: Sowohl die *Grammatik* dieses Fragments der deutschen Sprache entspricht der Logik erster Stufe – das Fragment enthält folglich Prädikate, Funktionen, Satz- und Individuenkonstanten, die Identität, Quantoren, Variablen, wahrheitsfunktionale Junktoren etc. – als auch die *beweistechnischen Begriffe* – etwa „Axiom", „Regeln", „Theorem", „Ableitbarkeit aus Prämissen", „Theorie", „Äquivalenz bzgl. einer Theorie" etc. - und zu guter letzt auch die *semantischen Begriffe* - wie z.B. „logische Wahrheit", „Schlüssigkeit", „semantische Äquivalenz bzgl. einer Theorie" etc.
   Außerdem haben wir uns (mit Belnap) vorzustellen, daß den Verwendern dieser Sprache bekannt ist, daß es eine enge Beziehung zwischen beweistechnischen und semantischen Begriffen gibt. Belnap sagt: Die Sätze der ersten Stufe träten in *inferentiellen Kontexten* („inferential contexts") auf. Belnap denkt an eine durch eine Gruppe von Mathematikern gebildete Sprachgemeinschaft, die in einem halbformalen Englisch (bzw. Deutsch) über ihre mathematischen Probleme diskutieren.

2. Es werden lediglich Definitionen von Prädikat-, Funktions- und Individualkonstanten betrachtet. (Außen vor bleiben z.B. Russells Behandlung von definiten Kennzeichnungen; auch die Definition von Quantoren (mit beschränktem Gegenstandsbereich) wird nicht behandelt.)

3. Es werden lediglich Definitionen betrachtet, die satzförmig sind und die in derselben Sprache formuliert sind, die die definierten und definierenden Ausdrücke enthält. (Das ist ein häufig zu findender Zug in den klassischen Definitionstheorien, den Belnap und Gupta mit ihrer (RTD) aber nicht machen.)

Die beiden Kriterien für gute Definitionen werden selbst in Form einer wenn auch partiellen Definition eines vierstelligen Prädikats gegeben: Im Falle des Eliminierbarkeitskriteriums wird das Prädikat 'x als Definition von y erfüllt das Eliminierbarkeitskriterium relativ zu z und $z_1$' definiert; im Falle des Kriteriums der Nicht-Kreativität ist es das Prädikat 'x als Definition von y erfüllt das Nicht-Kreativitätskriterium relativ zu z und $z_1$'. Die in diese partiellen Definitionen eingehenden vier Schlüsselentitäten werden unten aufgelistet; ich schließe mich Belnaps mnemotechnisch sinnvoller Wahl der Variablen für diese Schlüsselentitäten an:

- *Theorie:*
  Das ist eine Menge von Sätzen der zugrunde gelegten Sprache erster Stufe. *Theorie* soll für die Hintergrundtheorie, in dessem Kontext die Definition eingeführt wird, stehen. (Das kann eine Menge von Sätzen sein, die bzgl.

Implikation oder Ableitbarkeit abgeschlossen ist; oder einfach eine kodierte Menge von Axiomen oder die Menge aller wahren Aussagen.[19])

- *vorliegende Definitionen*
  Das ist wieder eine Menge von Sätzen. Sie soll die Menge der bereits vorliegenden Definitionen repräsentieren.

- *Definition*
  Das ist ein Satz aus der zugrunde gelegten Sprache. (Dieser Satz soll die neu einzuführende Definition darstellen.)

- *Zeichen*
  Das ist das neue Zeichen, das definiert wird. Also ein Prädikat-, Operator-, Individual-, (Namen-) oder Satz- Buchstabe.

Mit diesen Präliminarien lassen sich Eliminierbarkeits- und Nicht-Kreativitätskriterium (jeweils durch '(EK)' und '(NK)' symbolisiert) genau formulieren:

**Eliminierbarkeitskriterium (EK)**
Seien *Theorie*, *vorliegende Definitionen*, *Definition* und *Zeichen* so, wie oben angeführt.[20]
*Definition* als Definition von *Zeichen* erfüllt das *Eliminierbarkeitskriterium* bezüglich *Theorie* und *vorliegende Definitionen* $\Leftrightarrow_{Df}$
Für alle (möglicherweise offenen) Sätze B in der Sprache von *Theorie*, *vorliegende Definitionen* und *Zeichen* gibt es einen (möglicherweise offenen) Satz B', so daß folgendes erfüllt ist:

(a) B' ist in der Sprache von *Theorie* und *vorliegende Definitionen* formuliert und

(b) *Zeichen* erscheint nicht in B' und

[19]Ist tatsächlich mit Belnap auch die Menge *aller* wahren Aussagen gestattet? Sind denn alle wahren Aussagen in der Logik erster Stufe wiederzugeben? Die Menge aller mathematischen Aussagen vielleicht - und das auch nur, wenn man sich mit der Mengenlehre anfreunden kann. Wenn man wie Donald Davidson der Meinung ist, für die Beschreibung der Tiefenstruktur der natürlichen Sprache sei die Prädikatenlogik mit Identität ausreichend, so gewinnt das Eliminierbarkeitskriterium (wie auch (NK)) an Universalität: Die Bindung an eine konkrete Theorie aus einem bestimmten wissenschaftlichen Bereich entfällt und Definitionen, wenn sie denn die Kriterien erfüllen, sind gute Definitionen – ohne Bindung an irgendeine Theorie.

[20]Hiermit wird also eine Vorauswahl für die Dinge getroffen, für die in der damit partiellen Definition das vierstellige Prädikat erklärt wird. Anders gesagt: Die in den Formulierungen von Definitionen meist weggelassenen Quantoren, welche die im zu definierenden Prädikat erscheinenden Variablen binden, haben alle einen von vornherein festgelegten eingeschränkten Wertebereich. Man stellt bei genauerem Hinsehen sogar fest, daß die Quantoren über verschiedene Wertebereiche laufen: *Theorie* und *vorliegende Definitionen* laufen über Klassen von Sätzen, *Definition* über Sätze und *Zeichen* über Zeichen.

(c) B und B' sind äquivalent bzgl. *Theorie*, *vorliegende Definitionen* und *Definition*

**Nich-Kreativitätskriterium (NK)**
Seien *Theorie*, *vorliegende Definitionen*, *Definition* und *Zeichen* so, wie oben angegeben.
*Definition* als Definition von *Zeichen* erfüllt das *Nicht-Kreativitätskriterium* relativ zu *Theorie* und *vorliegende Definitionen* $\Leftrightarrow_{Df}$
Für alle Sätze B in der Sprache von *Theorie* und *vorliegende Definitionen* (aber ohne *Zeichen*) gilt: Wenn aus *Definition*, *Theorie* und *vorliegende Definitionen* zusammen B folgt, dann folgt schon aus *Theorie* und *vorliegende Definitionen* zusammen B.

In der Formulierung des Eliminierbarkeitskriteriums findet sich ein Ausdruck, der einen gewissen Interpretationsspielraum läßt, nämlich das Wort 'Äquivalenz'.[21] Für unsere an der Logik erster Stufe orientierten Formulierungen der Kriterien gibt Belnap eine syntaktische und eine semantische Lesart des Ausdrucks vor; aufgrund der Vollständigkeit und der Korrektheit des Kalküls ergeben sich keine Unterschiede bei den beiden Lesarten.

(i) $B$ und $B'$ sind äquivalent gdw
$B \leftrightarrow B'$ (bzw. der universelle Abschluß) ist aus *Theorie*, *vorliegende Definitionen* und *Definition* in einem Standardkalkül ableitbar.

(ii) $B$ und $B'$ sind äquivalent gdw.
B und B' haben in jeder Interpretation der nicht-logischen Konstanten (und eventuell freien Variablen), die alle Sätze aus *Theorie*, *vorliegende Definitionen* und *Definition* wahr macht, denselben Wahrheitswert.

Die hier vorliegende Äquivalenz ist eine zwischen (geschlossenen) Satz*schemata* bzw. offenen Satz*schemata*; Belnap möchte diese Äquivalenz als eine Idealisierung von „inferentieller Äquivalenz" ansehen. Daß bei den Formulierungen der Kriterien die stärkere „inferentielle" Äquivalenz und nicht die schwächere „materiale" Äquivalenz gewählt werden müsse, sei bereits daraus ersichtlich, daß es keinen Sinn mache, von einer Relativierung der materialen Äquivalenz auf eine Theorie und auf vorliegenden Definitionen zu reden. Worin besteht die materiale Äquivalenz zweier Satzschemata? Für Sätze gilt die altbekannte Definition:
Zwei Sätze 'P' und 'Q' sind material äquivalent, wenn das Bikonditional 'P genau dann, wenn Q' wahr ist.[22] Zwei Satzschemata B und B' sind dann nach Belnaps Verständnis material äquivalent *relativ zu einer Interpretation*, wenn das Bikonditional $B \leftrightarrow B'$ unter dieser Interpretation wahr ist. Mit diesem Verständnis

[21] Ein ähnlicher Interpretationsspielraum wäre auch für (NK) gegeben, wenn statt des semantisch-lastigen Ausdrucks der Folgerung eine entsprechende Übersetzung des neutraleren englischen Ausdrucks 'consequence', wie ihn Belnap gebraucht, verwendet worden wäre.

[22] Ich würde hier dann von der extensionalen Äquivalenz reden.

der Worte 'materiale Äquivalenz' macht die Relativierung keinen Sinn.[23] Würde man die Relativierung wegglassen und in (EK) die Klausel „Für alle B... gibt es ein B' so daß ... B zu B' material äquivalent ist" aufnehmen, dann bräuchte man lediglich zwei Werte von B': einen wahren und einen falschen.[24]

Der Begriff der Äquivalenz erscheint Belnap als das Herz des Eliminierbarkeitskriteriums; ohne einen Rekurs auf eine bestimmte Form der Gleichwertigkeit/Äquivalenz ist ein Kriterium kein Eliminierbarkeitskriterium. Und je undurchsichtiger der Äquivalenzbegriff ist, um so weniger weiß man mit der Eliminierbarkeit anzufangen.

Man beachte, daß (EK) in einem gewissen Sinne liberal zu sein scheint. Schaut man sich (EK) nämlich genau an, dann bemerkt man, daß die logisch-grammatische Kategorie des zu definierenden Zeichens überhaupt keine Rolle spielt; es gibt eine einheitliche Formulierung für Prädikatsymbole und Funktionssymbole. Eine strengere Fassung von (EK), die diesen Mangel beheben könnte, würde z.B. für die Definitionen eines einstelligen Prädikatsymbols G fordern, daß sich zu jedem Satz B, welches G enthält, ein Satz B' finden läßt, der aus B durch Ersetzung von Vorkommnissen von G durch Ausdrücke der G-freien Sprache ensteht, so daß B und B' gemäß der Theorie und den vorliegenden Definitionen sowie der Definition selbst äquivalent sind. Die Konstruktion könnte man sich folgendermaßen gebildet denken: Sind $t_1$, $...t_n$ Terme und sind die einzigen Vorkommnisse von G in B unter den Ausdrücken $\ulcorner Gt_1 \urcorner, ..., \ulcorner Gt_n \urcorner$ , dann entsteht B' durch Ersetzung von $\ulcorner Gt_1 \urcorner, ..., \ulcorner Gt_n \urcorner$ jeweils durch $\ulcorner A_1(t_1) \urcorner, ..., \ulcorner A_n(t_n) \urcorner$, wobei $\ulcorner A_i(t_i) \urcorner$ das Ergebnis der Ersetzung aller Vorkommnisse der Variablen 'x' des offenen Satzes $A_i(x)$ der G-freien Sprache durch den Term $t_i$ ist. Sie wirkt ein wenig unplausibel: Zwar wird jedes Vorkommnis des Definiendums G in B durch einen G-freien Ausdruck derselben Kategorie ersetzt, aber es ist nicht gefordert, daß dieser Ausdruck immer derselbe ist. Tatsächlich ist also die eigentliche strikte Formulierung die, welche man aus obiger erhält, indem man fordert: daß gilt: $\ulcorner A_1(x) \urcorner$ ist dieselbe Formel wie $\ulcorner A_2(x) \urcorner$ ist dieselbe Formel wie ... ist dieselbe Formel wie $\ulcorner A_n(x) \urcorner$ ist dieselbe Formel wie $A(x)$. Man könnte dann sagen, daß das Prädikat G durch den Ausdruck $A(x)$ strikt eliminierbar ist.

Ein Blick auf diese beiden Konstruktionen zeigt aber, daß sie miteinander äqivalent sind und abgemagert werden können zu folgender Forderung: Aus der Theorie, den vorhergehenden Definitionen und der Definition des einstelligen Prädikatbuchstabens G ist eine Aussage der Form $\ulcorner G(\alpha) \leftrightarrow A(\alpha) \urcorner$ mit einem G-freien, in der Variablen $\alpha$ offenen Ausdruck $A(\alpha)$ folgerbar/ableitbar. Diese Forderung ist aber gerade die nach der expliziten Definierbarkeit des Prädikats

[23]Ich glaube dennoch, daß Belnap an die materiale Äquivalenz von Sätzen denkt; da ist die Relativierung auf eine Theorie noch unsinniger.

[24]Beachte, daß die Formalisierungen zweier Sätze in QL (mit Identität), die intensional äquivalent sind, nicht inferentiell gleichwertig zu sein brauchen: So sind die beiden Sätze 'Kant ist Junggeselle' und 'Kant ist lediger Mann im heiratsfähigen Alter' zwar intensional gleichwertig, aber deren Schematisierungen etwa durch 'Fa' und 'Ga' sind nicht inferentiell äquivalent.

G, die weiter unten bei der Besprechung der Regeln für Definitionen erwähnt wird. Es läßt sich aber nachweisen, daß diese Forderung tatsächlich keine striktere als die der Eliminierbarkeit ist: Beides ist gleichwertig.

Verhält es sich mit den Funktionssymbolen und Namenbuchstaben genauso? Die Anwort ist: Nein! Daß die Verhältnisse nicht denen für Prädikatbuchstaben gleichen, dafür sorgt die weiter unten besprochene Regel für die explizite Definition von Funktionssymbolen (bzw. Namenbuchstaben). Ich möchte hier nicht vorwegnehmen, was genau diese Regel besagt; stattdessen will ich das anführen, was sie nicht ist: Diese Regel fordert nicht die Existenz eines f-freien Ausdrucks, der die Rolle des zu definierenden Funktionssymbols f übernimmt und von derselben logisch-grammatischen Kategorie ist. Diese intuitive Überlegung wird mit der strikten Eliminierbarkeit eingefangen. Normalerweise spricht man nicht von der strikten Eliminierbarkeit sondern strikten Definierbarkeit.

Die strikte Definierbarkeit von Funktions-und Namensbuchstaben kann man als Definition durch die Identität verstehen. Bisher waren es Äquivalenzen von (offenen) Sätzen, die die Definitionen bildeten – eine auf Prädikatbuchstaben gerade gut passende Form der Definition. Für Funktions- und Namenbuchstaben scheint aber die Definition per Identität naheliegender. Wenn z.B. in der Theorie der Arithmetik das Funktionssymbol '+' für Addition und der Namnebuchstabe '1' für die Eins vorliegen, dann könnte man das einstellige Funktionssymbol 'N', das für die Nachfolgeroperation auf der Menge der natürlichen Zahlen stehen soll über die Identität N(x) = x + 1 definieren. Nicht immer aber ist die strikte Definierbarkeit möglich.[25]

### 3.1.4 Eliminierbarkeit und Zirkularität

Mit dem Eliminierbarkeitskriterium wird nicht ausgeschlossen, daß wenn der Kanditat *Definition* für eine propere Definition von der Form eines Bikonditionals ist, links und rechts vom Doppelpfeil das zu definierende Zeichen *Zeichen* vorkommt, also in einem bestimmten Sinne zirkulär ist. Ist z.B. *Definition* der Satz '$(\forall x)$ ($x$ ist ein Junggeselle $\Leftrightarrow_{df}$ $x$ ist lediger Mann im heiratsfähigen Alter & ($x$ ist ein Junggeselle $\rightarrow$ x ist ein Junggeselle)), dann erfüllt *Definition* die Eliminierbarkeit (aus bestimmten Kontexten), obwohl der zu definierende Ausdruck 'ist ein Junggeselle' auch im Definiens vorkommt; diese Zirkularität ist aber, wenn man so will, eine harmlose Zirkularität, da der Ausdruck rechts vom obigen Bikonditional logisch äquivalent ist zu einem Ausdruck, der den Ausdruck 'ist ein Junggeselle' nicht enthält. Das wirft die Frage auf, welche Formen der Zirkularität gemäß (EK) gestattet sind. Das Junggesellenbeispiel liefert eine hinreichende

[25]Die strikte Definierbarkeit von Funktionssymbolen spielt eine prominente Rolle in der Beweisbarkeitstheorie. Kann man nämlich zeigen, daß die Normfunktion, also die Funktion, welche einem Ausdruck $E$ den Ausdruck $E$ gefolgt von der Gödelzahl von $E$ zuordnet, relativ zu einer Gödelkodierung strikt definierbar ist, dann ist das Sytem sehr reichhaltig, so reichhaltig nämlich, daß es Formeln beinhalten kann, die im Prinzip sagen 'Ich bin nicht beweisbar'.

Bedingung für im Sinne von (EK) harmlose Zirkel. Wollen wir, daß diese Bedingung auch notwendig ist, so legen wir uns auf eine bestimmte Form von harmloser Zirkularität fest, die ich l-harmlos (für **l**ogisch-harmlos) nennen möchte.

> **l-harmlose Zirkularität**
> Für Definitionen, die als Bikonditional formuliert sind, ist das Vorkommnis des Definiendums G im Definiens l-harmlos $\Leftrightarrow_{Df}$ es gibt einen zum Definiens logisch äquivalenten Ausdruck, der G nicht enthält.

Könnte man sich denn andere harmlose zirkuläre Definitionen vorstellen, die (EK) als propere, da eben mit harmloser Zirkularität ausgestattete Definitionen durchgehen läßt?[26] Wenn man die in die Formulierung der l-harmlosen Zirkularität eingehende Forderung, daß sich für das Definiens ein logisch äquivalenter G-freier Ausdruck finden läßt, abschwächt zu der Forderung, daß sich eine relativ zu der Theorie und den vorliegenden Definitionen äquivalente Formel finden lassen muß, dann erhält man eine andere Form von harmloser Zirkularität, welche ich T-harmlos (für Theorie-harmlos) nennen möchte. Dabei ist auch der Fall gestattet, daß die Theorie T die leere Menge ist.

> **T-harmlose Zirkularität**
> Für Definitionen, die als Bikonditional formuliert sind, ist das Vorkommnis des Definiendumds G im Definiens T-harmlos $\Leftrightarrow_{Df}$ es gibt einen G-freien Ausdruck, der relativ zur Theorie und den vorliegenden Definitionen äquivalent zum Definiens ist.

Definitionen, die T-harmlose Zirkularität aufweisen, erfüllen (EK). Ein Beispiel für eine derartige harmlose Zirkularität wird im folgenden gegeben werden. Im übrigen sind alle l-harmlosen zirkulären Definitionen auch T-harmlos zirkulär.

Erneut stellt sich die Frage, ob damit alle, von (EK) gestatteten harmlosen Zirkularitäten aufgezählt sind. Diesmal ist die Anwort: Ja. Es gibt aber Definitionen, von denen man sagen würde, daß sie, obwohl sie (EK) nicht erfüllen, eine harmlose Zirkularität enthalten. Hierunter sind implizite Definitionen und induktive Definitionen (s.u.)

---

[26] Hier läuft die Frage darauf hinaus, eine noch größere Menge an gemäß (EK) harmlosen zirkulären Definitionen zu erfassen. Es ließe sich natürlich aber auch die andere, vielleicht theoretisch nicht, aber doch möglicherweise praktisch interessante Richtung verfolgen, eine Differenzierung unter den l-harmlosen Zirkularitäten vorzunehmen. Es ließe sich z.B. (jetzt für die Aussagenlogik formuliert) die sl-harmlose Zirkularität (für stark logisch harmlos) formulieren, die eine Teilmenge der l-harmlosen Zirkularitäten bilden. Für Definitionen, die als Bikonditional formuliert sind, sind die Vorkommnisse des Definiendums G im Definiens sl-harmlos $\Leftrightarrow_{Df}$ es gibt eine allgemeingültige Teilformel des Definiens, in dem sich alle Vorkommnisse von 'G' befinden.
Z.B. sind die Vorkommnisse von 'G' im Definiens '$Hb$ & $(Gx \vee \neg Gx)$' sl-harmlos. In dem Definiens '$(Gx \rightarrow (Hx \,\&\, Fx)) \,\&\, (\neg Gx \rightarrow (Hx \,\&\, Fx))$' sind die Vorkommnisse von 'G' nur l-harmlos und nicht sl-harmlos.

Schauen wir uns ein Beispiel für eine zirkuläre Definition mit T-harmlosem, aber nicht l-harmlosem Vorkommnis des Definiendums im Definiens an. In seinem Artikel [44] bespricht Tichy die Bedingungen für harmlose Zirkel innerhalb von Definitionen, welche in einer Sprache erster Stufe formuliert sind, und gibt ein Verfahren an, mit dem sich das Vorkommnis des zu definierenden Zeichens im Falle eines harmlosen Zirkels aus dem Definiendum eliminieren läßt. Er beginnt seinen Artikel mit einem Beispiel einer properen Definition, die einen *harmlosen* Zirkel enthält, obwohl es keinen logisch äquivalenten Ausdruck für das Definiens dieser Definition gibt. Es wird sich aber zeigen, daß für diese Definition als ganzer eine logisch äquivalente Definition finden läßt, die kein Vorkommnis des Definiendums im Definiens enthält:[27] Also liegt hier für den Spezialfall, daß die Theorie T die leere Menge ist, eine T-harmlose, aber nicht l-harmlose Zirkularität vor.

Nehmen wir an, daß in der gegebenen Sprache erster Stufe die folgende mit einer metasprachlichen Abkürzung formulierte Formel (wie Tichy sagt:) analytisch wahr ist, d.h. (wohl) daß die Formel logisch allgemeingültig ist:

$(\exists x)\neg A(x)$

Dabei ist 'A(x)' ein Kürzel für ein in der Variablen 'x' offenes Satzschema. Zur Veranschaulichung kann man sich konkret für A(x) folgendes denken:

$$\begin{aligned} A(x) \quad \text{sei} \quad & (\forall w)(\forall y)(\forall z)(w + (y + z) = (w + y) + z) \;\&\\ & (\forall z)(z + 0 = z) \;\&\\ & (\forall z)(\exists y)(z + y = 0) \;\&\\ & (\forall z)(\neg z + x = z) \end{aligned}$$

In diesem meinem Beispiel liegt die Sprache der additiv geschriebenen Gruppen vor. Die ersten drei Konjunkte sind die Gruppenaxiome, das letze Konjunkt besagt, daß für alle Gegenstände z des Wertebereichs gilt: Die Addition von x auf z ist nicht identisch mit z ist. Unter allen Interpretationen der Formel $(\exists x)\neg A(x)$ ist diese Formel wahr: Ist eines der Gruppenaxiome unter der Interpretation I falsch, dann ist auch $A(x)$ unter I falsch. Sind unter I hingegen die Gruppenaxiome wahr, dann ist insbesondere das zweite Axiom wahr, das die Existenz eines neutralen Elementes 0 sichert, und dieses erfüllt nicht das letzte Konjunkt, d.h. für dieses Element ist $A(x)$ falsch , also $\neg A(x)$ wahr. Man betrachtet nun die folgende Definition des einstelligen Prädikatbuchstabens G:

$$Gx \quad \Leftrightarrow_{df} \quad (\exists y)(\neg Gy \;\&\; A(x))$$

Bei der obigen Ausbuchstabierung von $A(x)$ ist nun tatsächlich das Definiens dieser zirkulären Definition mit keinem G-freien Ausdruck äquivalent. Denn angenommen es wäre doch äquivalent zu einem G-freien Ausdruck B(x). Dann stellt man zunächst fest, daß $A(x)$ nicht unter jeder Interpretation falsch ist: Wählt man nämlich eine Interpretation, die die Gruppenaxiome wahr macht und außerdem der Variablen 'x' einen Gegenstand aus dem Wertebereich zuordnet, der nicht

[27] [44], S.19-20

gerade das Nullelement 0 ist, dann ist $A(x)$ unter dieser Interpretation wahr. Ist nun I1 eine Interpretation mit einem Wertebereich W, die G den gesamten Wertebereich W und der Variablen x einen Gegenstand aus D, der $A(x)$ erfüllt (- solch einen gibt es ja, wie gezeigt -), zuordnet, dann ist das Definiens unter dieser Interpretation I1 falsch. Unter der Interpretation I2 mit demselben Wertebereich W, die G allerdings nicht den gesamten Wertebereich W zuordnet , ansonsten aber mit I1 übereinstimmt, ist das Definiens wahr. Folglich haben wir für die beiden Interpretationen unterschiedliche Wahrheitswerte. B(x) aber hat für I1 und I2 denselben Wahrheitswert, da I1 und I2 sich nur in der Interpretation des Prädikatsbuchstabens G unterscheiden; dieser kommt aber nicht in B(x) vor (q.e.d.).
Obwohl das Definiens wie gesehen nicht zu einer G-freien Formel äquivalent ist, ist die Definition von G in einem bestimmten Sinne proper: Es gibt eine Interpretation der Definition, die die Definition wahr macht, und für je zwei Interpretationen I1 und I2, die Definition wahr machen und für alle von G verschiedenen nichtlogischen Zeichen aus A(x) übereinstimmen, gilt, daß sie auch für G über einstimmen, daß also I1 = I2 gilt.
Sei also eine Interpretation mit Wertebereich W gegeben, unter der A(x) die Teilmenge M von W zugewiesen bekommt. Zunächst wird gezeigt, daß dieses M dem G zugeordnet werden kann, so daß die Definition wahr wird. Aufgrund der Allgemeingültigkeit der Formel $(\exists x)\neg A(x)$ gibt es einen Gegenstand b aus W, der nicht in M liegt. Sei a ein beliebiger Gegenstand aus W. Ist a $\in$ W, so ist zu zeigen, daß a das Definiens erfüllt. Da b $\notin$ M, aber a $\in$ M, ist das Definiens durch a erfüllt. Ist umgekehrt a$\notin$ M, dann ist auch das Definiens nicht durch a erfüllt (q.e.d.).
Nun wird die Eindeutigkeit gezeigt. Angenommen N ist eine von M verschiedene Teilmenge von W; nehmen wir also an, daß sich N und M bzgl. des Punktes c unterscheiden; und weiter angenommen, daß bei der Zuordnung von N zum Prädikat G die Definition wahr wird. Für das obige Element b gilt: b$\notin$ N. Ist nun c $\in$ N, dann gilt c$\notin$ M, so daß c zwar das Definiendum erfüllt, nicht aber das Definiens. Daher ist c$\notin$N. Folglich ist c$\in$M, dann aber, da b$\notin$ N, erfüllt c das Definiens ohne das Definiendum zu erfüllen. Aufgrund dieses Widerspruchs folgt also N=M. (q.e.d.)

Wir haben uns vergewissert, daß es gemäß (EK) sogar T-harmlos zirkuläre Definitionen gibt. Wenn dem jetzt so ist, kann man dann noch sagen, Eliminierbarkeit impliziert das Zirkularitätsverbot? Zumindest die sehr starke Version des Zirkularitätsverbotes, die alle Definitionen in Form eine Bikonditionals mit einem Vorkommnis des Definiendums im Definiens verbietet, wird nicht impliziert. Dann vielleicht die abgeschwächte Version des Zirkularitätsverbots, die harmlose Vorkommnisse des Definiendums im Definiens durchgehen läßt? Für diesen Fall scheint die Eliminierbarkeit tatsächlich das Zirkularitätsverbot zu implizieren. Aber ein selbständiges, nicht auf die Eliminierbarkeit rekurrierendes Zirkularitätsverbot gibt es hier nun nicht mehr: Eine Definition von der Form eines Bikonditionals mit Vorkommnissen von *Zeichen* links und rechts vom Doppelpfeil

ist ja auf fatale Weise zirkulär genau dann, wenn es nicht das Eliminierbarkeitskriterium erfüllt.

# 4 Regeln für Definitionen in (STDL)

Im folgenden sind die Standardregeln für die Definition von Prädikatbuchstaben und Funktionsbuchstaben aufgeführt. Namenbuchstaben lassen sich als 0-stellige Funktionsbuchstaben ansehen und können so mit der Regel für die Definition von Funktionsbuchstaben eingefangen werden. Sie lassen sich natürlich auch separat behandeln; der Kürze wegen ziehe ich ersteren Weg vor. Definitionen, die den Regeln gemäß konstruiert wurden, erfüllen die beiden oben formulierten Kriterien. Außerdem gilt auch die umgekehrte Richtung, die man als eine Art Vollständigkeit ansehen kann; das zeige Belnap gemäß das Beth'sche Definierbarkeitstheorem.[28]

Die Regeln werden wieder in Form von Definitionen formuliert.

**Regeln für Prädikatbuchstaben**
*Definition* ist eine Standarddefinition eines Prädikatbuchstabens R relativ zu *Theorie* und *vorliegende Definitionen* $\Leftrightarrow_{df}$
es gibt eine natürliche Zahl n (die Stelligkeit von R), Variablen $v_1, ..., v_n$ und einen Satz A (das „Definiens"), so daß folgende Bedingungen erfüllt sind:

(a) *Theorie* und *vorliegende Definitionen* sind Mengen von Sätzen, R ist ein n-stelliger Prädikatbuchstabe (der zu definieren ist) und *Definition* ist ein Satz (die potentielle Definition).

(b) *Definition* ist ein n-mal universell quantifiziertes Bikonditional $(\forall v_1)...(\forall v_n)(Rv_1...v_n \leftrightarrow A)$.

(c) Die Variablen $v_1, ..., v_n$ sind verschieden.

(d) Das Definiens A hat keine anderen freien Variablen als $v_1, ..., v_n$.

(e) Jedes nicht-logische Symbol in A gehört der Sprache von *Theorie* und *vorliegende Definitionen* an.

(f) R kommt nicht in *Theorie* und *vorliegende Definitionen* vor.

[28] [5], S.139. Leider sagt Belnap nicht, wo genau dieser Definierbarkeitssatz eingeht. Der Bethsche Definierbarkeitssatz besagt, daß jedes Zeichen $\alpha$, welches in einer Theorie T implizit definierbar ist, auch explizit definierbar ist. Dabei heißt ein Zeichen $\alpha$ implizit definierbar in T gdw für je zwei Modelle von T (in der Sprache, die auch das $\alpha$ enthält) die Interpretation von $\alpha$ dieselbe ist. Tatsächlich sieht man, daß die Eliminierbarkeit eines einstelligen Prädikats G etwa äquivalent mit der expliziten Definierbarkeit von G ist - aber wo spielt dort der Bethsche Satz eine Rolle? Der Bethsche Satz wird mit Hilfe des Interpolationslemmas bewiesen, der besagt: Für alle Sätze $A, C$ der ersten Stufe mit Identität gilt: Ist $A \vdash C$, dann gibt es einen Satz C, der folgendes erfüllt: Er enthält nur solche nicht-logischen Konstanten, welche in A und B vorkommen, und es gilt $A \vdash C$ sowie $C \vdash B$. Nicht für alle Logiken gilt der Bethsche Definierbarkeitssatz.

Insbesondere folgt, daß R nicht in A vorkommt: zirkuläre Definitionen von Prädikatbuchstaben sind durch (e) und (f) ausgeschlossen.

Ganz ähnlich lautet die Standardregel für die Definition eines Funktionssymbols:

> **Regeln für Funktionssymbole**
> *Definition* ist eine Standarddefinition eines Funktionssymbols O relativ zu *Theorie* und *vorliegende Definitionen* $\Leftrightarrow_{df}$
> Es gibt eine natürliche Zahl n (die Stellenzahl von O) und Variablen $v_1, ..., v_n$, $w$ und einen Satz A (das Definiens), so daß folgende Bedingungen erfüllt sind:
>
> (a) *Theorie* und *vorliegende Definitionen* sind Mengen von Sätzen, O ist ein n-stelliges Funktionssymbol und *Definition* ist ein Satz.
>
> (b) *Definition* ist ein (n+1)-mal universell quantifiziertes Bikonditional $(\forall v_1)...(\forall v_n)(\forall w)((Ov_1...v_n = w) \leftrightarrow A)$.
>
> (c) Die Varablen $v_1...v_n, w$ sind verschieden.
>
> (d) Das Definiens A hat keine anderen freien Variablen als $v_1, ..., v_n, w$.
>
> (e) Jedes nicht-logische Symbol in A gehört der Sprache von *Theorie* und *vorliegende Definitionen* an.
>
> (f) O kommt nicht in *Theorie* und *vorliegende Definitionen* vor.
>
> (g) Der Satz $(\forall v_1)...(\forall v_n)(\exists y)(\forall w)((y = w) \leftrightarrow A)$ folgt aus bzw. ist ableitbar aus *Theorie* und *vorliegende Definitionen*. (D.h. aus der Theorie und den vorliegenden Definitionen folgt, daß für alle n-Tupel $(v_1, ..., v_n)$ genau ein w existiert, so daß A gilt. A ist sozusagen „funktional“.)

Insbesondere folgt aus (e) und (f), daß das Funktionssymbol O nicht in A vorkommt: also ist auch die zirkuläre Definition von Funktionssymbolensymbolen (und Namenbuchstaben) ausgeschlossen.

In diesem Zusammenhang gibt Belnap keine unabhängige Begründung für die einzelnen Regeln; da Belnap das Projekt der Regelformulierung für Definitionen als dem Projekt für die Formulierung von Kriterien für Definition untergeordnet ansieht, wäre es auch Fehl am Platze, eine derartige Begründung zu fordern. Man könnte sogar zu folgender Meinung gelangen: In Anbetracht der Tatsache, daß manche Vorkommnisse von zu definierenden Ausdrücken im Definiens harmlos sind in dem Sinne, daß die zugehörigen Definitionen trotz der Zirkularität das Elimierbarkeitskriterium erfüllen, scheint nicht mal von der Kriterienseite her die strenge Fassung begründet werden zu können. Hierzu läßt sich nur sagen: Daß das aus (e) und (f) resultierende Zirkularitätsverbot in diesem Rahmen, i.e. den durch die beiden Kriterien (NK) und (EK) vorgegebenen, also in jedem einzelnen Falle - sozusagen lokal - in seiner strengen Fassung nicht begründet werden kann,

bedeutet nicht, daß diese Regel im ganzen gesehen - sozusagen global - zu streng ist. Der Beth'sche Definierbarkeitsatz rechtfertigt die Regeln und weist sie als nicht zu streng aus: Da sich zu jedem Satz *Definition*, der die Kriterien (EK) und (NK) erfüllt[29], ein Satz mit der in den Regeln gegebenen kanonischen Form finden läßt, ist man schlichtweg auf keine harmlos zirkulären Definitionen angewiesen, also verliert man auch nichts, wenn man sie wegläßt.

Man vergesse bei all diesen Überlegungen nicht, daß (EK) die Zirkularität in einem anderen Sinne verbietet: Das Kriterium fordert für jeden Satz B, der (eventuell) das zu definierende *Zeichen* enthält, die Existenz eines Satzes B', welches nicht *Zeichen* enthalten darf. Wenn man es so sehen möchte, könnte man sagen, daß (EK) die „Erklärung" des Satzes B durch einen Satz B' der alten Sprache fordert, ohne daß dabei das zu definierende Zeichen benutzt wird. Auch wenn (EK) in dem Satz *Definition*, von dem wir annehmen können, es sei in Form eines Bikonditionals, nicht ausschließt, daß das zu definierende *Zeichen* auch auf der rechten Seite des Bikonditionals vorkommt, so verbietet es doch den Fall, daß in dem Bikonditional B↔B', welches aus *Theorie* und *vorliegende Definitionen* folgt bzw. ableitbar ist, auf der rechten Seite des Doppelpfeils, i.e. in B', *Zeichen* vorkommt.

## 4.1 Begründungsversuche für Zirkularitätsverbot

Für uns, die wir auf die Regeln für Definitionen im Rahmen der beiden Kriterien (NK) und (EK) schauen, mag das durch die Unterpunkte (e) und (f) implizierte Zirkularitätsverbot in dieser Fassung in lokaler Perspektive als zu streng erscheinen.[30] Welche Motivation für das Zirkularitätsverbot könnte es gegeben haben oder geben, wenn wir Definitionen als die Angabe der Bedeutung bestimmter sprachlicher Entitäten ansehen?

Leonard in [29] führt, nachdem als eine Regel das Verbot der Zirkularität vorgebracht wurde, gegen zirkuläre Definitionen an, sie müßten abgelehnt werden, „... because they do not explain the meaning of the definiendum: a person who did not already understand the definiendum could not understand the definiens."[31] Der Proponent des Zirkularitätsverbots könnte – so Yablo in [48] – zu seiner Ansicht durch eine zwar naheliegende, aber doch unüberlegte Antwort auf die Frage gelangt sein, wie eine Definition z.B. des einstelligen Prädikats G über $Gx \quad \Leftrightarrow_{df} \quad A(x)$ die Bedeutung des Definiendums (G bzw. Gx) determiniert/festlegt/bestimmt: nämlich die bildbefrachtete Antwort, daß G seine Bedeutung vom Definiens A(x) erbt. So gesehen könnte auch Leonards Begrün-

[29] Insbesondere auch für Bikonditionale, die links und rechts ein Vorkommnis des zu definierenden Zeichens enthalten.

[30] Über das Zirkularitätsverbot bei anderen mit dem Ausdruck 'Definition' verbundenen Projekten habe ich nichts gesagt: da mag sich das Zirkularitätsverbot auch in der strengen Fassung als völlig berechtigt herausstellen.

[31] Zitiert nach [48], S.364

dung einleuchten: Damit A(x) dem G seine Bedeutung vererben kann, muß es selbst eine Bedeutung haben - und das könnte es nicht, wenn es das noch nicht mit einer Bedeutung bestückte 'G' enthielte. Hier stellt sich natürlich sofort die Frage, ob jemand, der sich die Festlegung der Bedeutung eines Ausdrucks als eine Vererbung vorstellt, nicht mit bestimmten, oben schon vorgestellten harmlosen Zirkularitäten leben kann: Wenn ich die Bedeutung der einstelligen Prädikate 'F' und 'H' kenne, von 'G' zumindest weiß, daß es ein einstelliges Prädikat ist, dann ist die Bedeutung des Definiens (*) '$Fx \vee Hx$ & $(Gx \vee \neg Gx)$ schon determiniert, auch wenn 'G' noch keine Bedeutung zugewiesen bekommen hat. Der Proponent des Zirkularitätsverbots müßte dieses folglich modifizieren und diejenigen Fälle von Vorkommnissen des Definiendums im Definiens verbieten, die in der Position sind, etwas zur Bedeutung des gesamten Definienskomplexes beizutragen. Wie würde man diese schwammige Rede von einer bestimmten Position erhellen können? Die naheliegendste Lösung ist, daß man hier die oben schon angedeutete Äquivalenz ins Spiel bringt: Ist das Definiens äquivalent zu einem G-freien Ausdruck, dann war das Vorkommnis von G gestattet. Der Pragmatiker wird sich fragen: „Hilft das aber demjenigen weiter, der die Bedeutung von G nicht kennt und sich nun anhand der, wie es nun zufälligerweise der Fall ist, harmlos zirkulären Definition für G hierüber Klarheit verschaffen möchte?" Im Falle einer Definition mit einem Definiens wie in (*) ist für den G-Ungebildeten die Bedeutung des Definiens schnell einzusehen.

Pragmatische Aspekte von Definitionen stehen für Yablo, der die oben dargestellte Vermutung über die Motivation für das Zirkularitätsverbot angestellt hat, nicht zur Debatte. Für ihn zählt, daß es auch andere Wege gibt, mit Hilfe des Definitionsschemas '$Gx \quad \Leftrightarrow_{Df} \quad A(x)$' die Bedeutung des zu definierenden Prädikats G festzulegen. Yablo führt weitere Konventionen an, mit deren Hilfe sich aus dem Definitionsschema eine Festlegung der Bedeutung von G gewinnen läßt. Das führt er bis zu den Kriterien (E), (F) und (G) und nennt eine Definition konsistent, wenn sie alle drei dieser Kriterien erfüllt. Das Kriterium (G) fischt dabei sozusagen die meiste Information aus dem Definitionsschema heraus: (G) gestattet auch die krudesten Fälle (insbesondere bestimmte Fälle von Zirkularität, bei der das zu definierende Prädikat G sowohl positiv als auch negativ im Definiens vorkommt). Dabei hat er auch immer den klassischen Begriff der Bedeutung eines Prädikats im Auge, welcher zumindest dieses besagt, daß die Angabe der Bedeutung eines einstelligen Prädikats die Angabe der Anwendungsbedingungen des Prädikats involviert. Das zum Definitionsschema zugehörige (abquantifizierte) Bikonditional, mit dem Definiendum links vom Doppelpfeil und dem Definiens rechts davon, ist auch für Yablos vorgeschlagene anderen Weisen, wie man das Definitionsschema zur Festlegung der Anwendungsbedingungen eines Begriffs nutzen kann, wahr. Das wird in der (RTD) von Belnap und Gupta z.B. nicht mehr der Fall sein.

Man kann sich also überlegen, auf welche Weisen man mit einem Definitionsschema dem zu definierenden Ausdruck seine Bedeutung zuweisen möchte. Die Vererbungstheorie, die auf einer Art Gleichsetzungsverfahren beruht, ist eine

Möglichkeit. Es gibt aber andere Möglichkeiten, wie sie Yablo vorführt. Einer der bekannteren Möglichkeiten, das Definitionsschema auf eine andere Weise zu lesen als mit dem Gleichsetzungsverfahren, sind die rekursiven Definitionen.

## 4.2 Rekursive Definitionen

Rekursive Definitionen sind kontextuelle Definitionen. Ein Beispiel für eine rekursive Definition ist die übliche Definition des Additionsoperators '+' durch:

$$x + y = z \quad \Leftrightarrow_{Df} \quad [(y = 0 \;\&\; z = x) \vee (\exists v)(\exists w)(y = v' \& z = w' \& x + v = w]$$

mit der Nachfolgeroperation $'$. Diese Definition gestattet die Elimination des Zeichens '+' nur aus bestimmten Kontexten; in dem Kontext des folgenden Satzes (Satzschemas) ist das Additionszeichen z.B. nicht eliminierbar:

$$(\forall x)(\forall y)(x + y = y + x)$$

Denn an diesem Beispiel läßt sich weder die erste Klausel (erstes Disjunkt) anwenden - da die 0 nicht vorkommt - noch die zweite (zweites Disjunkt), da nirgends der Nachfolgeoperator in Erscheinung tritt. Dafür ist die Elimination aus dem Kontext des folgenden Satzes gewährleistet:

$$3 + 2 = 5$$

Dieser Satz wird gelesen als Abkürzung des Satzes

$$0''' + 0'' = 0''''$$

Mehrfache Anwendung der zweiten Klausel und am Ende der Anfangsklausel liefert einen Satz, der das Additionszeichen nicht mehr enthält:

$$0'''' = 0''''$$

Die obige Definition ist aber, obwohl sie meist in dieser Form angegeben wird, nicht vollständig. Denn es fehlt eine Endklausel, die besagt, daß nur diejenigen Tripel (x,y,z) zur Extension der Additionsfunktion gehören sollen, die sich durch endlichmalige Anwendung der beiden Klauseln erhalten lassen. Ohne die Endklausel würde nicht ausgeschlossen werden, daß auch andere nicht für die Additionsfunktion intendierte Tripel eingeschlossen werden.

Wenn man sich darauf verständigt hat, daß man die kleinste Menge wählt, die die das definitorische Schema erfüllt, dann gibt es keine Probleme. Die drohende Zirkularität ist dann mit dieser Vereinbarung nunmehr eine harmlose. Diese Vereinbarung läßt sich aber nicht in der Logik erster Stufe formulieren. Erst mit der Logik zweiter Stufe, die auch Quantifikation über Teilmengen des Wertebereichs gestattet, kann der Mangel behoben werden. Man kann kraft dieser Ausdrucksstärke jetzt sogar eine explizite Definition für die Additionsfunktion geben, in der

dann die Zirkularität im Definiens verschwindet. Insofern bedeutet das Eliminierbarkeitskriterium auf den Fall der rekursiven Definitionen angewandt keine wesentliche Beschränkung. Ist axiomatisch bereits das einstellige Prädikat '$N$', dessen intendierte Extension die Menge aller natürlichen Zahlen ist, festgelegt worden und steht das Symbol '0' für die Null sowie das Symbol $'$ für die Nachfolgeoperation zur Verfügung, so läßt sich die Addition folgendermaßen definieren:

$(\forall x)(\forall y)(\forall z)[x + y = z \quad \Leftrightarrow_{df}$
$(\forall F^3)[(\forall x)(\forall y)(\forall z)(F^3xyz \ \rightarrow \ N(x) \ \& \ N(y) \ \& \ N(z)) \ \&$
$(\forall y)[F^3y0y \ \& \ (\forall x)(\forall z)(F^3xyz \ \rightarrow \ F^3x'yz')) \ \rightarrow \ F^3xyz]]$

Mit dieser expliziten Definition, die im Definiens kein Vorkommnis des Definiendums enthält, wird die Additionsfunktion als die kleinste dreistellige Relation $F^3$ festgelegt, die die in der rekursiven Darstellung angegebenen Eigenschaften hat: Zum einen sind alle Tripel der Form (y,0,y) in $F^3$ - das entspricht der Anfangsklausel y+0=y - zum anderen gilt, daß wenn $(x, y, z)$ in $F^3$ ist, dann auch $(x, y', z')$ hierin sein muß - was der Rekursionsklausel '$x+y' = (x+y)'$' entspricht.

Die rekursive Formulierung in der Logik erster Stufe kann also in der ausdrucksstärkeren Logik zweiter Stufe vollständig formuliert. Was man damit verliert, ist die Einfachheit und bessere Handhabbarkeit; was man gewinnt, ist die explizite Aufführung von ansonsten verdeckten Annahmen oder Anweisungen. Außerdem erfüllt die explizite Definition des Additionszeichens in der zweiten Stufe das (für diese Stufe) formulierte Kriterium der Eliminierbarkeit.[32]

Zum Abschluß der Überlegungen zu (STDL), (EK), (NK) und dem Zirkularitäsverbot ein kleines Resümee: Das Zirkularitätsverbot in der Form, wie es in den Regeln zu (STDL) vorkommt, ist in einem gewissen Sinne zu streng, da es Definitionen verbietet, die harmlose Zirkularitäten aufweisen. Für diese Definition gibt es aber immer einen Ersatz, so daß man nicht auf die harmlos zirkulären Definitionen angewiesen ist. Einen praktischen Nutzen von l-harmlos und T-harmlosen zirkulären Definitionen kann ich nicht erkennen, daher sind sie auch von dieser Seite nicht zu retten.

Wenn man sich auf andere Weisen verständigen kann, auf denen ein Definitionsschema die Bedeutung eines Ausdrucks festlegen soll, dann ist das Zirkularitätsverbot nicht zu rechfertigen; Sinn macht es erst dann, wenn man fest vorschreibt, daß das Gleichsetzungsverfahren anzuwenden ist. Wenn man aber andere Wege kennt, dann gewinnt man daraus in praktischer Hinsicht viel. Z.B. rekursive Definition in der Logik erster Stufe enthalten in dem Sinne fatale Zirkularitäten, daß sie von dem für die Logik erster Stufe formulierten (EK) nicht

[32] Allerdings würde vermutlich der strenge Konstruktivist noch nicht sämtliche Zirkularität aus der Welt geschafft sehen: Es wird der Durchschnitt aller Mengen gebildet, die bestimmte Eigenschaften haben. Unter denen kommt aber auch die Menge vor, die man durch die Definition als die Extension des Definiendums fixieren möchte.

erlaubt werden; sie sind aber praktisch handhabbarer als ihre Pendants in der Logik zweiter Stufe, die überhaupt keine Zirkularität mehr aufweisen, aber entsprechend komplexer sind. Die Zirkularitäten, die in den rekursiven Definitionen – formuliert in der ersten Stufe – auftraten, waren aber nie in dem Sinne fatal, daß sie verhinderten, daß für jeden Gegenstand sich eindeutig ergab, ob es in der Extension des definierten Ausdrucks lag oder nicht. Denn mit der richtigen Anweisung versehen, die man in der Logik zweiter Stufe formulieren kann, verschwindet die Zirkularität. Die verschiedenen z.B. von Yablo in [48] besprochenen Weisen, wie man ein Definitionsschema zu lesen hat, laufen alle darauf hinaus, die Anwendungsbedingungen eines Begriffes im klassischen Sinne festzulegen. Die Zirkularitäten werden durch geeignete Interpretationen des Definitionsschemas ihrer Fatalität beraubt: Trotz der Zirkularität vermögen die Definitionen die Anwendungsbedingungen eines Begriffs festzulegen.

Einen ganz anderen Weg, mit dem Zirkularitätsverbot und zirkulären Definitionen umzugehen, bieten Belnap und Gupta mit ihrer Revisionstheorie der Definition (RTD). Hier wird das Definitionsschema ebenfalls anders intepretiert, allerdings wird im Unterschied zu Yablo in [48] überhaupt keine zirkuläre Definition aufgrund der Zirkularität als logisch inkorrekt verworfen.

# 5 Zirkuläre Definitionen in (RTD)

## 5.1 Was besagen (RTW) und (RTD)?

In Aufsatz „On rigorous Definitions“ greift Belnap unter dem Titel „Relaxing the criteria (or changing the rules)“ u.a. die Revisionstheorie als eine derjenigen Theorien der Definition auf, die das Eliminierbarkeitskriterium in seiner standardmäßigen Fassung verwerfen. Hier führt er an, daß Gupta in seinem Artikel [16] folgendes festgestellt habe:[33]

1. Einige Begriffe sind zirkulär. (Normale Resultate induktiver Definitionen, z.B. Multiplikationsoperation, sind *nicht* Beispiele für zirkuläre Begriffe.)

2. Die Standarderklärung von Definitionen sagt nichts Nützliches über zirkuläre Begriffe. Belnap fügt in Klammern hinzu: „I suppose it [sc. the standard account] denies their existence“. ([5], S.145)

3. Man erhält eine mächtige Theorie zirkulärer Begriffe, indem man die Theorie der Definitionen so überarbeitet, daß auch zirkuläre Definitionen zugelassen werden.

4. Wahrheit (z.B. im Englischen) ist ein zirkulärer Begriff.

5. Mit den Ideen der umgearbeiteten Theorie der Definitionen angewandt auf Wahrheit kann man die gewöhnlichen und die ungewöhnlichen (pathologischen, paradoxen) Phänomene des Wahrheitsbegriffs erklären, für die sich die Philosophen interessiert haben.

Einige Vorüberlegungen zu dieser Auflistung scheinen mir angebracht zu sein:

- Der erste Punkt legt nahe, Gupta könne mehr als ein Beispiel für einen zirkulären Begriff anführen – was auch immer ein zirkulärer Begriff nun sein mag.[34] Tatsächlich versucht Gupta hauptsächlich nachzuweisen, daß der Wahrheitsbegriff des Englischen zirkulär ist. Die Zirkularität anderer Begriffe wird dann durch Rückführung auf den Wahrheitsbegriff nachgewiesen. Man darf nicht erwarten, Gupta würde die Zirkularität auch solcher Begriffe der natürlichen Sprachen nachzuweisen versuchen, die vom Wahrheitsprädikat genuin verschieden sind.

[33] [5], S. 141-144

[34] Auch wenn Belnap Logiker ist und in deren Munde bekanntermaßen das Wort ‘einige’ bzw. ‘some’ fast immer nur durch ‘mindestens ein’ bzw. ‘at least one’ wiederzugeben ist, wird man nicht leugnen können, daß die Wortwahl suggeriert, Gupta habe plausibel machen können, es gebe mehr als einen zirkulären Begriff.

- Es stellt sich die Frage, ob der Begriff der Zirkularität ein technischer, durch Gupta und Belnap stipulativ festgesetzter Begriff ist. Tatsächlich setzten Gupta und Belnap nirgendwo fest, was es heißen soll, daß ein Begriff zirkulär ist. Man bekommt lediglich eine hinreichende Bedingung mitgeteilt. Damit ist es unwahrscheinlich, daß wir es mit einem rein stipulativ festgesetzten Terminus zu tun haben können: Welchen Zweck sollte es haben, stipulativ nur eine hinreichende Bedingung für die Zirkularität eines Begriffes festzusetzen? Das legt die Vermutung nahe, Gupta und Belnap würden die Meinung vertreten, die durch wesentlich zirkuläre Definitionen festgelegten Begriffe wiesen eine Zirkularität auf, die unserem intuitiven Verständnis von der Zirkularität eines Begriffes sehr nahe kommen oder genau auf diese passen würden. Damit habe ich ein Problem: Wir alle scheinen ein Verständnis dafür zu haben, wann eine Erklärung oder eine Begründung oder eine Definition mit einem fatalen Zirkel behaftet ist. Haben wir jedoch ein intuitives Verständnis des Ausdrucks 'zirkulärer Begriff' oder 'Zirkularität eines Begriffes'? Es ist vermutlich zuzugestehen, daß wir die einzelnen Komponenten dieser komplexen Ausdrücke verstehen – nämlich die Ausdrücke 'zirkulär' bzw. 'Zirkularität' und 'Begriff'. Aber den Begriff der Zirkularität haben wir immer nur im Zusammenhang mit Definitionen, Begründungen oder Sätzen kennengelernt – Begriffe wurden nicht als Kandidaten für Zirkularität erwogen. Wir könnten höchstens Vergleiche mit solchen Fällen ziehen, bei denen wir tatsächlich bestimmte Intuitionen aufweisen können, und dann schließlich diese Intuitionen auf Begriffe projizieren.

  Leider problematisieren Belnap und Gupta an keiner Stelle in ihrem Buch den Ausdruck 'concept' – den ich mit 'Begriff' übersetzt habe – als daß man sich ein genaueres Bild von zirkulären Begriffen machen könnte. Einer Grundsatzdiskussion über Begriffe gehen Belnap und Gupta aus dem Weg. Würde man dem hermeneutischen Wohlwollenprinzip keine Beachtung schenken, dann könnte man Belnap und Gupta böswillig unterstellen, sie wollten suggerieren, die Sachlage bei zirkulären Begriffen sei ähnlich der in der normalwissenschaftlichen Phase der Mathematik - um in Kuhnscher Terminologie zu sprechen: Man könnte sich das durchaus realistische Bild von einem Mathematiker machen, der z.B. ein auf Topologien gemünztes einstelliges Prädikat F auf die natürlichen Zahlen überträgt, indem er es zu einem Prädikat F' modifiziert, und dann nachweist, daß einige Zahlen das Prädikat F' erfüllen. Dadurch, daß er nachweist, daß manche natürliche Zahlen die durch F' ausgedrückte Eigenschaften besitzen, hat sich nichts Problematisches für den Begriff der natürlichen Zahl ergeben. Die durch F' ausgedrückte Eigenschaft hat den Begriff der Zahl nicht strapaziert, war mit diesem kompatibel.

  Verhält sich das mit dem zirkulären Begriffen genauso? Übertragen Gupta und Belnap die Eigenschaft zirkulär zu sein von den Definitionen auf Be-

griffe und machen plausibel, daß einige Begriffe zirkulär sind, ohne dabei den Begriff des Begriffes mit dieser Übertragung zu strapazieren?

- Wie hat man sich die unter dem ersten Punkt angesprochene Plausibilisierung oder gar den Nachweis vorzustellen, daß es zirkuläre Begriffe gibt? Man könnte aufgrund dieser Auflistung, welche unter 1. den Sachverhalt nennt, daß einige Begriffe wesentlich zirkulär sind, und erst danach unter 3., daß man mit (RTD) eine mächtige Theorie zirkulärer Definitionen erhält, zur Meinung gelangen, Gupta könne in seinem Artikel – und dann natürlich auch Gupta und Belnap in ihrem Buch RTT – einen guten Grund für die Existenz zirkulärer Begriffe angeben, ohne dabei auf den ganzen durch (RTD) zur Verfügung gestellten Apparat zirkulärer Definitionen zurückgreifen zu müssen. Tatsache ist, daß von Gupta und auch von Gupta und Belnap in RTT keine derartige unabhängige Plausibilisierung der Existenz von zirkulären Begriffen unternommen wird: Die Plausibilisierung besteht gerade in der Entwicklung einer Theorie zirkulärer Definitionen, die zirkuläre Begriffe abwirft. Wie man sich das genau vorzustellen hat, werde ich im Laufe der Arbeit verständlich zu machen versuchen.

- Auch in dem zweiten Punkte scheint sich mir die Überzeugung Belnaps zu zeigen, der von ihm und Gupta in der (RTD) aufgebrachte zirkuläre Begriff habe viel mit der Zirkularität, wie wir sie uns üblicherweise vorstellen würden, zu tun. Interessant ist Belnaps Klammerzusatz, die seine eigene – vielleicht auch Guptas – Meinung wiedergibt: Leugnet die Standardtheorie der Definition tatsächlich die Existenz zirkulärer Begriffe? Leider begründet Belnap seine Vermutung an dieser Stelle nicht, und in RTT wird diese Vermutung, soweit ich sehen kann, nicht einmal aufgeführt. Die Frage, ob die Standarderklärung von Definitionen die Existenz von zirkulären Begriffen leugnet, ist unabhängig von der Frage, wie sich bestimmte *Vertreter* der Standarderklärung zur Existenzfrage zirkulärer Begriffe geäußert haben. Eine Theorie nicht-zirkulärer Definitionen scheint auf den ersten Blick hin nicht inkompatibel mit der Existenz zirkulärer Begriffe zu sein. Es stellt sich die natürliche Frage, ob zirkuläre Begriffe, so es sie gibt, nicht auch anders als über den Weg zirkulärer Definitionen charakterisiert werden können. Mit der Theorie zirkulärer Definitionen legt man sich auf die Existenz von Gebilden fest, die Begriffen ähnlich zu sein scheinen und eine Zirkularitätseigenschaft besitzen. Mit der Standarderklärung folgt eine derartige ontologische Festlegung nicht; aber das heißt nicht, daß man sich mit der Standardtheorie der Definition auf das Gegenteil festlegt. Die Standardtheorie spricht überhaupt nicht von zirkulären Gebilden, also könnte es doch sein, daß sie die Existenzfrage nicht entscheidet: die Standardtheorie der Definition ist keine syntaktisch vollständige Theorie, als daß man dieses ausschließen könnte. Damit könnte nur Belnaps These bestätigt werden,

daß die Standardtheorie nichts Nützliches über zirkuläre Begriffe sagt, nicht aber daß sie deren Existenz leugnet. Tatsächlich ist aber nicht zu sehen, wie sich Ausdrücke klassisch definieren lassen sollten, die zirkuläre Begriffe ausdrücken sollen: Wir lernen zirkuläre Begriffe auf dem von Gupta und Belnap vorgezeichneten Wege über wesentlich zirkuläre Definitionen kennen; zirkuläre Definitionen sind aber in der klassischen Definitionstheorie verboten.[35]

In den folgenden Abschnitten werden über die schon genannten Dinge hinaus die folgenden Tatsachen über (RTD) dargestellt werden:

- (RTD) akzeptiert das Nicht-Kreativitätskriterium der Standardtheorie der Definition, verwirft aber das Eliminierbarkeitskriterium und insbesondere das Verbot der Zirkularität.
- (RTD) ändert die (logische) Grammatik zirkulärer Definitionen. Definitionen werden nicht mehr als Sätze derjenigen Sprache angesehen, für die sie aufgestellt werden.
- (RTD) macht einen Unterschied zwischen materialen Konditionalen und Definitionen: Für Definitionen gelten nicht die logischen Regeln, die für das materiale Konditional gelten.
- (RTD) wartet mit einer eigenen Semantik für zirkuläre Definitionen auf, die durch entsprechende beweistechnische Methoden ergänzt wird.

## 5.2 Ein erster Vergleich: Wahrheit und zirkuläre Definitionen

Die folgenden Abschnitte orientieren sich zum größten Teil an Kap. 4, S. 113-143 von RTT. Dieses Kapitel wie auch wesentliche Teile des siebten Kapitels entstammen im großen und ganzen Guptas Artikel „Remarks on Definitions and the concept of Truth“ ([16]).

In diesem einführenden Abschnitt wird streng der Darstellung von Gupta und Belnap folgend die Einführung von zirkulären Definitionen motiviert. Beim Vergleich des paradoxen Lügnersatzes mit Sätzen, die ein zirkulär definiertes einstelliges Prädikat enthalten, zeigt sich, daß beide sehr ähnliche Probleme mit sich führen. Diese Ähnlichkeit legt die Vermutung nahe, daß hier eine engere Verwandschaft zugrunde liegt, welche aufgedeckt eine neue, interessante Sichtweise auf die Lügnerparadoxie und den Begriff der Wahrheit gewähren könnte.

[35] Selbst wenn sich die Standardtheorie auf die Nicht-Existenz zirkulärer Definitionen festlegen sollte, bleibt weiterhin die Frage, ob der Zugang über zirkuläre Definitionen der einzige Weg ist. So versucht Martin in seiner Kritik in [31] zu argumentieren, daß es alternative Theorien für zirkuläre Begriffe geben könnte. Gupta in seiner Antwort ([17]) versucht nachzuweisen, daß diese alternativen Theorien gewisse wünschenswerte Eigenschaften nicht besitzen.

Zunächst will ich eine einfache Version der Lügnerparadoxie[36] vorstellen, die man sicherlich noch sauberer formulieren müßte.[37]

### 5.2.1 Der Lügner und der Wahrsager

Nach unseren umgangssprachlichen Intuitionen zum Wahrheitsbegriff werden wir vermutlich das folgende Zitattilgungsprinzip (ZTP) akzeptieren:

**(ZTP)** 'p' ist wahr $\leftrightarrow$ p

Der Buchstabe 'p' ist dabei ein Platzhalter für einen Satz der deutschen Sprache. Man betrachte den folgenden Satz, der zu einer Gruppe von Sätzen mit dem Titel „The Liar"[38] zu zählen ist:

(1) Der Satz in Zeile (1) ist nicht wahr.

Wendet man (ZTP) auf diesen Satz an, wird also p durch den Satz 'Der Satz in Zeile (1) ist nicht wahr' ersetzt, dann erhält man:

(2) 'Der Satz in Zeile (1) ist nicht wahr' ist wahr $\leftrightarrow$ Der Satz in Zeile (1) ist nicht wahr.

Ein Blick auf die Zeile (1) verrät, welcher Satz mit der Kennzeichnung 'Der Satz in Zeile (1)' denotiert wird:

(3) Der Satz in Zeile (1) = 'Der Satz in Zeile (1) ist nicht wahr'

Wenden wir auf (2) und (3) die Regel der Identitätsbeseitigung an, ersetzen wir also das zweite Vorkommnis des Ausdrucks 'Der Satz in Zeile (1)' in der Zeile (2) durch den Ausdruck 'Der Satz in Zeile (1) ist nicht wahr', so erhalten wir folgenden Satz:

(4) 'Der Satz in Zeile (1) ist nicht wahr' ist wahr $\leftrightarrow$ 'Der Satz in Zeile (1) ist nicht wahr' ist nicht wahr.

Im letzten Schritt der Ableitung erhält man abhängig vom Zitattilgungsprinzip einen Satz der aussagenlogischen Form $\ulcorner A \leftrightarrow \neg A \urcorner$, welches bekanntermaßen

[36] Die übliche Bemerkung zur Namensgebung dieser Paradoxie: Worauf es in dieser Paradoxie ankommt ist, daß ein Satz von sich selber sagt, es sei falsch. Eine Lüge ist aber nicht schon dann eine Lüge, wenn sie ein falscher Satz ist. Hier spielen weitere Komponenten eine Rolle - intentionale Aspekte etc.

[37] Belnap und Gupta geben eine ganz ähnliche Ableitung, allerdings nicht im 4., sondern im 1. Kapitel ([18], S.5).

[38] Teilweise benutzt man diese Bezeichnung auch für das Paradox.

ein nicht erfüllbares Satzschema ist. Folglich haben wir aus (ZTP) einen formal-analytisch falschen Satz abgeleitet. Wenn die angewandten Regeln korrekt[39] waren, dann stehen wir vor dem Problem, das intuitiv doch recht einleuchtende Zitattilgungsprinzip verwerfen zu müssen.

Die oben gegebene Ableitung ist eine recht kurze Version der Lügnerparadoxie, die zwar aufs genauste auf die Problematik aufmerksam zu machen vermag, aber vermulich nicht genau den Gedankengang wiederspiegelt, den jemand hat, der aufgefordert wird, den Wahrheitswert des Lügnersatzes zu bestimmen. Die meisten Personen, die im Kopf die Lügnerparadoxie durchspielen, scheinen in eine nicht enden wollende Schleife zu geraten, wenn sie den Wahrheitswert des Satzes in (1) zu bestimmen versuchen. Vemutlich wird man statt des Prinzips (ZTP) dabei eher mehr oder minder bewußt die beiden Regeln (WB) (Regel der **W**ahrheits**b**eseitigung) und (WE) (Regel der **W**ahrheits**e**inführung) verwenden:

$$\text{(WB)}\ \frac{\text{`A' ist wahr}}{\text{A}}$$

$$\text{(WE)}\ \frac{\text{A}}{\text{`A' ist wahr}}$$

(WB) stellt dabei eine „Richtung“ von (ZTP) dar, (WE) die andere.
Faßt man mit Gupta und Belnap die Regeln (WE) und (WB) als Prozeduren zur Beantwortung der Frage auf, ob ein Satz wahr ist oder nicht, dann gelangt man zu einer anderen Darstellung der Problematik mit dem Wahrheitsbegriff. Wenn jemand aufgefordert wird, den semantischen Status von ‘A ist wahr’ zu bestimmen, dann wird er versuchen, den semantischen Status von A zu bestimmen. Um den Status von ‘A ist nicht wahr’ zu bestimmen, wird man auf nicht-A[40] verwiesen. Mit dem für diese Zwecke eingeführten Kürzel

x
↓
y

[39]Man könnte durchaus auch die Meinung vertreten, daß eine der bei der Ableitung angewandten Regeln nicht korrekt ist. Die auf den ersten Blick angreifbar erscheinende Regel aus der Ableitung ist die Regel der Identitätsbeseitigung. Obwohl mir kein Lösungsversuch der Lügnerparadoxie bekannt ist, bei dem wesentlich mit einer Kritik an der Anwendung von $= B$ argumentiert wird, erscheint mir doch zumindest der Verdacht, $= B$ sei für die Paradoxie verantwortlich, nicht abwegig. Die Problematik mit der Anwendung von $= B$ auf natürliche Sprachen sind hinreichend bekannt. Sie resultieren aber – glaube ich – aus einer unvorsichtigen, inkorrekten *Anwendung* von $= B$, nicht daraus, daß $= B$ selbst inkorrekt wäre. Also bliebe für den Kritiker nur noch die Möglichkeit, eine inkorrekte Anwendung von $= B$ zu beanstanden. Dieses müßte dann natürlich nachgewiesen werden. Eine andere mögliche Diagnose für die Paradoxie ist natürlich, die Probleme auf semantische Prinzipien der klassischen Logik zurückzuführen.

[40]Gupta und Belnap verwenden den Ausdruck ‘non-A’ sicherlich, um ‘It is not the case that A’ abzukürzen ([18], S.115).

für ‘Zur Bestimmung des semantischen Status von x bestimme den semantischen Status von y’ lassen sich diese Sachverhalte so formulieren:

$(R1_W)$

‘A’ ist wahr
↓
A

$(R2_W)$

‘A’ ist nicht wahr
↓
nicht-A

Setzt man nun speziell für A den Lügnersatz (1) ein, so gerät man in eine endlose Schleife, die einen zur Bestimmung des Status von A jeweils nach zwei Schritten wieder auf die Bestimmung des semantischen Status von A führt:

$(E1_W)$

‘A’ ist wahr
↓
‘A’ ist nicht wahr
↓
‘A’ ist wahr
⋮

Die Problematik, die der Lügnersatz für den Wahrheitsbegriff aufbringt, stellt sich hier also dar als die Unmöglichkeit, den semantischen Status dieses Satzes bzw. des Satzes, welcher dessen Wahrheit behauptet, zu bestimmen.

Mit der obigen Darstellung läßt sich auch die Problematik veranschaulichen, die andere Sätze für den Wahrheitsbegriff darstellen. Für den Satz (5) etwa, der auch „The Truth-Teller“ genannt werden kann

(5) Der Satz in Zeile (5) ist wahr

ergibt sich eine ähnliche Schleife, in der zur Bestimmung des semantischen Status von A direkt auf die Bestimmung des semantischen Status von A verwiesen wird:

$(E2_W)$

‘A’ ist wahr
↓
‘A’ ist wahr
↓
‘A’ ist wahr
⋮

### 5.2.2 Vergleich mit zirkulären Definitionen

Betrachten wir nun ein Beispiel für eine zirkuläre Definition eines einstelligen Prädikatbuchstabens G:

(*) $Gx \quad =_{Df} \quad Fx \vee (Hx \,\&\, \neg Gx)$

Im Definiens für G kommt neben den einstelligen Prädikatbuchstaben H und F auch wieder G vor. In der klassischen Definitionstheorie gelten für eine Definition

$G(x) \quad =_{Df} \quad$ —$x$—

wobei '—$x$—' ein in x offenes Satzschema ist, folgende Regeln der **Def**iniendum-**B**eseitigung und **Def**iniendum -**E**inführung:

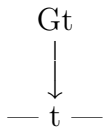
**(DfB)** Ist im Laufe einer Ableitung (etwa in einem Kalkül der Logik erster Stufe) 'G(x)' abgeleitet worden, so darf man zum Definiens '—$x$—' übergehen.

**(DfE)** Ist im Laufe einer Ableitung '—$x$—' abgeleitet worden, so darf man zum Definiendum 'G(x)' übergehen.

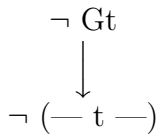
Denkt man sich in der Definition (*) das Zeichen '$=_{Df}$' durch das materiale Bikonditional '$\leftrightarrow$' ersetzt, so erscheinen die Regeln recht plausibel. In der klassischen Theorie der Definition gehorcht das Definitionszeichen genau denselben Ableitungsregeln wie das materiale Bikonditional. Mit einer derartigen Definition gerät man leicht in Paradoxien, die der Lügnerparadoxie nahekommen. Nehmen wir an, **a** ist ein durch 'a' denotierter Gegenstand, der zwar in der Extension von H, nicht aber in der von F vorkommt. Ist nun 'Ga' wahr oder falsch? Angenommen es ist wahr, dann nach (DfB) auch die a-Instanz des Definiens, also: '$Fa \vee (Ha \,\&\, \neg Ga)$' ist wahr. Da nach Voraussetzung (über die Wahl von F und H bzw. des Gegenstandes **a**) '$\neg Fa$' wahr ist, muß – wie etwa die Anwendung der klassischen Regel MTP zeigt – das zweite Disjunktionsglied wahr sein, also: '$Ha \,\&\, \neg Ga$' ist wahr. Damit ist aber auch '$\neg Ga$' wahr, was der Annahme, daß 'Ga' wahr ist, widerspricht.
Also muß wohl 'Ga' falsch sein. Aber auch das führt zu einem Widerspruch; denn damit wäre '$\neg Ga$' wahr und damit, da nach Voraussetzung 'Ha' wahr ist, auch das zweite Disjunktionsglied des Definiens, also '$Ha \,\&\, \neg Ga$'. Damit ist aber das gesamte Definiens wahr. Die Anwendung von (DfE) gestattet, zum Definiendum überzugehen und also zu behaupten, daß 'Ga' wahr ist. Wieder liegt ein Widerspruch vor.
Die beiden Regeln (DfB) und (DfE) entsprechen jeweils den oben aufgeführten Regeln (WB) und (WE). Auch sie führen im Zusammenhang mit bestimmten außergewöhnlichen Sätzen – hier also Definitionen – zu einem Paradox, und auch bei diesem Paradox stellt sich die Frage, ob nun eine der beiden Regeln oder der ungewöhnliche Satz – die ungewöhnliche Definition – in der Hauptsache für den abgeleiteten Widerspruch verantwortlich zu machen ist.

Die Ähnlichkeit läßt sich weiter verfolgen: Auch im Falle der Regeln (DfB) und (DfE) scheint es möglich, eine Verknüpfung mit bestimmten Anweisungen herzustellen. So kann man aus (DfB) und (DfE) die folgenden Anweisungen für den Fall, daß ein einstelliges Prädikat G mit dem Definiens '— t — ' definiert wird, herauslesen:

$(R1_D)$

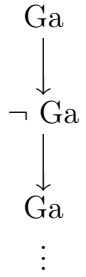



$(R2_D)$




Wenn wir nun für die Definition (*) aufgrund dieser Anweisungen zu bestimmen suchen, ob **a** zur Extension von G gehört, dann gerät man in eine unendliche Schleife:

$(E1_D)$




Für das folgendermaßen definierte Prädikat G

(**) $Gx \quad =_{Df} \quad Fx \vee (Hx \,\&\, Gx)$

gerät man für den Satz 'Ga' in eine Schleife ähnlich der für den Satz 'Dieser Satz ist wahr' in $(E2_W)$:

$(E2_D)$


Die Ähnlichkeiten des Wahrheitsprädikats und bestimmten zirkulären Definitionen ist aber nicht bloß auf „pathologische" Fälle – wie Gupta und Belnap es nennen – beschränkt. So wie sich beim Wahrheitsprädikat für viele Sätze, wie

etwa den berühmten Satz 'Schnee ist weiß', keine Widersprüche bzw. unendliche Anweisungsschleifen ergeben, so ist auch bei vielen zirkulären Definitionen eines einstelligen Prädikats G die Frage, ob ein Gegenstand b zur Extension von G gehört, ohne Widersprüche und ohne daß man in unendliche Schleifen gerät zu beantworten.

Diese Überlegungen motivieren Gupta und Belnap zu folgendem Urteil:

> Concepts with circular definitions, then, behave in ways that are remarkably similar to the behavior of the concept of truth. And like truth, they can be, and usually are, unproblematic over a range of cases. These similarities suggest, first, that the outright rejection of circular definitions in logic may be to precipitous, for their behavior is very much like that of a concept the use of which we *do* accept and *want* to accept.
>
> [...] Second, the similarities suggest that the perplexing behavior of the concept of truth might be explainable as arising from some circularity in its definition. ([18], S. 117)

## 5.3 Bedeutung zirkulärer Definitionen und $S_0$

Eine Theorie zirkulärer Definitionen sollte nach Gupta und Belnap zumindest folgendes zu spezifizieren versuchen, um dem oben dargestellten pathologischen Verhalten und dem Inhalt zirkulärer Definitionen gerecht werden zu können:

1. Die Bedeutung („meaning"), die Definitionen – zirkuläre eingeschlossen – ihren jeweiligen Definienda zuschreiben.

2. Logische Regeln für Definitionen.

Die beiden Spezifikationspunkte scheinen im Falle, daß man unter 'Definitionen' im großen und ganzen Sätze oder Sprechakte zur Angabe der Bedeutung sprachlicher Ausdrücke versteht, angemessene Ziele zu sein. Die hierbei gemachte Einschränkung von 'Definition' auf 'Angabe der Bedeutung' ist nicht redundant, da sich eine Theorie zirkulärer Definition, die 'Definition' Aristoteles gemäß als 'Wesensbestimmung' auffaßt, nicht in erster Linie um die Spezifikation der Bedeutung von sprachlichen Ausdrücken zu bemühen hätte; bzw. eine solche Theorie hätte es zumindest so zu gestalten, daß sich daraus etwas zur Wesensbestimmung des zu „definierenden" Gegenstandes ableiten ließe.

Was nun sind Definitionen oder welchen Zweck haben sie nach Gupta und Belnap? Die einzige wesentliche Behauptung, die Gupta und Belnap hierzu machen, ist die folgende:

> Let us accept the natural idea that a definition fixes completely the meaning of its definiendum. ([18], S.118)

Definitionen fixierten also vollständig die Bedeutung („meaning“) ihres Definiendums, d.h. des Ausdrucks, der definiert werden soll. Einer derartigen Ansicht würde auch Frege zustimmen: Wiewohl Definitionen nach Frege mehr zu leisten haben, so folgt doch aus dem „mehr“ – Fixierung des Sinns – auch die Forderung nach der vollständigen Fixierung der Bedeutung.[41]

Auch wenn man sich nicht in allen Punkten über die richtige Explikation von ‘Definition’ und/oder ‘Bedeutung’ bzw. ‘meaning’ einig sein sollten, so dürfte man dennoch mit Fug und Recht behaupten dürfen, daß es eine nicht abwegige Auslegung dieser Ausdrücke gibt, die die Aussage wahr macht, daß Definitionen die Bedeutung ihrer Definienda vollständig fixierten. Zumindest können wir das für Definitionen behaupten, die nicht echt zirkulär sind.

Im Falle von nichtzirkulären Definitionen, so Gupta und Belnap, könne man die Frage danach, wie man sich die Bedeutung vorzustellen habe, mit Verweis auf die traditionelle Erklärung beantworten:

> The meaning of a predicate[42] by this account, is a rule that gives the extension of the predicate in all possible situations. Or, equivalently, the meaning determines the conditions for the predicate's applicability. For example, the meaning of ‘red’, by the traditional account, is a rule that determines the conditions under which an object counts as red. ([18], S.118)

Ich kann nicht entscheiden, ob es sich nur um laxe Ausdrucksweise oder eine ernst gemeinte These handelt, wenn Gupta und Belnap im ersten Satz des Zitats behaupten, nach der traditionellen Erklärung[43] wäre die Bedeutung eines Prädikats nichts anderes als eine *Regel*. Über den Regelbegriff lassen sich Gupta und Belnap nicht aus, ebensowenig will ich mich auf eine Diskussion des Regelbegriffs einlassen. Wichtig scheint mir nur, für einige wichtige Fälle von Ausdrücken entscheiden zu können, ob sie gemäß dieser Lesart dieselbe Bedeutung haben. Haben z.B. die Prädikate ‘ist ein gleichwinkliges Dreieck’ und ‘ist ein gleichseitiges Dreieck’ gemäß der obigen Lesart dieselbe Bedeutung? Nach unserem intuitiven Verständnis würden wir sagen, daß diese Ausdrücke nicht denselben Sinn haben: Im ersten Ausdruck spielt die Gleichseitigkeit eines Dreiecks eine Rolle, im letzteren hingegen nicht. Gleichwohl ist es notwendigerweise der Fall, daß ein Dreieck genau dann gleichseitig ist, wenn es gleichwinklig ist. Letztere Tatsache würde uns vermutlich darin rechtfertigen zu sagen, daß die Regel, welche die Anwendungsbedingungen für das Prädikat ‘ist ein gleichseitiges Dreieck’ angibt, und diejenige Regel, welche die Anwendungsbedingungen für das Prädikat ‘ist ein

[41] „Bedeutung“ wird hier nichtfregisch gelesen.

[42] Die (RTD) wird für Definitionen von Prädikaten entwickelt. Guptas und Belnaps trockener Versicherung, ähnliche Ergebnisse würden sich auch im Falle sprachlicher Ausdrücke aus anderen logisch-grammatischen Kategorien – wie z.B. der Kategorie der Funktionsausdrücke oder der Namen – ergeben, will ich getrost Glauben schenken.

[43] Gupta und Belnap denken vermutlich an die sogenannte „Kalifornische Semantik“. Eine solche Erklärung gibt z.B. Rudolf Carnap in seinem Buch „Meaning and Necessity“ ([9]).

gleichwinkliges Dreieck' angibt, identisch sind. Folglich wären die Bedeutungen dieser Ausdrücke gemäß der obigen Erläuterung dieselben. Daraus würde folgen, daß 'meaning' hier nicht im Sinne von 'Sinn' gebraucht wird.

Gupta und Belnap suggerieren, daß durch den zweiten Satz des Zitats eine äquivalente Formulierung der traditionellen Erklärung der Bedeutung eines Prädikats gegeben sei. In dem heißt es aber nur, daß die Bedeutung eines Prädikats die Anwendungsbedingungen des Prädikats festlegt – davon daß die Bedeutung eine Regel ist, die die Anwendungsbedingungen eines Prädikats festlegt, ist nicht die Rede. Ich vermute aber, daß hier einfach unvorsichtig formuliert wurde; das wird auch durch das Beispiel nahegelegt, in dem es wieder heißt, daß die Bedeutung des Prädikats 'ist rot' eine *Regel* sei, die die Bedingungen festlege, unter denen ein Gegenstand als rot gelte.

Die erste Formulierung enthält die nicht sehr eindeutige Rede von „allen möglichen Situationen". Es scheint damit eine bloß stilistische Variante der nicht weniger problematischen, aber zumindest technisch besser traktierbaren Rede von "möglichen Welten" gemeint zu sein. Das wird durch Ausführungen an anderer Stelle des Buches nahegelegt. Im ersten Kapitel besprechen Gupta und Belnap die sogenannten T-Bikonditionale (s.u.) und führen dort an, daß sie die Intuition aufrechterhalten möchten, die T-Bikonditionale fixierten vollständig die *Bedeutung* der Wahrheit. Zur Klarstellung dieser Intuition fügen sie an:

> Another necessary qualification is that the relevant sense of 'meaning' here is that of *intension*. ([18], S.25)

In der hierzu gehörigen Fußnote 44 merken sie zum Begriff der Intension an:

> We use *intension* to isolate that aspect of meaning that concerns alethic modalities such as possibility. Roughly, the intension of an expression is determined by its extension through all possible worlds. ([18], S.25, Fn.44)

Das Wort 'meaning', so legt es das erstere dieser beiden Zitate nahe, heißt in dem dort aufgeführten Kontext nichts anderes als 'Intension'. Die Intension eines Ausdrucks wiederum, so erfährt man aus dem zweiten Zitat, wird durch die Extension des Ausdrucks in allen möglichen Welten festgelegt. Im Falle eines einstelligen klassischen Prädikats kann man sich folglich dessen Intension als eine Funktion denken, die zu jeder möglichen Welt eine Menge von Gegenständen zuordnet, welche man als Extension des Prädikats ansehen kann.[44]

Da an der oben zitierten Stelle 'meaning' im Sinne von 'Intension' gebraucht wird, ist es nicht abwegig anzunehmen, Gupta und Belnap würden 'meaning' an anderen Stellen ebenfalls so gebrauchen. Da Gupta und Belnap nicht angeben, wie genau eine mögliche anspruchsvollere Lesart von 'meaning' auszusehen hätte,

[44]Der Ausdruck 'Intension' ist ambig. Neben der von Gupta und Belnap zugrunde gelegten Verwendung kennt man in der Philosophie auch einen stärkeren Gebrauch von 'Intension', der in Richtung '(sprachlicher) Sinn' geht.

und da sie im Zusammenhang mit dem Wahrheitsbegriff 'meaning' im Sinne von 'intension' interpretiert wissen wollen, werde auch ich diese Lesart zugrunde legen.

Was diese Interpretation ein wenig stört, ist die im zweiten Zitat aufgenommene Anmerkung der Fußnote, welche davon spricht, daß der Begriff der Intension von den Autoren benutzt wird, um einen bestimmten *Aspekt* der Bedeutung zu isolieren. Dann scheinen aber Gupta und Belnap durchaus auch eine andere, umfassendere, nicht auf einen Aspekt reduzierte Verwendungsweise des Wortes 'meaning' in Betracht zu ziehen. Ansonsten müßte man man davon ausgehen, daß wenn man den Aspekt eines Dinges isoliert, man dieses Ding selbst erhält – was doch sehr merkwürdig klingt.[45]

Man sollte sich insbesondere hier schon klar machen, daß wenn Gupta und Belnap die These aufstellen werden, die T-Bikonditionale[46] würden den Wahrheitsbegriff *definieren*, damit nicht gemeint ist, es werde der sprachliche Sinn des Prädikats 'ist wahr' festgesetzt, sondern lediglich dessen Extension[47] in allen möglichen Welten.

Wie geschieht es nun, daß Definitionen die Bedeutung (Intension) des Definiendums fixieren? Gupta und Belnap schreiben:

> Now, a noncircular definition enables us to calculate the extension of the definiendum once we are given the extensions of the terms in the definiens. Hence, by the traditional account of meaning, it explains the meaning of the definiendum on the basis of the meanings of the terms in the definiens. ([18], S.118)

In nichtzirkulären[48] Definitionen wird die Bedeutung des Definiendums ganz einfach auf der Grundlage der Bedeutungen der im Definiens enthaltenen Ausdrücke

[45]Ein weiterer Hinweis darauf, daß Gupta und Belnap den Ausdruck 'meaning' nicht immer im Sinne von 'Intension' – so wie dieser Ausdruck in der zitierten Fußnote erläutert wurde – verwendet wissen wollen, findet sich an einer ganz anderen Stelle im Buch: Im dritten Kapitel von RTT, der eine ausführliche Kritik der Fixpunkttheorie der Wahrheit enthält, treffen Gupta und Belnap eine Unterscheidung zwischen einem T-Prädikat für eine Sprache **L** und einem Wahrheitsprädikat für eine Sprache **L**: „From now on it will prove useful to say that a predicate is a *truth predicate* for a language **L** iff it *means* „true in **L**“ . In contrast, a predicate is a T-predicate for **L** iff it is coextensive with truth for **L**.“([18], S.86) An diese Zeilen schließt sich der für mich interessante Gebrauch des Wortes 'intentional' an: „ 'Truth predicate' is an intentional notion:...“ ([18], S.86) Würden Gupta und Belnap 'meaning' lediglich als 'intension' lesen wollen, dann wäre es angebrachter gewesen zu sagen, daß 'Truth predicate' ein intens**i**onaler Begriff. Daß hier kein Schreibfehler vorliegt, wird durch eine weitere Verwendung des Wortes 'intentional' in der Fußnote 4 der Seite 86 bestätigt.

[46]T-Bikonditionale sind Sätze der Form ⌜ S ist wahr genau dann, wenn p ⌝, wobei 'S' Platzhalter für eine Standardbezeichnung eines Satzes und 'P' Platzhalter für einen bedeutungsgleichen Satz ist. Tarskis W-Äquivalenzen setzen eine Unterscheidung von Objekt- und Metasprache voraus, die mit den T-Bikonditionalen nicht gemacht ist. Diese Verwendung des Ausdrucks 'T-Bikonditional' scheinen mir Gupta und Belnap zugrunde zu legen, auch wenn sie anscheinend meinen, daß die T-Bikonditionale gerade die Tarskischen W-Äquivalenzen sind.

[47]Genauer muß man statt „Extension“ „Signifikation“ sagen. Dieser Begriff wird später erläutert werden.

[48]Nichtzirkulär heißt in diesem Zusammenhang eine Definition, die kein Vorkommnis des

bestimmt. Es lassen sich die Anwendungsbedingungen des zu definierenden Prädikats auf der Basis des Definiens durch Gleichsetzen angeben; genauer: in einer stipulativen nichtzirkulären Definition eines Prädikats F werden die Anwendungsbedingungen von F dadurch eindeutig festgelegt, daß sie als die Anwendungsbedingungen des Definiens angegeben werden. Ich will dieses Verfahren für meine Zwecke Gleichsetzungsverfahren nennen. Die Anwendungsbedingungen des Definiens wiederum lassen sich bestimmen, wenn die Anwendungsbedingungen der im Definiens vorkommenden Ausdrücke bekannt sind.

Für zirkuläre Definitionen ergibt sich nach Gupta und Belnap aber ein Problem:

> There is a problem with the traditional account, however, if we want to preserve the idea that a circular definition *also* fixes the meaning of its definiendum. A circular definition does not in general enable us to determine the extension of its definiendum. Like all definitions, it does provide a rule for determining this extension, once the extensions of *all* the terms in the definiens are given. The problem is that, as the definiendum occurs in the definiens, to apply this rule we need to know the very thing we are trying to determine, namely the extension of the definiendum. To capture the meaning that a circular definition ascribes to its definiendum, we need to think of meaning in a different way. ([18], S.118-119)

Für zirkuläre Definitionen gerieten also die Überlegung, Definitionen würden vollständig die Bedeutung des Definiendums fixieren, auf der einen Seite und die traditionelle Explikation der Bedeutung als einer Regel, die zu jeder möglichen Welt die Extension des zu definierenden Ausdrucks angibt, auf der anderen Seite miteinander in Konflikt. Das Problem bestehe darin, daß man, um die Extension des Definiendums bestimmen zu können, diese bereits kennen müßte. Mit dieser Problematik wird man nicht bei allen zirkulären Definitionen konfrontiert. Wenn das Vorkommnis des Definiendums l-harmlos ist, läßt sich das Gleichsetzungsverfahren ebenfalls anwenden: Für die Definition

> 'x ist ein Junggeselle $\Leftrightarrow_{Df}$ x ist ein lediger Mann im heiratsfähigen Alter und wenn x eine Junggeselle ist, dann ist x ein Junggeselle'

braucht man nur die Anwendungsbedingungen von 'lediger Mann im heiratsfähigen Alter' zu kennen, ohne daß die Anwendungsbedingungen von 'ist ein Junggeselle' „vorher" gegeben sein müßten. Solchen Fällen tragen Gupta und Belnap Rechnung, indem sie vorsichtig formulieren, daß man bei zirkulären Definitionen *im Allgemeinen* – aber eben durchaus nicht immer – die Extension „vorher" kennen muß, um das Gleichsetzungsverfahren anzuwenden.

---

Definiendums im Definiens enthält. Wie wir gleich sehen werden, ist das im Haupttext für nichtzirkuläre Definitionen beschriebene Gleichsetzungsverfahren für die Festlegung der Bedeutung des Definiendums auch auf solche Definitionen anwendbar, die l-harmlose Vorkommnisse des Definiendums im Definiens enthalten.

Gupta und Belnap möchten an der Idee festhalten, Definitionen würden eindeutig die Bedeutung des Definiendums fixieren – das soll also auch für zirkuläre Definitionen gelten. Den Schritt, an dieser Idee festzuhalten, können Gupta und Belnap m.E. erst dadurch rechtfertigen, daß sich hierdurch eine schöne Theorie zirkulärer Definitionen ergibt. Da sie den klassischen Begriff der Bedeutung verwerfen, das Wort 'Bedeutung' im folgenden dann auch anders verwenden werden, können sie nicht an etwaige sprachliche Intuitionen appellieren – was sie de facto auch nicht tun. Folglich glaube ich, daß man das Festhalten an dieser Idee als ein willkürliches verstehen muß.

Aus der Vorgabe, an der Idee festzuhalten, Definitionen fixierten vollständig die Bedeutung des Definiendums, folgern Gupta und Belnap, daß man den Begriff der Bedeutung für zirkuläre Definitionen modifizieren muß. Natürlich folgt aus dieser Vorgabe nicht im strengen Sinne gleich die Notwendigkeit einer Modifikation des Bedeutungsbegriffs. So kann man die Problematik etwa auf das Gleichsetzungsverfahren zurückführen und sich andere Methoden ersinnen, mit denen durch eine Definition die Bedeutung des Definiendums festgelegt werden soll. Das Gleichsetzungsverfahren geht davon aus, daß die Extension des Definiens vollständig gegeben ist, und setzt die Extension des Definiendums dadurch fest, daß es die des Definiens sein soll. In anderen Verfahren kann man sich überlegen, wie man mit Hilfe einer Definition die Extension des Definiendums festlegen kann, ohne davon auszugehen, daß die Extension des Definiens vorher gegeben ist. So könnte man z.B. die Überlegung haben, daß eine Definition als Gleichung aufgefaßt werden könnte: In einer Welt w soll die Extension des Definiendums G, welche eine Menge $\mathbf{G}$ ist, als Lösung einer Gleichung

$$\mathbf{G} = A_w(\mathbf{G})$$

verstanden werden, wobei mit '$A_w(\mathbf{G})$' die Extension des Definiens '$A(G)$' bezeichnet wird. Solch ein Verfahren aber geht davon aus, daß es eine eindeutige Lösung der Gleichung gibt, was durchaus nicht immer der Fall ist. Für induktive Definitionen z.B. muß man zu diesem Verfahren noch die Anweisung hinzunehmen, daß die kleinste Lösung der Gleichung diejenige sein soll, welche als die Bedeutung des Definiendums festgesetzt wird. Für manche zirkuläre Definition gibt es aber auch keine kleinste Lösung für die Gleichung und man muß sich ein anderes Verfahren ersinnen, um noch mehr zirkuläre Definitionen zu akzeptieren. Wie das gehen könnte, macht z.B. Stephen Yablo in [48] vor. Aber selbst bei Stephen Yablo werden nicht alle zirkulären Definitionen akzeptiert: Es bleiben immer einige zirkuläre Definitionen übrig, die schlichtweg nicht „lösbar" sind in dem Sinne, daß nicht in jeder möglichen Welt für das zu definierende Prädikat eindeutig eine Extension festgelegt wird. Nun könnte man sich überlegen, wie man die restlichen, von Yablos Überlegungen nicht erfaßten zirkulären Definitionen auslegt, so daß schließlich durch alle zirkulären Definitionen eine eindeutige Extension (in jeder möglichen Welt) für das Definiendum festgelegt wird. Da es durchaus möglich ist, diesen Weg zu gehen, indem man z.B. auf den Überlegungen

von Yablo aufbaut, ergibt sich durchaus nicht die Notwendigkeit, den Begriff der Bedeutung zu modifizieren, und man ist nicht darin gerechtfertigt zu sagen, man *müsse* sich die Bedeutung, die in zirkulären Definitionen festgesetzt werde, anders vorstellen – auch wenn im letzten Satz des obigen Zitats gerade das behauptet wird.

Wenn nun für zirkuläre Definitionen die klassische Erklärung der Bedeutung nicht mehr angewendet werden soll, wie wollen Gupta und Belnap sie dann verstehen?

> A circular definition, though it may not determine the extension of the definiendum, does provide a rule that can be used to calculate what the extension should be *once we make a hypothesis concerning the extension of the definiendum*. This is the key, in our view, to the problem of meaning before us. The meaning a circular definition ascribes to its definiendum, we wish to suggest, should be viewed as having a *hypothetical* character. ([18], S. 119)

Die Idee, die Gupta und Belnap haben, ist zu sagen: Zirkuläre Definitionen legen nicht absolute Anwendungsbedingungen, sondern sozusagen nur hypothetische Anwendungsbedingungen fest. Wird etwa ein Prädikat G zirkulär definiert, so haben wir eine beliebige Hypothese bzgl. seiner Extension aufzustellen, die übrigens in keiner Weise gerechtfertigt zu sein braucht, und dann unter dieser Vermutung zu schauen, was die Definition von G als neue Extension für G ergibt, wenn man dem Vorkommnis von G im Definiens diejenige Extension zuordnet, die man in der Hypothese für G angenommen hat, und dann das Gleichsetzungsvefahren anwendet. Dieses erweiterte Gleichsetzungsverfahren läßt sich selbstverständlich auf alle möglichen angenommenen Ausgangsextensionen für das Prädikat G anwenden; was man dabei erhält, ist keine Regel, die uns etwas über die Anwendungsbedingungen des Prädikats G zu sagen gestattet, sondern – um in der Terminologie von Gupta und Belnap zu sprechen – eine Revisionsregel: Die für das zu definierende Prädikat G angenommene Ausgangsextension wird durch die zirkuläre Definition für G revidiert und eine „neue“ als Output ausgegeben. Im Spezialfall nichtzirkulärer Definitionen ergibt sich für jede Ausgangsextension für das zu definierende Prädikat G dieselbe revidierte Extension, da das Definiens das Prädikat G nicht enthält und somit invariant gegenüber einer Änderung von Ausgangsextensionen für G ist.

Sollen sich zirkuläre Definitionen in irgendeiner Weise als ein nützliches, angemessenes Mittel zur Beschreibung unserer Sprachpraxis erweisen, dann muß die Theorie zirkulärer Definitionen mehr sagen als dies, daß zirkuläre Definitionen als Revisionsregeln aufgefaßt werden können. Sie muß erklären, wie man von den hypothetischen Aussagen bzgl. eines Prädikats G zu nicht-hypothetischen oder nicht-bedingten oder kategorischen („categorical“) Aussagen über G gelangen kann. Denn für die meisten Prädikate G unserer Sprache sind wir oftmals imstande zu sagen, ob ein Gegenstand nun in die Extension von G gehört oder

nicht, ohne dabei irgendeine vorangehende hypothetische Aussage über die Extension von G gemacht zu haben. Die zentrale Idee für den Übergang vom Hypothetischen zum Kategorischen sehen Belnap und Gupta durch die Vorstellung gegeben, zirkuläre Definitionen seien Revisionsregeln, die zu einer Ausgangsextension eine zum *Besseren*[49] hin revidierte Extension liefern. Beginnend mit einer Extension X für G erhalten wir unter der Anwendung der Definition D einen besseren Kandidaten $\delta_D(X)$ als Extension von G. Diese so erhaltene Extension unterwirft man erneut der durch die Definition gelieferten Revisionsregel, indem man nun $\delta_D(X)$ als hypothetische Extension für G annimmt. Man erhält einen noch besseren Kandidaten $\delta_D(\delta_D(X))$ als Extension von G. Dieses Verfahren läßt sich „beliebig weit“ fortsetzen. Stellt sich nun heraus, daß ab irgendeiner endlichmaligen Anwendung der Revisionsregel der Gegenstand a sich bei allen folgenden Anwendungen der Revisionsregel in den Extensionen von G befindet, dann wird man sagen wollen, daß der Gegenstand a unter der aller ersten Anfangshypothese X für G in die Extension von G gehört. Wenn sich nun gar zeigen läßt, daß für jede Ausgangshypothese X für G nach endlichmaligen Schritten der Anwendung der Revisionsregel sich schließlich (immer) Extensionen für G ergeben, in denen der Gegenstand a vorkommt, dann kann man sich von der Relativierung auf die Ausgangshypothese befreien und die absolute nichthypothetische Aussage treffen, daß – wie immer es auch um die anderen Gegenstände bestellt sein mag – zumindest der Gegenstand a ein G ist. In der weiter unten erläuterten Terminologie von Gupta und Belnap würde man sagen, daß der Satz ‘Ga’ gültig („valid“) bzw. kategorisch („categorical“) ist.

### 5.3.1 Beispiel für eine zirkuläre Definition und $S_0$

Zur Motivation der oben schon angeklungenen semantischen Begrifflichkeit für zirkuläre Definitionen betrachten wir die folgende formale Definition des Prädikats G:

$$G(x) \quad =_{Df} \quad (F(x)\&H(x)) \vee (F(x)\&\neg H(x)\&G(x)) \vee (\neg F(x)\&H(x)\&\neg G(x))$$

Wir nehmen an, daß wir in einer Sprache der klassischen Quantorenlogik arbeiten, für die eine Interpretation gegeben sei. Insbesondere gelte das Bivalenzprinzip, d.h. daß alle G-freien Sätze[50] in einer Interpretation entweder den Wahrheitswert wahr oder falsch erhalten. Gupta und Belnap reden statt von der „Interpretation“ von den Fakten ‘M’, die hier die folgenden sind:

*Gegenstandsbereich:* W = **{a, b, c, d}**
*Interpretation der Prädikatbuchstaben ‘F’ und ‘H’* : I(F) = **{a, b}**; I(H) = **{a, c}**
*Interpretation der Namenbuchstaben ‘a’, ‘b’, ‘c’ und ‘d’* : I(a) = **a**; I(b) = **b**; I(c)

[49] Die zweistellige Relation ‘ist besser als’ wird schwach gelesen als ‘ist echt besser als oder genau so gut wie’. Sie soll eine transitive Relation sein.

[50] Sätze sind wohlgeformte Formeln ohne freie Vorkommnisse von Variablen.

= **c**; I(d) = **d**.

Die Extension von G ist durch die Definition nicht festgelegt. Wir nehmen daher zunächst willkürlich an, daß kein Gegenstand des Gegenstandsbereichs ein G ist, indem wir die Interpretation I zur Interpretation I' erweitern, die mit I identisch – mit der Ausnahme, daß I' jetzt auch G interpretiert: I'(G) = $\emptyset$ (= leere Menge). Unter dieser Annahme können wir die aus der Definition resultierende, revidierte Extension für G errechnen, indem wir für das Vorkommnis von G im Definiens die leere Menge als Extension annehmen und nachschauen, welches der Objekte **a**, **b**, **c** oder **d** aus dem Gegenstandsbereich W die offene Satzformel auf der rechten Seite von '$=_{Df}$' unter dieser erweiterten Interpretation erfüllen. Es stellt sich heraus, daß es **a** und **c** sind.[51] Nimmt man als Ausgangsextension für G nun nicht die leere Menge $\emptyset$, sondern den gesamten Gegenstandsbereich W, so ergibt sich als neue Extension für G die Menge {**a**, **b**}. Für andere Teilmengen von W als mögliche Ausgangsextensionen für G bekommt man wieder neue, eventeuell mit der Ausgangsextension identische Teilmengen von W: die revidierten Extensionen von G. Somit ist durch die Definition für G und die relevanten Fakten M genau eine Funktion

$$\delta_{D,M} : \mathcal{P}(W) \rightarrow \mathcal{P}(W)$$

gegeben, die jeder Teilmenge von W als Element der Potenzmenge $\mathcal{P}(W)$ (= Menge aller Teilmengen von W) wieder eine Teilmenge von W zuordnet. Die Indizes 'D' und 'M' weisen auf die Definition und die relevanten Fakten M hin, durch die $\delta_{D,M}$ bestimmt ist. $\delta_{D,M}(X)$ ist somit die revidierte Extension für G, wenn 'X' der Platzhalter für die für G angenommene Ausgangsextension ist. Man kann durch einfache Überlegungen im Falle unseres Beispiels nachweisen, daß gilt: $C \in X \quad gdw. \quad C \notin \delta_{D,M}(X)$. Daher gilt für alle Inputs, daß $\delta_{D,M}(X) \neq X$. D.h. wir haben keine Ausgangshypothese X für G, so daß die revidierte Extension $\delta_{D,M}(X)$ mit der Ausgangshypothese identisch wäre. Folglich ordnet eine zirkuläre Definition im Allgemeinen dem zu definierenden Prädikat G keine Extension zu. Denn welche sollten wir aus all den ausgegebenen Extensionen auswählen, da doch keine von ihnen besonders heraussticht? Ein Fixpunkt X, also ein X mit $\delta_{D,M}(X) = X$, wäre ein guter Kandidat gewesen. Wie aber schon erwähnt und bereits in der Verwendung der entsprechenden Termini vorweggenommen, begegnen Gupta und Belnap diesem Problem mit der Überlegung, daß die Funktionen $\delta_{D,M}$ Revisionsregeln darstellen, die zu einer Ausgangsextension für das zu definierende Prädikat G eine zumindest gleich gute revidierte Extension ausgeben, d.h. unsere Anfangshypothese bzgl. der Extension von G wird durch die Revisionsregel korrigiert und eine neue Extension für G ausgegeben, die – so muß man Gupta und Belnap verstehen – deswegen besser oder zumindest genauso gut ist, weil sie nicht willkürlich gewählt wurde. Durch die Betrachtung mehrmaliger Anwendungen der Revisionsfunktion in sogenannten Revisionsfolgen oder

[51] **a** erfüllt das erste Disjunktionsglied, **c** das letzte.

-ketten können dann schließlich nicht-hypothetische Aussagen, die eventuell das definierte Prädikat G enthalten, getroffen werden.

Man definiert rekursiv zu einer beliebigen Teilmenge X des Gegenstandsbereichs die Folge von immer besseren bzw. zumindest gleich guten Extensionen für G, wobei wieder 'D' die Definition von G und 'M' die Fakten M symbolisiert:

$\delta^0_{D,M}(X) = X, \quad \delta^{n+1}_{D,M}(X) = \delta_{D,M}(\delta^n_{D,M}(X))$

Das ist die Revisionsfolge für die Ausgangshypothese X für G. Bei unserem Beispiel mit der Definition D und den relevanten Fakten M ergibt sich etwa für X = $\emptyset$ die Revisionskette:

$\delta^0_{D,M}(\emptyset) = \emptyset, \delta^1_{D,M}(\emptyset) = \delta_{D,M}(\delta^0_{D,M}(\emptyset)) = \delta_{D,M}(\emptyset) = \{\mathbf{a}, \mathbf{c}\}.$

Ebenso erhält man:

$\delta^2_{D,M}(\emptyset) = \{\mathbf{a}\}, \delta^3_{D,M}(\emptyset) = \{\mathbf{a}, \mathbf{c}\}, \delta^4_{D,M}(\emptyset) = \{\mathbf{a}\}$ usw.

Beginnen wir mit der Ausgangshypothese $X = W = \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$, dann ergibt sich folgende Revisionsfolge:

$\delta^0_{D,M}(W) = \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}, \delta^1_{D,M}(W) = \{\mathbf{a}, \mathbf{b}\}, \delta^2_{D,M}(W) = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\},$

$\delta^3_{D,M}(W) = \{\mathbf{a}, \mathbf{b}\}, \delta^4_{D,M}(W) = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\} \ldots$

Ebenso kann man andere Teilmenge X des Gegendstandsbereichs W wählen und die dazu gehörigen Revisionsfolgen berechnen.

Auf der Grundlage dieser Revisionsfolgen lassen sich schließlich Begriffe einführen, mit denen der erstrebte Sprung vom Hypothetischen zum Kategorischen erfolgt. In den folgenden Definitionen werden Begriffe erklärt, mit denen man die oben gemachten intuitiven Überlegungen technisch genau fassen kann. Mit diesen Begriffen ist aber nur eine von vielen Möglichkeiten gegeben, den Sprung vom Hypothetischen zum Kategorischen zu erreichen. Dasjenige semantische System, das durch die folgenden Definitionen charakterisisiert wird, nennen Belnap und Gupta $S_0$. Es ist ein relativ einfaches System, in denen transfinite Ordinalzahlen keine Rolle spielen, da keine Revisionfolgen betrachtet werden, deren Definitionsbereich über die natürlichen Zahlen hinausgehen. Für die Veranschaulichung der grundlegenden Überlegungen der Revisionstheorie reicht dieses System, wie auch der im nächsten Abschnitt beschriebene zugehörige Kalkül $C_0$, vollkommen aus, wird aber von Belnap und Gupta verworfen. Sie beschreiben andere semantische Systeme, die den gestellten Ansprüchen eher genügen.

In der folgenden Definitionen gehe ich von dem Fall aus, daß eine einzige Definition, nämlich eine Definition D eines einstelligen Prädikats G, gegeben ist. Korrekter müßte man bei den folgenden (Meta-)Definitionen davon ausgehen, daß eine ganze Menge von Definitionen vorliegt, um eventuellen gegenseitigen

Abhängigkeiten[52] unter den Definitionen Rechnung tragen zu können.

**Das semantische System** $S_0$

Zunächst wird erklärt, wann ein Satz A in der Sprache mit dem definierten Zeichen G wahr in dem Modell M+X ist, welches diejenige Erweiterung des Modells M darstellt, bei der der einstellige Prädikatbuchstabe 'G' durch die Teilmenge X des Gegenstandsbereichs von M interpretiert wird:

**(D1)** Ein Satz A ist *wahr in* $M + X$ $\Leftrightarrow_{Df}$
A ist wahr in M relativ zur Hypothese, daß X die Extension von G ist.

**(D2)** Ein Satz A ist *gültig („valid") in M* relativ zur Definition D im System $S_0$ $\Leftrightarrow_{Df}$
Es gibt eine natürliche Zahl p, so daß für alle $q \geq p$ und alle Teilmengen X des Gegenstandsbereichs gilt: A ist wahr in $M + \delta^q_{D,M}(X)$.[53]

**(D3)** Ein Satz $A$ ist *kategorisch in M* relativ zu D in $S_0$ $\Leftrightarrow_{Df}$ Entweder $A$ oder seine Negation $\neg A$ ist gültig in M relativ zu D in $S_0$.

**(D4)** Ein Satz A ist *pathologisch in M* relativ zu D in $S_0$ $\Leftrightarrow_{Df}$ A ist nicht kategorisch in M relativ zu D in $S_0$

Zur Aussonderung der logisch wahren, kurz: gültigen, Sätze im System $S_0$ quantifiziert man wie üblich über alle Interpretationen M:

**(D5)** Ein Satz $A$ ist *gültig* relativ zu D in $S_0$ $\Leftrightarrow_{Df}$ $A$ ist für alle klassischen Interpretationen M gültig in M relativ zu D in $S_0$.

Ein in einem Modell M gültiger Satz ist nach der Definition (D2) demnach wahr in M - welche Hypothese bzgl. der Extension des Definiendums G man auch immer treffen mag, d.h. daß man solche Sätze ohne wenn und aber bejahen kann. Die problematischen Fälle sind die pathologischen Sätze. Betrachten wir zwei Beispiele:

**Bsp.1:** Es liege hier wie auch im zweiten Beispiel wieder die Definition D des einstelligen Prädikats G aus dem ersten Beispiel vor. Der Satz G(a) ist falsch

[52] Gemeint ist eine Abhängigkeit der folgenden Form: Im Definiens einer Definition D1 kommt das Definiendum einer Definition D2 vor und im Definiens der Definition D2 kommt das Definiendum von D1 vor. Damit wären D1 und D2 „versteckt" zirkulär.

[53] Gupta und Belnap führen als eine alternative Formulierung von **D2** die folgende Definition an:

**(D2')** A ist gültig in M relativ zu D in $S_0$ $\Leftrightarrow_{Df}$ Es gibt eine natürliche Zahl p, so daß für alle Teilmengen X des Universums W gilt: A ist wahr in $M + \delta^p_{D,M}(X)$.

Diese Definition ist mit der ursprünglichen gleichwertig.

in $M+\emptyset$. Denn mit der Hypothese, die Extension von G sei $\emptyset$, erhalten wir, daß kein Ding in der Extension von G ist, insbesondere nicht der Gegenstand I(a). Der Satz G(a) ist sogar für alle $n > 0$ wahr in $M+\delta^n_{D,M}(\emptyset)$. Denn nach der ersten Revision für die Ausgangshypothese $\emptyset$ ist der Gegenstand I(a) in jedem $\delta^n_{D,M}(\emptyset)$ enthalten. Mit welchem X wir auch anfangen, A ist wahr in $M + \delta^n_{D,M}$ für alle $n > 0$, d.h. 'G(a)' ist gültig in M relativ zu D.

**Bsp.2:** Der Satz A' ='$H(a)$ & $G(c)$' ist wahr in $M + \delta^n_{D,M}(\emptyset)$, falls n ungerade ist, falsch wenn n gerade. Somit ist A' nicht gültig. A' ist aber in einem noch stärkeren Maße davon entfernt, gültig zu sein; denn A' oszilliert bzgl. seines Wahrheitswertes für jede Ausgangshypothese X für G. Sätze, die eine deratige Oszillation des Wahrheitswertes bzgl jeder Anfangshypothese zeigen, werden **paradox** in M relativ zu D in $S_0$ genannt.

Sätze von demjenigen Typ, wie er im ersten Beispiel vorliegt, könnte man trotz der zirkulären Definition des Prädikats G, welches in ihnen vorkommt, bedingungslos unterschreiben, ihre Negationen in entsprechender Weise bedingungslos ablehnen: Für den Gegenstand a ist durch die Revisionsfolgen bei der angegebenen semantischen Deutung zirkulärer Definitionen festgelegt, daß er ein G ist.
Unter derselben Definition für G und unter denselben Fakten M wie in Bsp. 1 ergibt sich für Sätze des Typs wie im zweiten Beispiel eine andere Sachlage: Wir können sie weder bedingungslos unterschreiben noch verneinen, da sich bei ihnen für jede Ausgangshypothese bzgl. der Extension von G eine Revisionsfolge ergibt, die sich nicht darauf festlegt, ob sie wahr sind oder nicht.

Belnap und Gupta weisen darauf hin, daß sich mit der Revisionssemantik eine noch feinere Unterscheidung von Sätzen treffen läßt als jene, die aus der obigen Einteilung in kategorische und pathologische Sätze und bei letzteren in paradoxe und nicht-paradoxe Sätze resultiert. Eine solche Einteilung führt Yaqub in seinem Buch [50] (S.61-62) vor.

### 5.3.2 Intermezzo: Individuierung der Revisionsregeln

Eine interessante Frage ist, wodurch Revisionsregeln eigentlich individuert sind. Diese Frage werde ich wieder für den Fall behandeln, daß es für eine Sprache nur die Definition eines einstelligen Prädikatbuchstabens G gibt. Der allgemeine Fall von mehreren Definitionen wäre gesondert zu betrachten, da es nach der Lesart von Belnap und Gupta eigentlich für diese gesamte Menge an Definitionen nur eine Revisionsregel gibt.[54] Stellen wir uns eine mögliche Welt $w_1$ vor mit einer Sprache $L_1$ und einer einzigen Definition $D_1$ für ein einstelliges Prädikat $G_1$, welche zirkulär ist. Die nichtsemantischen[55] Fakten über diese Welt werden

[54][18], S.145-146.

[55]Zum Begriff eines nichtsemantischen Faktums siehe z.B. [18], S.17-18. Er wird im 'Was (RTW) leisten soll' betitelten Abschnitt meiner Arbeit erwähnt.

durch ein Modell M repräsentiert, dessen Gegenstandsbereich W sei. In diesem Modell sind sämtliche extensionale Fakten gespeichert. Ist z.B. H ein einstelliges Prädikat aus $L_1$, so wird in M die Extension, die es in $w_1$ hat, nämlich eine Teilmenge von W, gespeichert; oder ist a ein Name aus $L_1$, dann wird seine Extension, nämlich ein Gegenstand aus W, in M gespeichert. Insgesamt seien alle Ausdrücke in $L_1$ bis auf $G_1$ in M interpretiert. Wie oben sei $M + X$ dasjenige Modell der gesamten Sprache $L_1$, bei dem nun zusätzlich $G_1$ als Extension $X$ zugeordnet wird. Außerdem stehe in $L_1$ ein Prädikat 'ist ein gleichseitiges Dreieck' zur Vefügung. Es habe den Sinn, den es auch jetzt im Jahre 2000 in der deutschen Sprache hat. $D_1$ ist die folgende Definition

$(D_1)$ $\quad G_1(x) =_{Df} x$ ist ein gleichseitiges Dreick oder $\neg G_1(x)$

In einer anderen möglichen Welt $w_2$ seien die Verhältnisse identisch mit denen aus $w_1$ – mit der Ausnahme, daß hier statt des Prädikats $G_1$ das Prädikat $G_2$ vorliege. Ansonsten sind aber alle anderen Elemente des Vokabulars genau so wie in $L_1$ und die nichtsemantischen Fakten M sind gleich. Auch die einzige in $L_2$ vorkommende Definition $D_2$

$(D_2)$ $\quad G_2(x) =_{Df} x$ ist ein gleichseitiges Dreick oder $\neg G_2(x)$

ist genau so wie $D_1$, außer daß die Position von '$G_1$' durch '$G_2$' eingenommen wird.

In einer dritten möglichen Welt $w_3$ sind die Verhältnisse identisch mit denen aus $w_3$ – mit der einzigen Ausnahme, daß hier statt des Prädikats 'ist ein gleichseitiges Dreieck' ein Prädikat 'ist ein gleichwinkliges Dreieck' vorliege, dessen Sinn der ist, den dieser Ausdruckstyp auch in der deutschen Sprache im Jahre 2000 hat. Außerdem ist Definition $D_3$ die folgende:

$(D_3)$ $\quad G_1(x) =_{Df} x$ ist ein gleichwinkliges Dreieck oder $\neg G_1(x)$

Die Revisionsregeln für die einzelnen Definitionen lauten folgendermaßen:

$\delta_{D_1,M} : \mathcal{P}(W) \rightarrow \mathcal{P}(W)$ mit
$(\forall d \in W)(\forall X \in \mathcal{P}(W))(d \in \delta_{D_1,M}(X) \quad \Leftrightarrow \quad d$ erfüllt 'x ist ein gleichseitiges Dreieck oder $\neg G_1(x)$' in $M + X$.

$\delta_{D_2,M} : \mathcal{P}(W) \rightarrow \mathcal{P}(W)$ mit
$(\forall d \in W)(\forall X \in \mathcal{P}(W))(d \in \delta_{D_1,M}(X) \quad \Leftrightarrow \quad d$ erfüllt 'x ist ein gleichseitiges Dreieck oder $\neg G_2(x)$' in M+X.

$\delta_{D_3,M} : \mathcal{P}(W) \rightarrow \mathcal{P}(W)$ mit
$(\forall d \in W)(\forall X \in \mathcal{P}(W))(d \in \delta_{D_1,M}(X) \quad \Leftrightarrow \quad d$ erfüllt 'x ist ein gleichwinkliges Dreieck oder $\neg G_1(x)$' in $M + X$.

Man kann sich nun davon überzeugen, daß sich für alle diese drei Definitionen im folgenden Sinne ein und dieselbe Revisionsregel für M ergibt: Wenn man den Begriff einer Funktion mengentheoretisch auffaßt, dann sind die Funktionsausdrücke '$\delta_{D_1,M}$', '$\delta_{D_2,M}$' und '$\delta_{D_3,M}$' nur verschiedene Bezeichnungen derselben Funktion. Im mengentheoretischen Sinne ist eine Funktion nämlich nicht mehr als sein Graph, d.h. eine Funktion ist in diesem Falle eine besondere Menge von Paaren. Es gibt aber andere Lesarten von 'Funktionen'; in einer dieser Lesarten ist nämlich noch relevant, wie die Funktionsvorschrift aussieht.[56] Z.B. haben die auf der Menge von Paaren von natürlichen Zahlen definierten Funktionen $f$ mit $f(x,y) = (x+y)^2$ und $g$ mit $f(x,y) = x^2 + 2xy + y^2$ zwar denselben Graphen. Aber die beiden Funktionsterme sind in nichttrivialer Weise verschieden: Sie haben zwar dieselbe Intension, aber sie sind in berechnungstheoretisches Hinsicht verschieden. Man kann auch sagen, sie haben einen unterschiedlichen Sinn. Zwei Funktionsausdrücke, die meiner Meinung nach in einem sehr uninteressanten Sinne verschieden sind, wären z.B. 'x+x' und 'y+y'; hier sind nur die Variablen verschieden. Funktionen $f_1$ und $f_2$ mit gleichem Definitions-und Wertebereich und $f_3(x) = x + x$ und $f_4(y) = y + y$ wären nur in einem sehr kruden Sinne verschieden; ob ich nun 'x' oder 'y' oder ein anderes Zeichen als Variable wähle, ist in diesem Falle egal.

Wenn man sich diese Lesart von 'Funktion' zu eigen macht, dann kann man einsehen, daß die Revisonsregel $\delta_{D_1,M}$ verschieden ist von der Revisionsregel $\delta_{D_3,M}$. Bei ersterem wird man zur Bestimmung des Werts $\delta_{D_1,M}(X)$ u.a. damit beauftragt zu prüfen, ob d ein gleichseitiges Dreieck ist: Wir müssen dann schauen, ob d ein Dreieck ist und gleich lange Seiten hat. Bei letzterem hingegen muß man zur Berechnung des Wertes $\delta_{D_3,M}(X)$ u.a. prüfen, ob der Gegenstand d ein gleichwinkliges Dreieck ist; hier muß ich prüfen, ob d ein Dreieck ist, mit drei gleichen Winkeln. Für einen Computer, der nicht um die intensionale Äquivalenz der Ausdrücke 'ist eine gleichwinkliges Dreieck' und ist 'ein gleichseitiges' Dreieck weiß, ergeben sich erhebliche Unterschiede bei der Prüfung des einen oder anderen. Den Weg, Revisionsregeln auf diese Weise zu individuieren, beschreitet Orilia in seinem Aufsatz [36].[57] Belnap und Gupta folgen (zumindest) in RTT *nicht* dieser Individuierung.[58] Diese Individuierung läßt sich für $\delta_{D_1,M}$ und $\delta_{D_3,M}$ einsehen. Orilia würde aber auch behaupten, daß die Regeln $\delta_{D_1,M}$ und $\delta_{D_2,M}$ unterschieden werden müssen.[59] Da er auch sagen würde, daß der Sinn der stipulativ zirkulär definierten Prädikate $G_1$ und $G_2$ ihre Revisionsregeln sind, würde das bedeuten, daß $G_1$ und $G_2$ unterschiedlichen Sinn haben. Ich meine, daß der hier zugrunde gelegte Begriff des Sinn nicht unseren Intuitionen entsprechen kannn. Das Verhältnis der Revisionsregeln $\delta_{D_1,M}$ und $\delta_{D_2,M}$ ist doch ähnlich dem Verhältnis der oben aufgeführten Funktionen $f_1$ und $f_2$. Hier werden nur andere Buchstaben

[56] Siehe z.B. [35], S.41
[57] [36], S. 163
[58] Das folgt aus einem Beispiel, welches sie in [18], S.130 geben.
[59] [36], S.164

verwandt, um die Funktionsvorschrift anzugeben. Der Sinn dieser beiden Funktionen ist doch derselbe. Ebenso ist der Sinn von $G_1$ in $w_1$ genau der Sinn, den $G_2$ in $w_2$ hat. Ich wüßte keine Auslegung von 'Sinn', die es rechtfertigen könnte zu sagen, $G_1$ und $G_2$ hätten einen unterschiedlichen Sinn.

## 5.4 Regeln zu zirkulären Definitionen: Der Kalkül $C_0$

Nachdem Gupta und Belnap im zweiten Abschnitt (II) des vierten Kapitels kurz die Semantik zirkulärer Definitionen dargestellt haben, wird im dritten Abschnitt einiges zu den logischen Regeln von zirkulären Definitionen nachgetragen. Diese legen grob gesagt fest, wie man mit zirkulären Definitionen bei einer Ableitung in einem vorgegebenen Kalkül umzugehen hat. Daß hier nicht die üblichen logischen Regeln der Standardtheorie angebracht sein können, welche das „logische Verhalten" von Definitionen durch die unten noch einmal angegebenen Einführungs- und Beseitigungsregeln (DfE) (für Definiendum-Einführung) und (DfB) (für Defininiendum-Beseitigung) als dasjenige von Bikonditionalen bestimmen, wird etwa durch folgendes Beispiel bestätigt:
Nehmen wir an, das logische Verhalten von $=_{Df}$ sei insofern das des Bikonditionals $\leftrightarrow$, als bei gegebener Definition

$G(x) \quad =_{Df} \quad —x—$

(mit einem in x offenen Satz(schema) '$—x—$' ) folgende Regeln gestattet sind:

**(DfB)** Ist im Laufe einer Ableitung (etwa in einem Kalkül der Logik erster Stufe) 'G(x)' abgeleitet worden, so darf man zum Definiens '$—x—$' übergehen.

**(DfE)** Ist im Laufe einer Ableitung '$—x—$' abgeleitet worden, so darf man zum Definiendum 'G(x)' übergehen.

Man betrachte nun konkret folgende Definition für G:[60]

$G(x) \quad =_{Df} \quad Fx \vee (Hx \,\&\, \neg Gx)$

Hieraus läßt sich allein mit Hilfe der üblichen Einführungs- und Beseitigungsregeln des Gentzen- Lemmon-Kalküls zuzüglich der beiden oben aufgeführten Regeln (DfB) und (DfE) die Formel $(\forall x)(Hx \rightarrow Fx)$ ableiten, die dann auch nur

[60] Das ist auch Guptas und Belnaps Beispiel, mit dem sie ihre Regeln für zirkuläre Definitionen motivieren ([18], S.127).

von der Definition abhängt.[61] Interpretiert man H durch 'ist mit sich identisch' und F durch 'ist identisch mit dem Einen', dann hätte man a priori den Monismus bewiesen – eine, wie man meinen sollte, unliebsame Konsequenz. Aus diesem Grunde können (DfE) und (DfB) nicht zu den logischen Regeln gehören, wenn wir an sämtlichen zirkulären Definitionen festhalten möchten, ohne dabei gleichzeitig die These einzukaufen, der Monismus ließe sich aus einer bestimmten Definition a priori beweisen. Bei der vorzunehmenden Modifikation der Regeln lassen sich Gupta und Belnap von dem bereits erläuterten semantischen Schema zirkulärer Definitionen leiten:
Erfüllt der Gegenstand a das Definiens A(x,G)[62] in der Revisionsstufe i – also nach i-maliger Anwendung des Operators $\delta_{D,M}$ auf eine beliebige Ausgangshypothese –, dann ist a ein G in der i+1-ten Revisionsstufe.
Umgekehrt gilt in ähnlicher Weise, daß falls ein Gegenstand a ein G ist in der i+1-ten Revisionsstufe, dann dieser Gegenstand auch A(x,G) erfüllt, nämlich dann in der i-ten Stufe.
Es wird folglich beim Übergang vom Definiens zum Definiendum und umgekehrt sozusagen ein Protokoll über die dabei durchlaufenen Revisionsstufen zu führen sein, was sich durch eine Indizierung der Formeln bewerkstelligen läßt. Somit kann man die modifizierten Regeln $(DfB_r)$ und $(DfE_r)$ folgendermaßen aufschreiben:

$$(\mathbf{DfE_r})\ \frac{A(t,G)^i}{[G(t)]^{i+1}}$$

$$(\mathbf{DfB_r})\ \frac{[G(t)]^i}{A(t,G)^{i-1}}$$

Der Kalkül, in dem die zirkulären Definitionen mit diesen beiden Regeln eingebettet wird, ist ein üblicher Kalkül des natürlichen Schließens, in dem die allseits

[61] Die betreffende Ableitung könnte so aussehen:

| | | | |
|---|---|---|---|
| 1 | (1) | $G(x) \quad =_{Df} \quad Fx \vee (Hx \,\&\, \neg Gx)$ | $Definition$ |
| 2 | (2) | $\neg Fa \,\&\, Ha$ | $A$ |
| 3 | (3) | $Ga$ | $A$ |
| 1,3 | (4) | $Fa \vee (Ha \,\&\, \neg Ga)$ | $1,3,DfB$ |
| 1,2,3 | (5) | $\neg Ga$ | $2,4,Theorem$ |
| 1,2,3 | (6) | $Ga \,\&\, \neg Ga$ | $3,5,\&E$ |
| 1,2 | (7) | $\neg Ga$ | $3,6,RAA$ |
| 2 | (8) | $Ha$ | $2,\&B$ |
| 1,2 | (9) | $Ha \,\&\, \neg Ga$ | $7,8,\&E$ |
| 1,2 | (10) | $Fa \vee (Ha \,\&\, \neg Ga)$ | $9,\vee E$ |
| 1,2 | (11) | $Ga$ | $1,10,DfE$ |
| 1,2 | (12) | $Ga \,\&\, \neg Ga$ | $7,11,\&E$ |
| 1 | (13) | $\neg(\neg Fa \,\&\, Ha)$ | $2,12,RAA$ |
| 1 | (14) | $Fa \vee \neg Ha$ | $13,Theorem$ |
| 1 | (15) | $Ha \rightarrow Fa$ | $14,Theorem$ |
| 1 | (16) | $(\forall x)(Hx \rightarrow Fx)$ | $15,\forall E$ |

[62] A(x,G) ist eine in x offene Formel, die möglicherweise das Prädikat G enthält.

bekannten Ableitungsregeln verwendet werden – mit der zusätzlichen durch die Indizierung bedingten Einschränkung, daß zum einen alle Formeln, die die Prämissen einer Regelanwendung bilden, denselben Index besitzen müssen und daß zum anderen die aus der Anwendung einer von $(DfE_r)$ und $(DfB_r)$ verschiedenen Regel resultierende Formel denselben Index wie die Prämissen aufzuweisen hat.
Als eine zusätzliche Regel kommt der Index-Shift (IS) hinzu, der die Änderung des Indexes einer Formel zu einem beliebigen ganzzahligen Index gestattet, sofern diese nicht das Definiendum G enthält. Die intuitive Begründung für diese Regel ist, daß es für Formeln, die G nicht enthalten, nicht von Belang ist, welche Extension für G in einer Revisionstufe gerade ansteht: Da besagtes G nicht in ihr vorkommt, bleibt ihr Wahrheitswert über die Revisionsstufen gleich.
Zusammengefaßt enthält somit der dem semantischen System $S_0$ entsprechende Kalkül $C_0$ als Regeln:

> **Regeln des Kalküls** $C_0$
> $(DfE_r)+(DfB_r)+$Index Shift$+$Klassische Logik (mit der oben erwähnten Einschränkung bzgl. der Indizes.)

In diesem Kalkül wird die „Ableitbarkeit“ unter Berufung auf den aus der klassischen Logik bekannten Begriff der Ableitung erklärt:
Ein Satz(schema) $A$ ist *ableitbar* auf der Basis einer Definition D $\Leftrightarrow_{Df}$ Es gibt eine Ableitung von $A^0$ in $C_0$ auf der Basis von D.[63]

**Bsp.:** Wieder liege für G die Definition $G(x) =_{Df} Fx \vee (Hx \;\&\; \neg Gx)$ vor. Dann läßt sich z.B. das Satzschema ‘$(\forall x)(Fx \rightarrow Gx)$’ auf der Basis von D ableiten:

| | | | |
|---|---|---|---|
| $1^*$ | (1) | $(Fa)^0$ | $A$ |
| $1^*$ | (2) | $(Fa)^{-1}$ | $1, IndexShift$ |
| $1^*$ | (3) | $(Fa \vee (Hx \;\&\; \neg Gx))^{-1}$ | $2, Theorem$ |
| $1^*$ | (4) | $(Ga)^0$ | $3, DfE_r$ |
| | (5) | $(Fa \rightarrow Ga)^0$ | $1, 4 \rightarrow E$ |
| | (6) | $(\forall x)(Fx \rightarrow Gx)^0$ | $5, \forall E$ |

Da im Kalkül $C_0$ nicht mehr die klassischen Regeln (DfE) und (DfB) gelten,

[63] Diese Formulierung der Definition entspricht nicht ganz dem Wortlaut des Orginals. Dort heißt es ([18], S.127), ein Satz A sei *ableitbar aus D* gdw [...] Diese Redeweise führt zu Problemen, wenn die Ableitbarkeit eines Argumentschemas erklärt werden soll und man weiterhin die übliche Rede der „Ableitbarkeit einer wohlgeformten Formeln aus anderen Formeln“ aufrechterhalten möchte: Dann müßte man z.B. sagen, die Formel K sei aus den Prämissenformeln A, B und C und aus der Definition D ableitbar. Das erweckt die Assoziation, Definitionen – ob nun zirkulär oder nicht – seien im Prinzip nichts anderes als womöglich besonders gekennzeichnete Sätze bzw. Satzschemata. So aber wollen Gupta und Belnap Definitionen welcher Art auch immer nicht verstanden wissen.

kann nun auch nicht mehr das aus der klassischen Definitionstheorie bekannte Verfahren übernommen werden, eine in L formulierte Definition

$G(x) =_{Df} A(x, G)$

als einen besonderen Satz der Sprache L aufzufassen, welches logisch dieselbe Rolle einnimmt wie der offene Satz

$(G(x) \leftrightarrow A(x, G))$

bzw. wie der Satz

$(\forall x)(G(x) \leftrightarrow A(x, G))$

welcher das klassische materiale Bikonditional enthält. Es muß folglich explizit zwischen '$=_{Df}$' und '$\leftrightarrow$' bzg. des logischen Verhaltens unterschieden werden.

Nähere Untersuchungen bzgl. $C_0$ zeigen:

- $C_0$ ist bzgl. $S_0$ korrekt und vollständig. Der Nachweis der Korrektheit und Vollständigkeit kann als eine Legitimation der Regeln bzgl. der durch $S_0$ beschriebenen Semantik verstanden werden. Eine zusätzliche, über den Korrektheits- und Vollständigkeitsnachweis gehende Rechtfertigung der Regeln von $C_0$ scheint daher nicht mehr nötig. Als Folge aus der Korrektheit sind zirkuläre Definitionen nicht kreativ; als Folge aus der Vollständigkeit sind sie aber auch nicht zu schwach.

- In kategorischen Kontexten kann mit den alten Regeln (DfE) und (DfB) gearbeitet werden.

- Bei nicht-zirkulären Definitionen kann man auf die Indizes verzichten.[64]

## 5.5 Nicht-Kreativität und Eliminierbarkeit in (RTD)

Fragen wir uns noch einmal, was die Motivation dafür war, die für übliche nicht-zirkuläre Definitionen geltenden Regeln (DfB) und (DfE) zu modifizieren. Die Antwort gibt die oben ausgeführte Ableitung: Würde man (DfB) und (DfE) unverändert in die Theorie zirkulärer Definitionen übernehmen, so könnte allein auf der Grundlage einer zirkulären Definition a priori der Monismus bewiesen werden. Wenn mir annehmen, daß sich der Monismus ohne diese Definition nicht ableiten läßt, dann würde das bedeuten, daß die zirkuläre Definition dafür verantwortlich zu machen ist, daß nach ihrer Einführung die Monismusthese abgeleitet werden

[64] Das beweist die folgende Ableitung:

| | | |
|---|---|---|
| (1) | $G(t)^i$ | |
| (2) | $A(t, G)^{i-1}$ | $DfB_r$ |
| (3) | $A(t, G)^{j-1}$ | $IndexShift$ |
| (4) | $G(t)^j$ | $DfE_r$ |

kann: Alles ist identisch mit dem Einen. Wie weit man auch den Definitionsbegriff strapaziere und abschwäche, zumindest das darf nach Guptas und Belnaps Intuitionen eine Definition nicht sein: etwas, das eine vorher nicht beweisbare Aussage abzuleiten gestattet.[65] Definitionen sollen nicht mehr leisten, als die Bedeutung von sprachlichen Ausdrücken zu fixieren. Das Nicht-Kreativitätskriterium der Standardtheorie von Definitionen wird von Gupta und Belnap ohne Vorbehalte in ihr System zirkulärer Definitionen übernommen. Daß das nicht auch für das Eliminierbarkeitskriterium gelten kann, scheint offensichtlich zu sein. Denn da die Eliminierbarkeitsforderung der (STDL) manche zirkuläre Definitionen verbietet, in (RTD) aber sämtliche zirkuläre Definitionen als logisch korrekt akzeptiert werden, können Gupta und Belnap das Eliminierbarkeitskriterium nicht gelten lassen. Gegen den Vorwurf, mit der Ausschließung des Eliminierbarkeitskriteriums eine ad-hoc-Maßnahme zu ergreifen, sind Gupta und Belnap solange gefeit, wie dieses Kriterium sich nicht als unbedingt notwendig für Definitionen erweisen läßt. Und hier meinen Gupta und Belnap bisher keine vernünftige Begründung gesehen zu haben - außer der vielleicht, möchte man ergänzen, daß man bisher gut mit ihr gefahren ist. Und so versuchen denn Gupta und Belnap auch dafür zu argumentieren, weshalb das Eliminierbarkeitskriterium im Rahmen einer Theorie der Definition nicht akzeptiert zu werden braucht und sogar nicht besonders plausibel ist. Dazu führen sie die Behauptung an, daß das Eliminierbarkeitskriterium einen ganz anderen Status als das Nicht-Kreativitätskriterium habe, daß hier eine Asymmetrie vorliege: Dieses müsse als eine absolute Forderung, jenes hingegen eher als eine relative Forderung angesehen werden, der man in unterschiedlichen Graden nachkommen kann:

> The two requirements, it seems to us, are not of equal status (at least on certain sorts of definitions). The requirement of noncreativity ensures that definitions only fix meanings and do not conceal substantive assertions. Noncreativity is undoubtedly correct as a requirement on „pure“ definitions, the only kind with wich we are concernd. [...] The requirement of eliminability, on the other hand, has nothing to do with the „purity“ of definitions. It is *not* implied by the idea that definitions only fix meanings and do not involve substantive assertions. [...] Eliminability, we believe, is best viewed not as an absolute requirement but as a relative one, one that can be met in varying degrees. The degree of eliminability required of a definition depends upon the context in which the definition is used. ([18], S.128-129)

Worin genau also besteht nach Belnap und Gupta die Relativität des Eliminierbarkeitskriteriums?

Prinzipiell könnten mit der Behauptung, das Eliminierbarkeitskriterium sei ein relatives, eines der beiden folgenden Dinge gemeint sein:

[65] Es genügt natürlich schon der Hinweis, daß sich aus zirkulären Definitionen Widersprüche ableiten lassen, um zu erläutern, weshalb (DfB) und (DfE) nicht uneingeschränkt für zirkuläre Definitionen gelten können.

- Es gibt verschiedene Kontextsorten: Die Sorte der extensionalen, der intensionalen und hyperintensionalen Kontexte. (EK) könnte in der Hinsicht als relative Forderung angesehen werden, daß je nach dem, was von einer Definition erwartet wird, die Eliminierbarkeit aus sämtlichen Kontexten bestimmter vorgegebener Kontextsorten verlangt wird. Z.B. könnten wir verlangen, daß mit einer Definition der Sinn eines Ausdrucks festgesetzt werden soll. Dann fordert (EK), daß alle Vorkommnise des definierten Ausdrucks in mindestens allen extensionalen und intensionalen Kontexten eliminierbar sein müssen. Die Eliminierbarkeit in allen Kontexten [66] ist eine hoffnungslose und so gut wie von keiner Definition erfüllbare Bedingung. So ist z.B. in dem Anführungskontext des Satzes

  > „Das Wort 'Bruder' beginnt mit 'B'"

  die Inskription 'Bruder' nicht salva veritate durch 'männliches Geschwister' ersetzbar, obwohl der Satz

  > 'Für alle Dinge x gilt: x ist ein Bruder gdw x ein männliches Geschwister ist'

  eine adäquate Definition des Prädikats 'ist ein Bruder' ist. Es muß daher die Formulierung des Eliminierbarkeitskriteriums eine Spezifikation der Kontextsorte, aus dem die Eliminierbarkeit gewährleistet werden soll, angegeben werden. Und diese Spezifikation kann von Fall zu Fall variieren; manchmal genügt es, die Elimination der definierten Ausdrücke aus intensionalen Kontexten zu fordern, in anderen könnte es nötig sein, die Eliminierbarkeit aus hyperintensionalen (intentionalen)[67] Kontexten zu fordern.

- Ein Vorkommnis des Definiendums mag aus einem Kontext K eliminierbar sein, aus einem anderen Kontext C aber, der zum selben Typ von Kontext wie K oder einem anderen Kontext angehört, nicht. Dieser Punkt ist eine Verallgemeinerung des ersten Punktes. Die von (EK) vorgenommene Spezifikation ist hier eine andere: Hier wird nicht eine Kontextgruppe angegeben, die durch bestimmte ihnen allen gemeinsame Eigenschaften charakterisiert sind, sondern eher willkürlich aus der Kontexgruppe ausgewählt, z.B. solche, aus denen Eliminierbarkeit nicht gewährleistet sein muß.

Da Belnap und Gupta im Rahmen von rein extensionalen Sprachen arbeiten, ist die Kontextsorte vorgegeben: Es sind extensionale Kontexte. Worauf Belnap und Gupta abzielen, ist also nicht die Relativität von (EK) bzgl. der bekannten

[66] „Kontextïst an dieser Stelle in der weitest möglichen Lesart zu nehmen.

[67] Z.B. kommt ein Satz P im Satz ⌜ Gerhard Schröder glaubt, daß P ⌝ in einem hyperintensionalen Kontext vor. In diesem Kontext ist nicht mehr gewährleistet, daß der Wahrheitswert des gesamten Satzes gleich bleibt, wenn man einen Ausdruck aus P durch einen anderen intensional gleichwertigen Ausdruck derselben grammatischen Kategorie ersetzt.

Kontextsorten, sondern auf die Relativität von (EK) bzgl. Kontexten derselben Sorte. Wenn z.B. ein zirkulär definiertes Prädikat G vorliegt, dann wird G nicht aus allen Kontexten derselben (i.e. extensionalen) Sorte eliminierbar sein. Bei partiell definierten Prädikaten haben wir eine ähnliche Sachlage vorliegen. Die Definition der partiell definierten Divisionsoperation gestattet keine Elimination des Divisionszeichens ':' aus dem folgenden Satzkontext '1:0 =5'. Hier läßt sich vernünftig begründen, warum (EK) nicht die Eliminierbarkeit von ':' aus diesem Kontext fordern darf. Denn für das Zahlenpaar (0,1) ist die Divisionsoperation schlichtweg nicht erklärt worden. Ebenso, scheinen Belnap und Gupta sagen zu wollen, folge, daß man eigentlich nur partielle Eliminierbarkeit für zirkuläre Definitionen fordern dürfe. Die Analogie ist naheliegend: Wozu brauchen wir partielle Definitionen? Es soll eben ein Begriff festgesetzt werden können, der genuin nur auf bestimmte Objekte anwendbar sein soll. Genauso stehe es mit zirkulären Definitionen: Belnap und Gupta haben die Intuition, daß es zirkuläre Begriffe gibt; diese kann man nicht vernünftig behandeln, wenn man keine zirkulären Definitionen gestattet; von letzteren aber, da zirkulär, kann man nicht die Eliminierbarkeit aus allen (extensionalen) Kontexten einfordern. Ein wesentlicher Unterschied zwischen partiellen und zirkulären Definitionen ist aber, daß sich bei ersteren absehen läßt, wann Eliminierbarkeit gewährleistet ist und wann nicht; das erkennt man grob gesagt daran, für welche Dinge der partiell definierte Ausdruck erklärt wurde. In der obigen Defintion des Zeichens wurde nicht erklärt, was '(1:0)' heißen soll. Also kann man auch keine Eliminierbarkeit aus diesem Kontext erwarten. Bei letzteren hingegen kann die Eliminierbarkeit von kontingenten Fakten abhängen.[68]

Die Eliminierbarkeitsforderung erhält in (RTD) selbstverständlich noch ihre Berechtigung, allerdings wird jetzt der Tatsache Rechnung getragen, daß es zirkuläre Begriffe gibt. Die sind für einige Gegenstände pathologisch; wenn die involviert sind, kann man auch keine Eliminierbarkeit mehr fordern. In kategorischen Kontexten bleibt die Eliminierbarkeit aber noch erhalten. Mit (RTD) wird also (EK) nur soweit eingeschränkt, wie es für die Behandlung von zirkulären Begriffen nötig ist. Insofern ist der von (RTD) erzwungene Schritt, das Eliminierbarkeitskriterium aus (STDL) zu beschränken, plausibel.

Im folgenden Abschnitt will ich – wenn auch auf sehr spekulativer Basis – der Frage nachgehen, ob man (NK) als absolute Forderung ansehen muß. Dieser Abschnitt kann übersprungen werden.

[68] Z.B. ist das für das Wahrheitsprädikat der Fall: Wie Kripke betont hat, kann man einem Satz nicht unbedingt sofort ansehen, ob er paradox ist oder nicht. Z.B. ist das für kontingent paradoxe Sätze der Fall. Für einen solchen Satz s, der das Wahrheitsprädikat enthält, ist 'ist wahr' nicht eliminierbar – das wird man aber nicht im voraus sagen können. Beispiel: Aus dem Satz 'Etwas von dem, was Jones äußert, ist wahr' ist das Prädikat 'ist wahr' für die T-Bikonditionale aufgefaßt als partielle Definitionen nicht eliminierbar, wenn Jones tatsächlich nur folgendes geäußert hat: Etwas von dem, was Jones äußert, ist nicht wahr. (Siehe hierzu auch das Kapitel 'Wahrheit' in meiner Arbeit.)

### 5.5.1 Intermezzo: Asymmetrie von (EK) und (NK)

Die von Gupta und Belnap vertretene Asymmetrie der beiden Kriterien wird hier mittels eines Relativitätsnachweises von (EK) geführt. Nicht unherbliche Unterschiede zwischen beiden Kriterien lassen sich bereits in der klassischen Theorie der Definition dingfest machen. Ich werde zunächst diese Unterschiede aufführen, welche als Zeichen dafür gesehen werden können, daß (EK) und (NK) in einem gewissen Sinne tatsächlich einen unterschiedlichen Status haben. Ob diese Unterschiede ausreichend sind, um den Schritt rechtfertigen zu können, (EK) zu verwerfen, aber an (NK) festzuhalten, steht nicht zur Debatte: Diese zu besprechenden Unterschiede führen Belnap und Gupta nicht an, um plausibel zu machen, daß es zwischen (EK) und (NK) einen wesentlichen Unterschied gibt.

Veikko Rantala führt z.B. in seinem Aufsatz „Definitions and Definability“ an, daß (NK) in einem gewissen Sinne schwächer als (EK) ist.[69] Er erklärt in seinem Aufsatz bestimmte Formen von Definierbarkeit, die schwächer sind als die explizite Definierbarkeit, jedoch mit dieser eine Globalitätseigenschaft[70] gemeinsam haben. Für Sätze, die im Sinne dieser schwachen globalen Definierbarkeit definierbar sind, gilt das Theorem, daß sie das Kriterium (NK) erfüllen, nicht aber unbedingt (EK). Dieses Theorem könnte zur Bewertung der Asymmetriefrage hinreichend sein. Die dahinter steckende Idee scheint mir plausibel und von allgemeiner Bedeutung zu sein: Bezogen auf eine Klasse von Objekten eines bestimmten Typs F haben zwei Dinge A und B, von denen man vernünftig sagen kann, sie handeln von Objekten der Sorte F – etwa bestimmte Sätze – denselben Status bezüglich einer Skala S. Diese Skala könnte z.B. der erkenntnistheoretische Wert von Dingen desselben Typs wie A oder B sein oder irgendeine andere Bewertung sein. Ist man nun aber der Meinung, daß A und B dennoch, obwohl sie den gleichen Status auf der Skala S haben, asymmetrisch sind, dann erweitere man die Klasse des Typs F zu einer Klasse F‘ von zusätzlichen Dingen, die bestimmte, aber nicht sämtliche Eigenschaften mit den Dingen des Typs F teilen. Ergeben sich dann Unterschiede von A’ und B’, die nun von F’ handeln, bezüglich der Skala S, dann sind A und B in einem bestimmten (vielleicht) schwächeren Sinne asymmetrisch.[71]

Rantala gibt verschiedene schwache Versionen der Eliminierbarkeit, die ich hier nur erwähnen, nicht besprechen möchte. Zunächst wird eine sogenannte lokale Eliminierbarkeit definiert, lokal deswegen, da auf ein Modell M bezogen. Für ein einstelliges Prädikat G, eine elementare Sprache L bzw. die um G erweiterte Sprache L(G) erklärt man:

[69] [37], S.151

[70] Das meint, daß diese Formen von Definierbarkeit sich nicht auf ein bestimmtes Modell beziehen.

[71] Vergleiche das etwa mit zwei nicht isomorphen Strukturen, die zwar in der Logik erster Stufe genau dieselben Sätze wahr machen, also bzgl. der durch die Logik erster Stufe gestellten Skala gleich sind, aber wegen der Nichtisomorphie nicht als symmetrisch anzusehen sind.

> G ist im Modell M für L(G) eliminierbar $\Leftrightarrow_{Df}$ Für jede in x offene Formel $A(x)$ von L(G) gibt es eine Formel $A'(x)$ in L, so daß gilt: $M \models (\forall x)(A(x) \leftrightarrow A'(x))$

Hiermit lassen sich dann verschiedene schwächere Versionen der Eliminierbarkeit erklären.[72] Allein die Möglichkeit, solche schwachen Lesarten anzugeben, die auf der lokalen Eliminierbarkeit, i.e. auf der Eliminierbakeit in einem Modell, beruhen, reicht als Nachweis für eine Asymmetrie. Denn für das Nicht-Kreativitätskriterium werden derart plausible Abschwächungen de facto nicht gemacht – und scheinen auch auf den ersten Blick zumindest nicht machbar. Das scheint auch ein von Rantala bewiesenes Theorem ([37], S.150) nahezulegen, welches besagt, daß Definitionen, die z.B. restringiert eliminierbar sind, das Nicht-Kreativitätskriterium erfüllen. D.h. daß das Nicht-Kreativitätskriterium die für die Eliminierbarkeit vorgenommene Abschwächung nicht mitmacht: Es ist bereits ein schaches Kriterium, schwächer zumindest als das Eliminierbarkeitskriterium.

Einen weiteren Unterschied zwischen Eliminierbarkeit und Nicht-Kreativität gibt die Tatsache her, daß sich erstere im Gegensatz zur letzteren (nun nicht relativieren, sondern) auch verallgemeinern läßt. Rantala gemäß läßt sich die Eliminierbarkeit auch so formulieren:[73] Ist T(G) eine Theorie in der Sprache L(G) mit einem einstelligen Prädikatbuchstaben 'G', dann ist G in T(G) eliminierbar $\Leftrightarrow_{Df}$ Für jede Formel $A$ von L(G) gibt es eine Formel $A'$ von L, so daß gilt: $T(G) \vdash A \leftrightarrow A'$ ([37], S.148). Ein typischer Fall für die Anwendung dieses Kriteriums wäre der, für den man normalerweise die Eliminierbarkeit formuliert: T(G) ist hier dann nämlich der logische Abschluß einer Menge $\Sigma$ von Sätzen aus der Sprache L vereinigt mit der Menge {D}, wobei 'D' die G enthaltende Definition ist. Die Formulierung von (NK) gestattet keine derartige Verallgemeinerung – zumindest gibt Rantala keine an. Bei (NK) wird explizit davon ausgegangen, daß ein ausgezeichneter Satz D gegeben ist, der der Kandidat für die Definition eines Zeichens G ist.

Wie müßte eine lokale Version von (NK) aussehen? Sei L eine Sprache erster Stufe, $D$ ein Satz aus der um das einstellige Prädikatsymbol G erweiterten Sprache L(G) und $M(G)$ ein Modell für L(G). Was sollte nun die Nicht-Kreativität

---

[72] Eine mögliche schwächere Version der Eliminierbarkeit ist die restringierte Eliminierbarkeit:

> G ist restringiert eliminierbar in einer in L(G) formulierten Theorie T(G) $\Leftrightarrow_{Df}$ Für jede Formel $A(x)$ von L(G) gibt es Formeln $A_i(z,x)$ (i= 0,1,...,n), so daß gilt: $T(G) \vdash \bigvee_{1\leq i\leq n}(\exists z)(\forall x)(A(x) \leftrightarrow A_i(z,x))$ ([37], S.152)

Eine andere Form der Elimierbarkeit gibt die folgende Definition:

> G ist stückweise eliminierbar in T(G) $\Leftrightarrow_{Df}$ Für jede Formel $A(X)$ aus L(G) gibt es eine Formel $A_i(x)$ (i=1,...,n) aus L, so daß gilt:
> $T(G) \vdash \bigvee_{1\leq i\leq n}(\forall x)(A(x) \leftrightarrow A_i(x))$ ([37], S.153)

[73] Rantala allerdings führt dieses Faktum nicht als einen Nachweis dafür an, daß (NK) und (EK) einen unterschiedlichen Status haben.

von $D$ bezüglich des Modells $M$ bedeuten? Der einzige naheliegende Vorschlag kann doch nur darin bestehen zu sagen, daß jeder Satz aus L, der wahr in M ist, auch vor der Einführung von $D$ wahr in M(G) (bzw. M) war. Das ist aber eine immer erfüllte Tatsache. Für (NK) scheint also eine „Lokalisierung“ wie bei (EK) nicht zu funktionieren.

Nun könnte jemand erwidern, die Forderung, eine Definition dürfe bei ihrer Einführung in eine Theorie nicht zu Inkonsistenzen führen, sei eine schwächere Form der Nicht-Kreativitätsforderung und dieses mit dem historischen Hinweis ergänzen, die Nicht-Kreativitätsforderung sei im Anschluß an die Konsistenzforderung formuliert worden – sozusagen als Verallgemeinerung von (NK). Hierüber läßt sich natürlich streiten; der Streit dürfte aber nicht besonders fruchtbar sein. Man kann lediglich hiergegen einwänden, daß die Konsistenzforderung nur eine notwendige Bedingung für (NK) ist, aber eine notwendige Bedingung ist nicht mit einer Relativierung oder Lokalisierung gleichzusätzen. Denn z.B. die Betrachtung des Begriffes des Unverheiratetseins stellt keine Relativierung des Begriffs eines Junggesellen dar. Außerdem fordert eine Relativierung irgendeine zusätzliche Variable, die die Relativierung realisiert. Bei der Konsistenzforderung kommt nichts dergleichen vor.

Die folgenden Überlegungen sind sehr spekulativ. Mit ihnen erwäge ich eine mögliche Relativierung von (NK). Wenn der Vorschlag plausibel erscheint, dann wäre die vermeintliche Asymmetrie zwischen den Kriterien (NK) und (EK), die in der (STDL) immer als ein gleichermaßen plausibles Kriterienpaar aufgeführt werden, zumindest nicht mit einem Relativitätsnachweis zu führen.

Die Relativierung, die ich mir für (NK) vorstellen könnte, ist eine auf eine bestimmte Sorte von Sätzen einer quantorenlogischen Sprache. Ich könnte mir also die Einführung eines Prädikats ‘Der Satz x erfüllt das Nicht-Kreativitätskriterium bzgl. der Satzmenge y’[74] vorstellen, das sich als interessant erweisen könnte. Ich gebe ein Beispiel für die Satzmenge y in diesem Prädikat, welches eine konkrete Relativierung darstellt.

Stellen wir uns vor, wir arbeiten in dem Zermelo-Fraenkelschen Axiomensystem ZF der Mengenlehre. Es läßt sich beweisen, daß

‘2+2=5’ ist kein Theorem von ZF

kein Theorem in ZF ist.[75] Tatsächlich läßt sich zeigen, daß es zu jedem Theorem S ein Theorem von ZF gibt, welches sich so interpretieren läßt, daß es besagt:

[74]Genaugenommen ist das Prädikat kein zwei-stelliges, sondern ein fünfstelliges - ergänzt nämlich um die für die bei der Formulierung von (NK) zusätzlich eingehenden Entitäten freien Positionen - nämlich: ‘x als Definition des Zeichens s erfüllt relativ zu der Hintergrundtheorie T und den vorangehenden Definitionen U bzgl. der Satzmenge y das Nicht-Kreativitätskriterium’.

[75]Der Beweis läßt sich mit Hilfe eines bestimmten modallogischen Systems (GL) führen, indem man den Notwendigkeitsoperator ‘$\Box$’ als ‘es ist beweisbar, daß’ (und entsprechend den Möglichkeitsoperator ‘$\Diamond$’ als ‘ist konsistent mit’) liest und schließlich eine geeignete Übersetzung von dem modallogischen System nach ZF findet. Es läßt sich dann z.B. zeigen, daß der Satz ‘$\neg\Box\,\bot$’ kein Theorem in (GL) ist. Für weitere Details konsultiere man etwa Boolos Buch [7].

S ist ein Theorem.

Es gibt aber für kein Nicht-Theorem S ein Theorem von ZF, welches besagte, daß S kein Theorem ist.

Sei nun G eine Menge von Sätzen, die nicht in ZF ableitbar sind, d.h. die keine Theoreme in ZF sind. Zu jedem Satz S aus G sei ein Satz gebildet, der behauptet, daß S kein Theorem in ZF ist. Die Menge aller so gebildeten Sätze sei H. Nach dem obigen Metatheorem, der zum Inhalt hat, daß es für kein Nicht-Theorem S ein Theorem in ZF gibt, der besagt, daß S ein Nicht-Theorem ist, folgt, daß keines der Sätze in H in ZF ein Theorem ist.

Stellen wir uns nun vor, y sei die Komplementmenge zu der Menge H. Die auf diese Menge y relativierte Nicht-Kreativitätsforderung würde damit fordern, daß mit der Einführung einer Definition in ZF keine Sätze aus der Menge y abgeleitet werden dürfen, die nicht schon vorher in ZF ableitbar waren. Diese auf y relativierte Nicht-Kreativitätsforderung würde es folglich einer Definition nicht verbieten, im Zusammenhang mit der Hintergrundtheorie und den vorangehenden Definitionen solche Sätze abzuleiten zu gestatten, die in H sind – die also die Nicht-Ableitbarkeit von nicht-ableitbaren Sätzen aus G behaupten. Diese relativierte Nicht-Kreativitätsforderung ist natürlich nicht mit (NK) vereinbar. Ich meine aber, daß hier immer noch die Rede von Nicht-Kreativität gestattet ist. Ich glaube, daß die mit der Einführung mancher behauptungslastiger Definitionen eingekauften Behauptungen sozusagen eine Explizitmachung dessen ist, was in der Theorie und dem System schon implizit zu finden ist, daß das System aber nicht über die nötige Ausdruckskraft verfügt (und nicht verfügen kann), mit der sie derartige selbstbezügliche Aussagen treffen könnte. Es werden durch solche Definitionen also nicht genuin neue Aussagen eingeführt. Sie verhelfen der Theorie nur zur Ausdrucksstärke bzgl. bestimmter Aussagen.[76] Und das wollen ja Gupta und Belnap für eine Definition nicht ausschließen, geben sie doch selber auf die Frage, was zirkuläre Definitionen nützen würden, die Antwort, zirkuläre Definitionen würden die Ausdrucksstärke der Sprachen vergößern. Außerdem sind die Aussagen aus H, die die Nichtableitbarkeit bestimmter nichtableitbarer Aussagen behaupten, nicht derart, daß die Theorie nicht über sie – zumindest in einem schwachen Sinne – entscheiden würde: Dadurch, daß in der Theorie diejenigen Aussagen, deren Nichtableitbarkeit die Aussagen aus H behaupten, nicht in der Theorie abgeleitet werden können, bestätigt die Theorie sozusagen das, was die Sätze in H behaupten, sie kann aber über diese Bestätigung nicht reden.

Es ließen sich in Reaktion auf meinen Vorschlag zweierlei Dinge sagen: 1.) Der von mir zugrunde gelegte Begriff der Ausdrucksstärke ist nicht korrekt.
2.) Ich habe kein konkretes Beispiel für eine Definition angegeben, die eine Vergrößerung der Ausdrucksstärke in meinem Sinne ermöglichen könnte. Würde man ein solches Beispiel angeben, dann würde man einsehen, daß hier die Rede von

[76] Gleichzeitig werden mit dieser Definitionen aber wieder mehr Nicht-Theoreme eingebracht, so daß man sich fragen muß, ob die Relativierung tatsächlich irgendetwas nützt.

„Definition“ nicht gerechtfertigt ist.
Ich müßte folglich den Begriff der Ausdrucksstärke genauer explizieren und ein konkretes Beispiel finden, bei dem die Rede von ‘Definition’ noch gerechtfertigt ist. Aber die Überlegung, daß man (NK) derart abschwächen kann, daß sie bestimmte Sätze der alten Sprache doch abzuleiten gestattet, widerspricht – glaube ich – nicht dem Kerngedanken, daß Definitionen nicht mehr tun sollen, als die Bedeutung eines Ausdrucks festzulegen.

## 5.6 Sind zirkuläre Definitionen nötig?

Die Entscheidung für oder wider die Theorie zirkulärer Definitionen hängt nach Gupta und Belnap von der Beantwortung der beiden folgenden Fragen ab[77]:

1. Sind zirkuläre Begriffe in irgendeiner Weise nützlich?

2. Sind einige unserer natürlichsprachlichen Begriffe zirkulär?

Beide Fragen beantworten Gupta und Belnap positiv. Als Begründung für die Anwort auf die erste Frage wird die Tatsache angeführt, daß zirkuläre Definitionen die Ausdrucksstärke der Sprachen vergrößern würden. Das soll durch das Beispiel 5D.19[78] veranschaulicht werden:

> Angenommen die Sprache L enthält den Namenbuchstaben ‘0’, welches in der Struktur M durch die Null interpretiert wird, und das Funktionssymbol ‘′’, das in M durch die Nachfolgerfunktion interpretiert wird. Der Wertebereich von M sei die Menge W $= \{\ldots -3, -2, -1, 0, 1, 2, 3, \ldots\}$, also die Menge der ganzen Zahlen. Die Menge der Definitionen $\mathcal{D}$ bestehe lediglich aus der Definition (1):
>
> (1) $G(x) =_{Df} (x = 0 \vee (\exists y)(G(y) \,\&\, x = y')) \,\&\, \neg((\exists z)(G(z) \,\&\, z' = 0)$
>
> Wieder sei $\delta_{\mathcal{D},M}$ die zur Definition (1) und zur Struktur gehörige Revisionsregel. In dem im 5. Kapitel definierten semantischen System $S^{\#}$ wird durch die Definition (1) im Modell M die Menge der natürlichen Zahlen definiert. Bekanntermaßen ist aber die Menge der natürlichen Zahlen in der Quantorenlogik erster Stufe nicht allein mit Hilfe der Zeichen ‘0’ und ‘′’ definierbar[79] – eine Ausdrucksschwäche, die mit der Revisionstheorie der Definition behoben wird. Welche Bedeutung hat nun konkret die Vergrößerung der Ausdrucksstärke für die Praxis? Mit (RTD) bedarf es keines Übergangs zur komplizierten Logik zweiter Stufe, in denen nicht so geordnete Verhältnisse herrschen wie in der Logik erster Stufe. Womit allerdings

[77] [18], S.129
[78] [18], S.190
[79] Allerdings läßt sich das in der Logik zweiter Stufe bewerkstelligen.

erkauft man sich den Verbleib in der Logik erster Stufe? Durch nichts anderes als einer Theorie der Definition, die mit einer Semantik aufwartet, für deren Entwicklung auf die Theorie der Ordinalzahlen aus der Mengenlehre zurückgegriffen wird – und das ist nicht minder kompliziert. Wenn schon auf die Mengenlehre zurückgegriffen wird, dann ließe sich auch der Schritt rechtfertigen, die natürlichen Zahlen in der Mengenlehre, die sich in der Logik erster Stufe entwickeln läßt, zu definieren, indem man für die natürlichen Zahlen und die Nachfolgerfunktion in der Mengenlehre geeignete Darstellungen findet. Obwohl man damit doch eine gewisse Ausdrucksstärke der Logik erster Stufe nachgewiesen hätte, hätte man es dann jedoch mit mengentheoretischen Abbildern der natürlichen Zahlen zu tun: Die Null wäre dann z.B. die leere Menge $\emptyset$, die Eins die Einermenge $\{\emptyset\}$, die Zwei die Menge $\{\emptyset, \{\emptyset\}\}$ usw. Und schließlich würde die Nachfolgerfunktion $'$ über $n' := n \cup \{n\}$ definiert werden. Bei der Revisionstheorie greift man zwar auf die Mengenlehre zurück, geht aber nicht den Weg, die natürlichen Zahlen als Gebilde ansehen zu müssen, die im Prinzip nicht mehr sind als Mengen, die die leere Menge enthalten.

Zur Begründung der Antwort auf die zweite Frage wird der Wahrheitsbegriff angeführt. Auf andere Begriffe der Umgangssprache, die sich als zirkulär herausstellen könnten, wird mit einer kurzen Erläuterung hingewiesen. Ich werde den „Nachweis“ für den Wahrheitsbegriff eingehender betrachten.

# 6 Der Begriff der Zirkularität bei Gupta und Belnap

Angesichts der Tatsache, daß es mit einer der Hauptziele der RTT ist, nachzuweisen, daß der Wahrheitsbegriff der ordinary language[80] ein zirkulärer ist, muß es ein wenig erstaunen, daß sich nirgendwo im Buch RTT eine explizite Besprechung des allgemeinen *Phänomens der Zirkularität* finden läßt. Noch erstaunlicher aber ist, daß man mit seinen Intuitionen darüber, was ein *zirkulärer Begriff* ist, allein gelassen wird und keine befriedigende Verständigung über diesen stattfindet.

Ein intuitives Verständnis des Ausdrucks 'zirkulärer Begriff' scheinen Belnap und Gupta an den Tag zu legen, wenn sie zu verschiedenen Gelegenheiten sagen, das Prädikat G drücke einen zirkulären Begriff aus, da man zur Bestimmung derjenigen Dinge, die G erfüllen, eine Hypothese bzgl. der Extension von G treffen müsse.[81]

Wie sieht ihre Explikation des Ausdrucks 'Zirkularität eine Begriffs' im Rahmen von (RTD) aus? Auch hier bleiben Belnap und Gupta dem Leser eine Antwort schuldig. Zumindest geben Gupta und Belnap in ihrer Antwort [19] zur Rezension ihres Buches durch Robert Koons [23] eine hinreichende Bedingung für die Zirkularität eines Begriffs G an, welches in einer Sprache L durch: $x$ ist $G =_{Df} A(x, G)$ definiert wird:

(1) Enthält das Definiens „A(x,G)" „G" wesentlich und ist es intensional adäquat bzgl. des Definiendums „x ist G", dann ist G ein zirkulärer Begriff.[82]

Um diese hinreichende Bedingung (1) zu verstehen, muß geklärt werden, was die Ausdrücke 'intensional adäquat' und 'G kommt wesentlich in 'A(x,G)' vor' bedeuten.

Zuvor will ich aber kurz andeuten, wie der Ausdruck 'Begriff' („concept") zu verstehen ist. Da Belnap und Gupta in RTT nicht explizit angeben, wie sie den Ausdruck 'concept' verstanden wissen wollen, trage ich das zusammen, was sich aus den verstreuten Bemerkungen in RTT, welche den Ausdruck 'concept' enthalten, für die Interpretation dieses Terminus ergibt: Hiernach sind Begriffe Dinge, die durch Prädikate (auch Funktionsausdrücke?) ausgedrückt werden

[80] Genauer muß man sagen, daß es der logische, schwache, absolute Wahrheitsbegriff der natürlichen Sprachen ist, dessen Zirkularität nachgewiesen werden soll. Diese Einschränkung wird im Abschnitt zur Wahrheit besprochen.

[81] Siehe z.B. [18], S. 255. An dieser Stelle schreiben sie, daß der Begriff „kategorisch in L" zirkulär ist, wenn man annimt, daß er in L enthalten ist. „On the contrary, we should observe that this notion is just as circular as truth: We can determine the the categorical sentences of L *only on the basis of a prior hypothesis concerning the extension of „categorical in L"*."

[82] „If the definiens „A(x, G)" contains „G" essentially and is intensionally adequate to the definiendum „x is G" then G is a circular concept." ([19], S.634) Bis auf die äußeren Anführungsstriche sind die Anführungszeichen genau so zitiert worden, wie es im Orginal zu finden ist. Die noch in ihrem Buch RTT zu findende Unterscheidung zwischen einfachen und doppelten Anführungszeichen geht in ihrem Anwortschreiben [19] verloren.

können. Z.B. drückt das Prädikat ‘ist wahr’ der deutschen Sprache den Begriff „Wahrheit“ für die deutsche Sprache aus. Begriffe scheinen abstrakte, objektive, nicht-mentale Entitäten zu sein. Begriffen läßt sich ähnlich Prädikaten eine Extension und eine Intension zuordnen. Liest man ‘Bedeutung’ als ‘Intension’, dann ist die Rede davon, daß Begriffe die Bedeutungen von Prädikaten (Funktionsausdrücken) sind, nicht mit dem obigen Verfahren zu vereinbaren, Begriffen eine Intension zuzuordnen. Tatsächlich sprechen Gupta und Belnap an einigen Stellen davon, daß ein Ausdruck den und den Begriff bedeutet.[83] Ich glaube, daß hier die Rede von ‘Bedeutung’ systematisch schwankt: Wenn Belnap und Gupta davon reden, daß die Bedeutung eines Prädikats ein Begriff ist, dann heißt ‘Bedeutung’ mehr als ‘Intension’. Manchmal reden Belnap und Gupta auch von der Bedeutung von Begriffen.[84] In diesem Falle ist ‘Bedeutung’ dann vermutlich als ‘Intension’ zu lesen. Die Individuierung von Begriffen gemäß Belnap und Gupta geschieht auf einer höheren Ebene als der der Intension, d.h. Begriffe A und B sind nicht schon dann ein und derselbe Begriff, wenn sie dieselbe Intension haben. Man muß also vorsichtig sein, wenn man mit Belnap und Gupta die Rede übernimmt, daß eine nicht-stipulative Definition einen Begriff definiert: Damit ist nicht gemeint, daß mit der Definition eineindeutig der Begriff fixiert wird – wie auch immer die Fixierung eines Begriffes durch eine Definition aussehen mag – sondern lediglich, daß die Intension eines Begriffes fixiert wird. Wenn man sich auf diese Weise verständigt hat, was es heißen soll, daß eine nicht-stipulative Definition einen Begriff definiert, dann sollte man sich auch darauf verständigen, was es heißt, ein bereits gängiges Prädikat in einer nicht-stipulativen Definition zu definieren – in jenem schwachen Sinne wie Gupta und Belnap es wünschen. Ich vermute, daß es nicht mehr heißt, als daß die Intension des Prädikats festgelegt wird. Wenn man hier also sagt, eine Definition würde vollständig die Bedeutung eines Prädikats fixieren, dann ist gemeint, daß die Intension des Prädikats fixiert wird.

## 6.1 Adäquatheit vs. Äquivalenz

Der Begriff der intensionalen Adäquatheit macht erst im Rahmen von nicht-stipulativen Definitionen, d.h. von Definitionen, die einen in einer bestimmten Sprachgemeinschaft schon vorhandenen Ausdruck zum Gegenstand definitorischer Erläuterung machen, einen Sinn. Sie kann als prüfende Instanz für nicht-stipulative Definitionen angesehen werden, deren Maßstab durch den Terminus ‘intensional’ vorgegeben wird. Ganz allgemein – also sowohl für zirkuläre als auch nicht-zirkuläre Begriffe – würde die Prüfung auf intensionale Adäquatheit der Definition eines bereits im Gebrauch befindlichen Prädikats G daraus hinauslaufen, die Definition zunächst als stipulative Definition eines Prädikats G’ anzusehen

[83] Siehe etwa [18], S.86

[84] So zumindest verstehe ich das, wenn sie z.B. in [18], S.25 von der „meaning of truth“ sprechen.

und dann zu überprüfen, ob G' und G intensional äquivalent sind.[85] Für klassische nichtzirkuläre Definitionen läßt sich der Begriff der intensionalen Äquivalenz Carnap gemäß folgendermaßen erklären[86]:

> Zwei einstellige Prädikate $G$ und $F$ sind intensional äquivalent genau dann, wenn es notwendigerweise der Fall ist, daß gilt: $(\forall x)(G(x) \leftrightarrow F(x))$.

Für zirkuläre Definitionen ergibt sich nicht nur, daß diese beiden Begriffe divergieren, sondern auch, daß der Begriff der 'intensionalen Äquivalenz' jetzt nicht wie üblich erklärt werden kann. Es muß eine Erklärung im Rahmen der Revisionssemantik gegeben werden, und diese eröffnet die Möglichkeit, verschiedene Grade von intensionaler Äquivalenz zu unterscheiden – eine Möglichkeit, die es im Falle nicht-zirkulärer Definitionen offensichtlich nicht gibt.

Gupta und Belnap betrachten hierzu folgendes Beispiel (hier allerdings) für extensionale Adäquatheit[87]. Angenommen eine Sprachgemeinschaft führt stipulativ durch (A) ein Prädikat G ein:

$$\text{(A)}\ Gx \quad =_{Df} \quad Fx \vee (Ga\ \&\ x = b) \vee (\neg Fx\ \&\ Hx\ \&\ \neg Gx)$$

Wir nehmen der Einfachheit halber an, daß die Prädikate F und H klassisch sind. Das Prädikat F habe lediglich den Gegenstand a in seiner Extension. Weiter nehmen wir an, jemand, dem die Definition (A) nicht bekannt ist, gebe seine eigene Definition des Prädikats G an:

$$\text{(B)}\ Gx \quad =_{Df} \quad (Fx \vee b = x) \vee (\neg Fx\ \&\ Hx\ \&\ \neg Gx)$$

Es stellt sich heraus, daß die durch (B) bestimmte Revisionsregel (im oben festgelegten Sinne) eine andere ist als die durch (A) bestimmte; denn für die Ausgangshypothese $\emptyset$ für G ergeben (A) und (B) unterschiedliche Werte.

Trotz dieses Unterschiedes verhalten sich die Revisionsregeln von (A) bzw. (B) aber relativ ähnlich. Nach der ersten Anwendung des Revisionsoperators liefern sie dieselben Werte. Nach einem schwächeren Standard für extensionale Äquivalenz wäre die Definition (B) also extensional adäquat, da die Einteilung der Sätze, welche G enthalten, in gültige, kategorische, pathologische, paradoxe Sätze (B) gemäß dieselbe ist wie die Einteilung, welche aus (A) für G enthaltende Sätze resultiert. Nach einem stärkeren Standard wäre (B) nicht extensional adäquat,

---

[85] [18], S.130f. Entsprechend würde die Prüfung auf extensionale Adäquatheit bzw. Adäquatheit bzgl. des Sinns verlaufen, wenn die Begriffe der extensionalen Äquivalenz bzw. kognitiven Äquivalenz gegeben wären.

[86] Extensionale Äquivalenz und kognitive Äquivalenz (für einstellige Präikate) werden klassich folgendermaßen erläutert:
Zwei Prädikate $G$ und $F$ sind extensional äquivalent genau dann, wenn gilt: $(\forall x)(G(x) \leftrightarrow F(x))$
Zwei Prädikate $G$ und $F$ sind kognitiv äquivalent genau dann, wenn gilt:
Man kann den Satz $(\forall x)(F(x) \leftrightarrow G(x))$ nicht verstehen, ohne zu glauben, daß $(\forall x)(G(x) \leftrightarrow F(x))$

[87] [18], S. 130

da die Revisionsregeln[88] von (B) und (A) nicht identisch sind.
Man kann sich neben diesen beiden Standards durchaus weitere vorstellen, für die es sicher auch gute Motivationen geben mag.[89]

Machen wir uns die Begrifflichkeit an einem weiteren Beispiel klar. Angenommen jemand gibt folgende Definition für das Prädikat 'ist ein Junggeselle', mit dem Anspruch, hiermit das im Deutschen verwendete Wort Junggeselle zu definieren:

(∗) x ist ein Junggeselle $=_{Df}$ x ist ein Junggeselle.

Die Definition (∗) ist im Rahmen von (RTD) formal absolut korrekt; Zirkularität wie in diesem Extremfall ist gestattet. Allerdings ist die Definition nicht extensional adäquat. Weshalb nicht? Wir denken uns die Definition (∗) als stipulative Definition eines Prädikats 'ist (∗)-Junggeselle' und überprüfen es auf extensionale Äquivalenz mit unserem Prädikat 'ist ein Junggeselle'. Wie wir gesehen haben, gibt es verschiedene Standards, die man an die extensionale Äquivalenz anlegen kann. Gehen wir von der schwächsten aus, so muß durch (∗) zumindest dieselbe Einordnung von Sätzen in die semantischen Sparten „kategorisch-wahr (-falsch)"„paradox" und „nicht-kategorisch" gewährleistet sein. Nun wird man doch ohne Wenn und Aber der Behauptung 'Kant ist Junggeselle' beipflichten wollen. Wäre (∗) eine extensional adäquate Definition, so müßte sie zumindest eine Revisionsregel festsetzen, die in der oben angegebenen Semantik den Satz 'Kant ist (∗)-Junggeselle' als kategorisch-wahr herausstellt. Das ist aber nicht der Fall: Je nach dem, ob die Ausgangshypothese für die Extension des definierten Prädikats 'ist (∗)-Junggeselle' die Person Kant enthält oder nicht, ergibt sich in den folgenden Revisionsschritten immer nur die Wahrheit des Satzes 'Kant ist (∗)-Junggeselle' oder immer nur deren Falschheit. Obwohl die Definition (∗) zirkulär, aber eben nicht inhaltlich adäquat ist, ist das Prädikat 'ist Junggeselle' auch nicht als ein zirkuläres ausgewiesen. Was dieses Beispiel u.a. zeigt ist, daß der Begriff der extensionalen Adäquatheit vom Begriff der extensionalen Äquivalenz

[88]Dieser Unterschied besteht auch dann noch, wenn wir die Revisionsregeln, welche Funktionen sind, als lediglich durch ihre Graphen individuiert ansehen.

[89]Das Beispiel ist hier ein sehr ideales Beispiel, bei dem wir auf eine Sprachgemeinschaft stoßen, die ein Prädikat G stipulativ einführt. Wenn es sich nicht zufälligerweise um eine Gruppe von Mathematikern handelt, dürfte man i.A. eher selten auf stipulative Definitionen treffen. Der weitaus häufigere und philosophisch interessantere Fall ist der eines im Gebrauch befindlichen Prädikats, welches nicht durch eine stipulative Definitionen eingeführt wurde. Für das Wahrheitsprädikat liegt ein derartiger Idealfall nicht vor. Wir können eine zirkuläre Definition für ein Wahrheitsprädikat, nennen wir es T-Wahrheit, anführen. Wie allerdings prüfen wir die extensionale oder intensionale Adäquatheit? Dazu müßten wir die extensionale bzw. intensionale Äquivalenz der T- Wahrheit mit dem Wahrheitsprädikat etwa des Deutschen vergleichen; aber für dieses haben wir keine Revisionsregel vorgegeben. Tatsächlich werden Belnap und Gupta hier eine nicht weiter begründete These, die Signifikationsthese, anführen, die besagt, daß die durch die zirkuläre Definition definierte T-Wahrheit intensional-äquivalent ist mit dem Wahrheitsbegriff der Umgangssprache.

von Definiendum und Definiens im Rahmen der Theorie zirkulärer Definitionen abweicht: (∗) gibt zwar trivialerweise ein extensional gleiches Definiendum und Definiens aus, ist aber mitnichten extensional adäquat.

Hat man kein vernünftiges Kriterium dafür, welchen Standard für extensionale und intensionale Äquivalenz man zur Anwendung kommen lassen will, dann ergibt sich das Problem, daß man man im Grunde nur eine notwendige Bedingung und eine hiervon verschiedene hinreichende Bedingung für (extensionale/intensionale) Adäquatheit hat: Ergibt sich, daß die in Frage stehende Definition nicht einmal gemäß des schwächsten Standards extensional adäquat ist, dann ist sie nicht extensional adäquat. Alles dazwischen bleibt unter der Annahme, es gebe kein vernünftiges Kriterium, eine Erneuerungsfrage von Fall zu Fall. Das wirkt sich auch auf die Frage nach der Zirkularität eines Begriffes aus: Kann ich intensionale Adäquatheit nicht im stärksten Sinne zeigen, so ist die Frage nach der Zirkularität noch offen.

## 6.2 Wesentliche Zirkularität

Weshalb ist in der hinreichenden Bedingung (1) für die Zirkularität eines Begriffes der Zusatz „wesentlich“ nötig? Betrachten wir folgende Definition des Prädikats ‘ist eine Primzahl’:

(P) x ist eine Primzahl $\Leftrightarrow_{Df}$ ( x ist nur durch sich und 1 ohne Rest teilbar) & (x ist eine Primzahl $\vee$ $\neg$(x ist eine Primzahl))

Offensichtlich ist die Definition (P) intensional adäquat: Mit ihr werden in allen möglichen Welten die und nur die Gegenstände als Primzahlen ausgewiesen, die wir in der Mathematik als Primzahlen kennengelernt haben. Dennoch wird nach Belnap und Gupta durch diese Definition kein zirkulärer Begriff definiert: Der durch diese Definition festgesetzte Begriff des Primzahlseins ist kein zirkulärer. Wie sieht genau die Begründung hierfür aus?

Die Diagnose für (P) fällt leicht, da man sieht, daß das Prädikat ‘ist eine Primzahl’ im Definiens insofern unwesentlich vorkommt, als das Definiens logisch äquivalent zu einer Formel ist, die das Prädikat nicht enthält, nämlich ‘x ist nur duch sich und die 1 ohne Rest teilbar’. In der Terminologie des ersten Kapitels formuliert: Das Vorkommnis des Definiendums ‘ist eine Primzahl’ im Definiens ist l-harmlos. Diese Begründung scheinen auch Gupta und Belnap zu intendieren, wenn sie schreiben: „The reason is that circularity, though present, is eliminable and inessential ..."([19], S.635)

Es wäre durchaus aber auch möglich gewesen, die nicht wesentliche Zirkularität der Definition folgendermaßen zu begründen: Durch (P) wird dem Prädikat ‘ist eine Primzahl’ in allen möglichen Welten eine definite Extension zugewiesen wird; sie besteht gerade aus denjenigen Gegenständen a, für die die Aussage ‘a ist eine Primzahl’ kategorisch-wahr ist. Da ‘a ist eine Primzahl’ für alle Namen

'a' nach der Definition (P) kategorisch ist, kann man eindeutig sagen, ob ein Gegenstand in die Extension des Prädikats 'ist eine Primzahl' gehört oder nicht. Es gibt also keine pathologischen Fälle.

Wir haben schon bei der Besprechung der Definitionskriterien im ersten Teil der Arbeit gesehen, daß es Beispiele für Definitionen gibt, die keine l-harmlose Vorkommnisse des Definiendums enthalten, obwohl die Vorkommnisse dennoch in einem anderen Sinne harmlos sind: wir hatten ein Beispiel für eine zirkuläre Definitionen mit einem Definiendum 'Gx', das nicht zu einer Formel äquivalent ist, welches 'G' nicht enthält, obwohl sich durch die Definition für 'G' eine eindeutige Extension ergab. In den semantischen Systemen der Revisionstheorie würde sich für solch eine Definition ergeben, daß sie eine Revisionsfunktion festlegt, nach der jeder Satz 'a ist G' sich als kategorisch herausstellt - was also bedeuten würde, daß es keine pathologischen Fälle gibt. Wir hatten aber auch gesehen, daß es für diese Definition eine andere Definition gibt, die sich als logisch äquivalent zu der ursprünglichen herausstellt und die kein Vorkommnis des Definiendums enthält. Wir hätten für diese Definition also auch die Möglichkeit, das nicht wesentliche Vorkommen des Definiendums dadurch zu begründen, daß wir auf die Existenz einer nichtzirkulären äquivalenten Definition verweisen. Tatsächlich scheinen Gupta und Belnap diese letztere Begründung zu intendieren, wenn sie für ein ähnliches Beispiel in 5A.11 ([18], S.151) die bereits oben zitierte Begründung angeben, daß die Zirkularität eliminierbar und unwesentlich ist.[90]

Für induktive Definitionen gibt es keine logisch äquivalenten Definitionen mehr, die ein G-freies Definiens besitzen. Aber auch für induktive Definitionen ist das Vorkommnis des Definiendums im Definiens nicht wesentlich und somit der in solchen Definitionen festgelegte Begriff kein zirkulärer. Wie sieht hierfür die Begründung aus? Im ersten Teil der Arbeit hatte ich bereits angedeutet, wie hier die Begründung lauten könnte. Tatsächlich scheinen Gupta und Belnap diese Begründung angeben zu wollen, wenn sie schreiben:

> Inductive definitions are not strictly circular; the apparent circularity in them is eliminable through higher-order quantification in the following well-known way. Suppose for simplicitiy that in the definition
>
> (2) $G(x_1, \ldots, x_n) =_{Df} A_G(x_1, \ldots, x_n)$,
>
> G is the only definiendum occurring in $A_G$. Let $A_H$ be obtained from $A_G$ by replacing all occurrences of G in it by an n-place predicate variable H. Construed inductively, (2) is equivalent to the definition

[90] Die in 5A.11 angegebene Definition lautet:
$G(x) =_{Df} [x = a \;\&\; (G(a) \vee G(b)] \vee [x = b \;\&\; \neg G(a) \;\&\; \neg G(b)]$ Das Definiens dieser Definition ist mit keiner G-freien Formel logisch äquivalent . Nichtsdestotrotz ist in den Systemen $S_n$ diese Definition regulär, d.h. daß diese Definition dem definierten Prädikat G eine definite Extensionen zuweist.

$$G(x_1, \ldots, x_n) =_{Df} (\forall H)[(\forall x_1) \ldots (\forall x_n)(H(x_1, \ldots, x_n) \leftrightarrow A_H(x_1, \ldots, x_n)) \rightarrow H(x_1, \ldots, x_n)],$$

> which is noncircular. Hence, inductive definitions ascribe definite extensions to their definienda; they do not result in any pathological behavior. ([18], S.194)

Für induktive Definitionen gibt es zwar keine logisch äquivalente, nicht-zirkuläre Definition in der Logik erster Stufe, aber dafür eine „äquivalente“ nicht-zirkuläre Definition in einer Logik, die höherstufige Quantifikation gestattet.

Man beachte hier, daß zu einer induktiven Definition in der Logik erster Stufe zu dem definierenden Satz selbst noch immer der Hinweis gehört, daß der kleinste Fixpunkt zu wählen ist – ein Hinweis, der im definierenden Satz in der Logik erster Stufe nicht ausgedrückt werden kann. Wendet man auf den definierenden Satz einer induktiven Definition in der Logik erster Stufe die Revisionssemantik an, dann ergeben sich pathologische Fälle: Wurde etwa ein einstelliges Prädikat ‘G’ in der Logik erster Stufe induktiv definiert, dann gibt es einen Gegenstand a, für den der Satz ‘Ga’ nicht kategorisch ist. So betrachtet würde der definierende Satz einer induktiven Definition in der Logik erster Stufe ohne den Hinweis, daß der kleinste Fixpunkt zu wählen ist, einen Begriff definieren, der zirkulär ist. Nun gehört zu einer induktiven Definition immer auch dazu, daß der kleinste Fixpunkt zu wählen ist; das läßt sich nur in der Logik zweiter Stufe korrekt wiedergeben: Da aber ist die Definition nicht mehr zirkulär, also wird durch eine induktive Definition auch kein zirkulärer Begriff definiert.

Da in den Begründungen von Gupta und Belnap dafür, daß eine Definition nicht wesentlich zirkulär ist, auch die Tatsache erwähnt wird, daß es keine pathologischen Fälle gibt, liegt die Versuchung nahe, folgende hinreichende und notwendige Bedingung für die wesentliche Zirkularität einer Definition aufzustellen:

> **Wesentliche Zirkularität**
> Eine Definition eines einstelligen Prädikats G ist genau dann wesentlich zirkulär, wenn es in einer möglichen Welt (repräsentiert durch ein klassisches Modell) einen Gegenstand a gibt, für den die Aussage ‘Ga’ nicht kategorisch wahr ist.[91]

Zunächst bemerkt man, daß in diesem Kriterium durch den Begriff ‘kategorisch’ eine Abhängigkeit vom semantischen System gegeben ist: Theoretisch ist damit die Möglichkeit nicht ausgeschlossen, daß sich gemäß einer revisionstheoretischen Semantik S1 keine pathologischen Fälle ergeben, gemäß einer anderen S2 aber schon. Das wiederum hätte zur Folge, daß eine Definition in einem semantischen System als wesentlich zirkulär ausgewiesen wird, in dem anderen hingen als nicht

[91] Dieses Bikonditional ist für den sehr idealisierten Fall formuliert, daß nur eine Definition vorliegt, nämlich die des einstelligen Prädikats G.

wesentlich zirkulär. Die Hoffnung der Revisionstheoretiker ist es natürlich, das *eine* korrekte, adäquate semantische System zu finden, so daß 'kategorisch' nur gemäß dieses semantischen Systems verstanden und damit eine drohende Divergenz vermieden wird.

Die eine Richtung des Bikonditionals, nämlich die Richtung von rechts nach links, dürfte intuitiv klar sein. Wenn es für eine Definition einen Gegenstand a gibt, für den die Aussage 'Ga' nicht kategorisch ist, dann kann das nur der Fall sein, wenn im Definiens das Definiendum G enthalten ist. Die Definition ist dann auf jeden Fall zirkulär, und diese Zirkularität ist fatal, da sie für den pathologischen Fall a verantwortlich gemacht werden kann; also ist die Definition wesentlich zirkulär.

Wie steht es allerdings mit der umgekehrten Richtung? Ist das Vorhandensein von pathologischen Fällen notwendig für die wesentliche Zirkularität einer Definition? Anders gefragt: Kann die Zirkularität in einer Definition auch wesentlich sein, ohne daß es pathologische Fälle zu geben braucht? Belnap und Gupta besprechen keinen derartigen Fall, was die Vermutung nahelegt, daß sie Pathologizität notwendig für wesentliche Zirkularität einer Definition halten. Eine dahingehende explizite Behauptung findet sich nicht. Es gibt allerdings einige Textstellen, die darauf schließen lassen, daß Belnap und Gupta Pathologizität in mindestens einer möglichen Welt als notwendig für Zirkularität erachten.[92] Dies läßt sich auch plausibilisieren, wenn man versucht zu erklären, was es sonst heißen könnte, daß eine zirkuläre Definition wesentlich zirkulär ist (s.u.).

Eine explizite Behauptung dafür, daß Pathologizität nicht notwendig für Zirkularität ist, findet man bei Aladdin M. Yaqub in seinem Buch [50] gemacht, allerdings betrifft es dort nicht Definitionen, sondern Begriffe. Yaqub meint, daß der Begriff der Wahrheit auch dann zirkulär wäre, wenn es keine problematischen Fälle wie den Lügnersatz oder den Wahrsagersatz geben würde.

Die Zirkularität eines Begriffs erläutert Yaqub folgendermaßen:

> We first need to explain, if only crudely, what it means to say that some concept C is circular. In a broad sense, C is circular if there is a nonempty collection of objects such that for each object o in this collection, the conditions under which o is subsumed under C involve *ultimately* reference to C itself. This broad sense includes the stronger notion of circularity whereby such conditions involve ultimately the condition that o is subsumed under C or that o is excluded from C. ([50], S. 36)

Was es genau heißt, daß die Bedingungen dafür, daß ein Objekt o unter einen Begriff C subsumiert wird, einen *letzendlichen (ultimativen) Bezug* auf den Begriff C involvieren, erfährt der Leser nicht.

Yaqubs so erklärter Ausdruck 'Zirkularität eines Begriffs' scheint die Folge einer kanonischen Übertragung des Ausdrucks 'Zirkularität einer Definition' zu

[92] Z.B. [18], S.117

sein. Im ersten Teil meiner Arbeit hatte ich Humberstones ganz ähnlich lautende Erklärung für die Zirkularität einer Definition vorgestellt. Yaqubs Vorschlag, die Zirkularität eines Begriffs wie im obigen Zitat zu erläutern, scheint daher der naheliegendste Weg zu sein, die üblichen Intuitionen wiederzugeben.

Als ein konkretes Beispiel führt Yaqub die folgenden beiden Sätze an, die zusammen genommen den bivalenten Charakter des Wahrheitsbegriffs ausdrücken:

NC. Kein Satz ist sowohl wahr als auch falsch.
EM. Jeder Satz ist wahr oder falsch.

Jedes dieser beiden Sätze würde bereits als Nachweis dafür, daß der Begriff der Wahrheit zirkulär ist, ausreichen; dafür sei kein Bezug auf problematische Sätze wie den Lügnersatz nötig – und NC sowie EM seien offensichtlich nicht als pathologische Fälle anzusehen. Yaqub gibt hierzu an, wie solche gesetzartigen Aussagen über den Wahrheitsbegriff üblicherweise gelesen werden. Bei beiden Lesarten, welche Yaqub anführt, stelle sich heraus, daß ein ultimativer Bezug auf den Wahrheitsbegriff involviert sei.[93]

Nach der ersten Lesart ist NC eine Abkürzung oder ein Repräsentant einer unendlichen Konjunktion von Sätzen der Form ⌜ Es ist nicht der Fall, daß gilt: p und nicht-p ⌝, wobei 'p' Platzhalter für einen Satz ist. Die Wahrheitsbedingungen von NC sind also durch sämtliche Sätze dieser Form gegeben. Da aber NC selbst als zulässiger Wert für 'p' eingesetzt werden kann, beinhalten die Wahrheitsbedingungen von NC ultimativ das, was NC besagt. NC aber enthält wieder den Bezug auf den Wahrheitsbegriff. Man gerät folglich in eine unendliche Schleife, in der man immer wieder auf NC, der einen Bezug auf den Wahrheitsbegriff beinhaltet, verwiesen wird.

Nach der zweiten von Yaqub favorisierten Lesart hat man NC (und EM) als eine Beschreibung bestimmter Charakteristiken der Extension des Wahrheitsprädikats bzw. des Begriffs der Wahrheit aufzufassen: NC etwa behauptet, daß die Extension und die Antiextension des Wahrheitsprädikats disjunkt sind. Auch hier sei ein wesentlicher Bezug auf den Wahrheitsbegriff involviert.

Ich glaube, daß man Yaqub folgendermaßen verstehen muß, wenn er sagt, daß Pathologizität nicht notwendig für die Zirkularität eines Begriffs ist: In unserer Welt (neben den anderen möglichen Welt) braucht es keine pathologischen Sätze zu geben, damit ein Begriff als zirkulär gelten kann. Eine Ansicht, die Belnap und Gupta auf jeden Fall teilen würden. Daraus folgt aber nicht, daß es auch Begriffe geben kann, die in keiner möglichen Welt pathologische Fälle aufweisen.

Worin bestünde denn noch die echte Zirkularität eines solchen Begriffs, welches ich für diese Zwecke 'np-zirkulär' ('nicht-pathologisch-zirkulär') nenne, wenn es in keiner möglichen Welt pathologische Fälle besäße? Übertragen wir die Frage auf Definitionen und stellen uns hier die Frage: Wodurch wären echt zirkuläre Definitionen, die einen np-zirkulären Begriff definieren, in der revisionstheoretischen

[93]Yaqub gibt die beiden Lesarten für NC an; entsprechendes läßt sich zu EM sagen.

Semantik zu charakterisieren? Wie würde sich hier also darstellen, daß eine Definition wesentlich zirkulär ist, ohne daß sie pathologische Fälle aufweist? Hierüber kann ich nur spekulieren; auszuschließen ist der Fall, daß genau diejenigen zirkulären Definitionen ein np-zirkuläres Prädkat G definieren, die folgendes erfüllen: a) alle Gegenstände (des Gegenstandsbereichs) erfüllen G kategorisch und b) es gibt keine Definition in der ersten Stufe, die nichtzirkulär ist und die ein Prädikat G' definiert, welches intensional äquivalent mit G gemäß des schwächsten Standards ist. Denn so gelesen würden auch die natürliche Zahl ein zirkulärer Begriff sein. Das folgt aus dem oben besprochenen Beispiel im Abschnitt 'Sind zirkuläre Definitionen nötig'. Dann muß vielleicht Klausel b) abgeändert werden, so daß sie besagt, daß es nicht einmal in der Logik zweiter (dritter, vierter, ... ?) Stufe eine nichtzirkuläre Definition gibt, die ein Prädikat G' definiert, welches gemäß des schwächsten Standards intensional äquivalent mit G ist. Wie das allerdings formal genau auszubuchstabieren ist, da doch über verschiedene Logiken hinweg np-Zirkularität erklärt werden soll, weiß ich nicht zu sagen. Ich sehe das aber auch nicht als ein Zeichen dafür an, daß (RTD) nicht die Intuitionen zur Zirkularität eines Begriffs wiederzugeben vermag – sofern wir denn überhaupt über solche verfügen sollten. Die vage Rede von dem „ultimativen Bezug" oder die Rede davon, daß man eine Hypothese bzgl. der Extension eines zirkulären Begriffs aufstellen muß, kann ersetzt werden durch die von der Pathologizität – im technischen durch (RTD) vorgegebenen Sinne. Wenn man über die Pathologizität hinaus noch von der Zirkularität eines Begriffs sprechen möchte, dann muß man das durch eine genauere technische Explikation fundieren. Ich glaube also, daß es einen keinen Grund gibt, an die Existenz np-zirkulärer Begriffe zu glauben.

Wie sieht nun eine hinreichende und notwendige Bedingung für die Zirkularität eines Begriffs aus? Der einfachste, mögliche Versuch, die bereits vorliegende hinreichende Bedingung zu ergänzen, bestünde in folgendem:

> Ein durch ein Prädikat G ausgedrückter Begriff ist zirkulär gdw es eine Definition von G gibt, die intensional adäquat und wesentlich zirkulär ist.

Diese hinreichende und notwendige Bedingung ist aber sehr unvollständig: Definitionen werden Gupta und Belnap gemäß für bestimmte Sprachen L[94] aufgestellt und sind, wenn auch nicht Sätze dieser Sprache, so doch – bis auf das Definitionszeichen $=_{Df}$ – mit Hilfe der logischen und nichtlogischen Zeichen des Vokabulars von L formuliert: Wo also bleibt der Bezug zu einer solchen konkreten Sprache L?

Bei Gupta und Belnap können Definitionen in einem Modell M einen Revisionsoperator festlegen. Im Speziallfall eines Prädikats G und einer Definition dieses Prädikats wird dieser nämlich so erklärt, daß er für jede Input-Teilmenge des Wertebereichs in einem Modell als Output diejenige Teilmenge des Wertebereichs ausgibt, deren Elemente das Definiens – unter der Interpretation von

[94]Der Einfachheit halber nehme ich an, daß L eine Sprache der klassischen Quantorenlogik ist

G mit der Inputteilmenge – erfüllen. Tatsächlich können aber solche Operatoren ein definitionsunabhängiges Eigenleben führen. Die Zirkularität eines Begriffes G müßte dann eher mit Hilfe dieser Operatoren erläutert werden. Dazu erklären Gupta und Belnap nach Besprechung der semantischen System $\mathbf{S}^{\#}$ und $\mathbf{S}^{*}$:

> The semantical schemes $\mathbf{S}^{\#}$ and $\mathbf{S}^{*}$ can be applied to languages **L** that contain predicates governed by rules of revision but in which no *definition* is formulable for these predicates. A simple kind of case is this: A predicate G may be introduced into **L** through the stipulation that it has the same meaning as the predicate H in another language **L'**, where H is given (in **L'**) a circular definition that is not formulable in **L**. ([18], S. 196-197)

Wenn es also in einer Sprache aufgrund ihrer Ausdrucksschwäche keine (nicht einmal) zirkuläre Definition für ein Prädikat geben sollte, ist damit nicht ausgeschlossen, daß dieses Prädikat einen zirkulären Begriff ausdrückt. Ein Beispiel geben Gupta und Belnap in der oben zitierten Passage an: Der durch das einstellige Prädikat G ausgedrückte Begriff aus **L** läßt sich nicht zirkulär in **L** definieren, ist aber dennoch ein zirkulärer Begriff, da das Prädikat G per Stipulation dieselbe Bedeutung wie ein Prädikat H aus einer Sprache **L'** besitzt, welches in **L'** durch eine wesentlich zirkuläre Definition definiert ist und damit nach der hinreichenden Bedingung einen zirkulären Begriff ausdrückt. Wie hat man sich die Stipulation vorzustellen und was meint die Bedeutungsgleichheit? Die Stipulation muß in einer metasprachlichen Definition erfolgen, da man schließlich mit Prädikaten aus zwei verschiedenen Sprachen **L** und **L'** arbeitet. Die in der Stipulation festgesetzte Bedeutungsgleichheit meint vermutlich die Intensionsgleichheit der Prädikate G und H im stärksten Sinne, d.h. daß für G der Revisionsoperator bestimmend ist, welcher auch das Verhalten von H bestimmt. Bei diesem Beispiel wird letztlich wieder auf zirkuläre Definitionen rekurriert, wenn auch auf Definitionen von anderen Prädikaten als dem, der jenen Begriff ausdrückt, dessen Zirkularität zu zeigen ist. Das Beispiel legt folgende Modifikation der obigen hinreichenden und notwendigen Bedingung für die Zirkularität eines Prädikats einer Sprache L nahe:

> Ein Prädikat G einer Sprache **L** drückt einen zirkulären Begriff aus gdw es mit einem Prädikat H einer Sprache **L'** intensional äquivalent ist, welches über eine stipulative wesentlich zirkuläre Definition in **L'** definiert wurde.

Ein Problem bleibt immer noch mit dieser hinreichenden und notwendigen Bedingung. Hier wird von der einen Intensionsgleichheit und von der einen intensionalen Adäquatheit gesprochen. Wir haben gesehen, daß man in der Revisionssemantik verschiedene Grade von intensionaler Äquivalenz bzw. Adäquatheit ausmachen kann. Der schwächste Standard ist hinreichend, sofern man hier nicht von np-zirkulären Begriffen reden möchte. Denn im schwächsten Standard wird die Unterscheidung zwischen kategorischen und nicht-kategorischen Fällen aufrechterhalten wird, was für die Entscheidung dafür, ob ein Begriff zirkulär ist

oder nicht, ausreichend ist.[95]

Ein weiteres Problem ist natürlich, daß hier immer noch von der Existenz einer Definition (wenn auch einer beliebigen Sprache) ausgegangen wird. Kann es nicht auch zirkuläre Begriffe geben, obwohl die in keiner bekannten Sprache definiert werden können? Auch wenn wir zirkuläre Begriffe kennengelernt haben über zirkuläre Definitionen, muß das nicht bedeuten, daß es für sie eine tatsächlich vorhandene Sprache gibt, in der man zirkulär ein Prädikat definieren kann, das sie ausdrückt. Ich glaube daher, daß die unverfänglichste Rede von der Zirkularität eines Begriffs die folgende ist:

> **Zirkularität eines Begriffs**
> Ein Begriff C ist zirkulär, wenn es eine mögliche Welt M gibt, in der die „Extension" von C durch einen solchen Revisionsoperator $\delta_M$ korrekt wiedergegeben wird, für den es Dinge a gibt, so daß Ca nicht kategorisch wahr in M ist.

Allerdings ist das bestimmt nicht als ein gutes Kriterium anzusehen, mit dem sich die Zirkularität eines Begriffs leicht feststellen ließe.

[95] Für np-zirkuläre Begriffe bin ich mir nicht im Klaren darüber, welcher Standard der angemessene ist, da ich mir nicht sicher bin, wie man solche Begriffe bzw. die Definitionen, mit denen sie definiert werden, in (RTD) charakterisieren könnte.

# 7 Ausbau der Semantik: $S^*$ und $S^\#$

Dieser Abschnitt bespricht, weshalb das bei der Darstellung der Revisionstheorie benutzte semantische System nicht den Zwecken genügt, die Gupta und Belnap von einer Definitionstheorie fordern: $S_0$ erfüllt zwar bestimmte Nicht-Kreativitätsforderungen, ordnet den Definienda aber nicht hinreichend Inhalt zu, wie es manchmal zu wünschen wäre. Auch die anderen System $S_n$ ($n > 0$), die so ähnlich wie $S_0$ formuliert sind, sind nicht befriedigend: Sie verletzten bestimmte Nicht-Kreativitätskriterien. Das veranlaßt Gupta und Belnap, semantische Systeme zu entwickeln, die beide Kriterien erfüllen. Das Ergebnis sind die revisionstheoretischen Systeme $S^*$ und $S^\#$. Ich werde die zur Definition dieser Systeme notwendigen Begriffe in kompakter Form darstellen und einige wenige Bemerkungen zu diesen machen. (Der folgende Abschnitt kann übersprungen werden. Er gewinnt lediglich Relevanz für einige wenige Einwände, die ich im kritischen Teil besprechen möchte.)

## 7.1 Logische Schwäche und Nicht-Kreativität

Im 5. Kapitel ihres Buches besprechen Gupta und Belnap mehrere semantische Systeme und Kalküle, die sich für zirkuläre Definitionen eignen könnten . Unter diesen ist auch das semantische System $S_0$ und der Kalkül $C_0$, welche weiter oben zur Veranschaulichung der Hauptideen von (RTD) dargestellt wurden. Sie sind, wie die Indizierung andeutet, eines der unendlich vielen semantischen Systeme $S_n$ und der Kalküle $C_n$. Obwohl sehr eingängig und leicht zu verstehen, halten Gupta und Belnap keines der semantischen Systeme $S_n$ für einen geeigneten Rahmen, in dem sich zirkuläre Definitionen angemessen behandeln ließen. Die Gründe für die Ablehnung von $S_0$ und $S_n$ ($n \geq 1$) sind unterschiedlich. $S_0$ wird abgelehnt, da es – obwohl stark nicht-kreativ – logisch schwach ist: Es ordnet den Definienda einer fest vorgegebenen Menge an Definitionen weniger Inhalt zu als zu wünschen wäre. Die Systeme $S_n$ ($n \geq 1$) werden abgelehnt, da sie – obwohl logisch stark – nicht einmal schwach nicht-kreativ sind. Um die Begründung verständlich zu machen, sind die Ausdrücke ‘ist schwach/stark nicht-kreativ’ und ‘ist (logisch) schwach’ zu klären.

### 7.1.1 Schwach und stark nicht-kreativ

Ich gebe zunächst eine informelle Erläuterung dieser Begriffe, um danach kurz zu erklären, wie die semantischen System $S_n$ definiert sind, und schließlich anzugeben, was Nicht-Kreativität für diese semantischen Systeme bedeutet.

Ein System $S_n$ ist stark nicht-kreativ, wenn alle Sätze A, die keines der Definienda enthalten und die gemäß der durch $S_n$ vorgegebenen Semantik wahr in einem Modell M sind, auch wahr bzgl. der üblichen Semantik sind. Intuitiv gesprochen würden wahre Sätze ohne ein definiertes Zeichen nach der Einführung

von Definitionen gemäß $S_n$ wahr bleiben – wenn die $S_n$ denn stark nicht-kreativ wären.

Ein System $S_n$ ist schwach nicht-kreativ, wenn alle Sätze A, die kein definiertes Zeichen enthalten und die gültig gemäß $S_n$ sind, auch klassisch allgemeingültig sind.
In der Tat sollten semantische Systeme, mit denen man die Semantik von Definitionen einfangen will, nicht dazu führen, den semantischen Status solcher Sätze in der klassischen Logik zu ändern, die überhaupt gar kein definiertes Zeichen enthalten. Definitionen sollen nicht kreativ sein. Dahinter steckt die Grundintuition, daß Definitionen nicht mehr tun sollen, als die Bedeutung eines Ausdrucks anzugeben.[96]

Die genauen Formulierungen setzen weitere Definitionen und Begriffsklärungen voraus: Gupta und Belnap entwickeln die Begrifflichkeit für den allgemeinen Fall, daß mehrere Zeichen definiert werden – und nicht wie im informellen Teil ein einzelnes Zeichen G. Sie beschränken sich dabei auf die Definition von Prädikatsym-bolen. Hypothesen werden nun gleichzeitig für alle definierten Zeichen getroffen. Eine Hypothese h ist eine Funktion, die als Input Paare (G, d) hat, deren linkes Glied eines der definierten Prädikatsymbole ist und deren rechtes Glied ein n-Tupel (mit n = Stellenzahl von G) bestehend aus Gegenständen des Wertebereichs D ist. Als Output liefert h einen Wahrheitswert. Damit würde z.B. h(G,d) = W bedeuten, daß unter der Hypothese h das Tupel d von Gegenständen in der Extension des Prädikates G liegt. Der Revisionsoperator $\delta_{\mathcal{D},\mathcal{M}}$ hat damit diese komplexen Funktionen h als Input und Output: Im Modell $M$ und bzgl. der Menge $\mathcal{D}$ an Definitionen von einstelligen Prädikatbuchstaben ist somit $\delta_{\mathcal{D},\mathcal{M}}(h)$ diejenige Hypothesenfunktion, die man erhält, wenn man den Prädikatbuchstaben ihre durch die Definientia $A_G$ in M und unter h bestimmten Extensionen zuordnet. Damit erklärt man:

> Eine Hypothese h ist n-reflexiv auf $\mathcal{D}$ in M (für $\delta_{\mathcal{D},\mathcal{M}}$) $\quad \Leftrightarrow_{Df} \quad \delta^n_{\mathcal{D},\mathcal{M}}(h) = h$.
>
> Sei A ein Satz aus der um die Definienda erweiterten Sprache. A ist gültig auf $\mathcal{D}$ in $M$ im System $S_n$, kurz: $M \models_{\mathcal{D},\mathcal{M}} A$, $\quad \Leftrightarrow_{Df}$ Es gibt eine natürliche Zahl p, so daß für alle n-reflexiven h gilt: A ist wahr in $M + \delta^p_{\mathcal{D},\mathcal{M}}(h)$.
>
> A ist gültig auf $\mathcal{D}$ in $S_n$, kurz: $\models_{\mathcal{D},n} A$) $\quad \Leftrightarrow_{Df}$ Für alle klassischen Modelle M von L gilt: $M \models_{\mathcal{D},n} A$.

Hiermit lassen sich nun formulieren:

> **Stark nicht-kreativ**
> System $S_n$ ist stark nicht-kreativ $\quad \Leftrightarrow_{Df} \quad$ Für alle Definitionen $\mathcal{D}$, alle

[96] Für einige Sätze aber, wie ich in einem Intermezzo-Abschnitt angegeben habe, würde ich das nicht unbedingt fordern wollen.

Modelle M und alle Sätze A von L (= Sprache ohne die Definienda) gilt: Wenn $M \models_{\mathcal{D},n} A$ dann ist A wahr in $\mathcal{L}$ ($=< L, M, \tau >$)[97].

**Schwach nicht-kreativ**
System $S_n$ ist schwach nicht-kreativ $\Leftrightarrow_{Df}$ Für alle Definitionen $\mathcal{D}$ und alle Sätze A von L gilt: Wenn $\models_{\mathcal{D},n} A$ dann ist A klassisch allgemeingültig.

Ein einfaches Beispiel zeigt, daß die Systeme $S_n$, $n > 1$ nicht einmal schwach nicht-kreativ sind. ([18], S.152, Beispiel 5A.14). Das ist Grund genug, keines der Systeme $S_n$ als die korrekte Semantik für die Beschreibung zirkulärer Definitionen zu wählen.

### 7.1.2 Logische Schwäche

Der Begriff der logischen Schwäche scheint kein technischer Begriff zu sein. Zumindest geben Gupta und Belnap für diesen keine technische Definition an. Stattdessen führen sie an einem Beispiel vor, daß $S_0$ logisch schwach ist ([18], S. 154). Intuitiv scheint ein semantisches System für zirkuläre Definitionen logisch stark zu sein, wenn sich möglichst viele Sätze, die eines der Definienda enthalten, als gültig in diesem System herausstellen. In dem Beispiel wird ein einstelliges Prädikat G definiert und gezeigt, daß die Aussage '$(\forall x)Gx$' in $S_0$ nicht aus einer bestimmten Axiomenmenge folgt. Anscheinend soll aber diese Aussage aus der Axiomenmenge folgen. Gupta und Belnap scheinen nahelegen zu wollen, daß in $S_0$ mit der Definition das Intendierte nicht erreicht wird; und das Intendierte, das scheint entweder für Gupta und Belnap aus dem Kontext so offensichtlich hervorzugehen, daß sie es nicht näher erläutern, oder es geht aus der Aufstellung der Definition von G hervor. Betrachten wir dieses Beispiel genauer:

Das nichtlogische Vokabular der Sprache L enthalte den Namenbuchstaben '0', ein einstelliges Funktionssymbol $'$ (für die Nachfolgeroperation) und einen zweistelligen Prädikatbuchstaben $<$. AX sei die Konjunktion der folgenden Axiome:

$$\begin{aligned}
&(\forall x)(\neg x < x)\\
&(\forall x)(\forall y)(\forall z)((x < y \;\&\; y < z) \;\rightarrow\; x < z)\\
&(\forall x)(\forall y)(x < y \vee x = y \vee y < x)\\
&(\forall x)\neg(x < 0)\\
&(\forall x)(x < x' \;\&\; (\forall y)(x < y \;\rightarrow\; y = x' \vee x' < y)).
\end{aligned}$$

Die ersten drei Axiome besagen, daß '$<$' eine strikte lineare Ordnung ist.

[97]Das Dreiertupel besteht aus der syntaktisch konstruierten Sprache L, dem Modell M und dem klassischen Schema $\tau$, welches die Semantik der logischen Zeichen festlegt. Hier ist $\tau$ das klassische Schema, das den logischen Zeichen die Bedeutung zuordnet, welche sie in der klassischen bivalenten Quantorenlogik haben.

Die Worte 'G ist abgeschlossen' seien synonym mit folgender Formel

$(\forall x)((\forall y)(y < x \ \rightarrow \ Gy) \ \rightarrow \ Gx)$

Die Menge $\mathcal{D}$ von Definitionen für L enthalte als einzige Definition

$$Gx \quad =_{Df} \quad (G \text{ ist abgeschlossen } \& \ x = x) \vee (\neg(G \text{ ist abgeschlossen } \& \ (\forall y)(y < x \ \rightarrow \ Gy))$$

Das Definiens ist eine Disjunktion, die zwei nicht vereinbare Disjunkte enthält. Im ersten Disjunkt wird angenommen, daß G abgeschlossen ist, im zweiten, daß es nicht abgeschlossen ist. Sollte das durch diese Definition definierte G abgeschlossen sein, dann müßten aufgrund des zweiten Konjunktionsgliedes des ersten Disjunktionsgliedes alle Gegenstände im Wertebereich zur Extension von G gehören. Sollte das durch die Definition festgelegte G nicht abgeschlossen sein, dann müßten alle Dinge, die in der Extension von G liegen, das zweite Konjuktionsglied des zweiten Disjunktions erfüllen, um gemäß der Definition in der Extension von G zu sein - und diese besagt, daß alle Vorgänger eines Elementes x aus der Extension von G zu G gehören, was zur Folge hätte, daß G also doch abgeschlossen ist. Also dürfte gar nicht der Fall eintreten, daß G nicht abgeschlossen ist, folglich müßten alle Dinge des Wertebereichs zur Extension von G gehören. Das scheint die Intuition zu sein, die Gupta und Belnap bzgl. dieser Definition haben. In $S_0$ wird diese Intuition nicht erfüllt. Denn für diese Definition gilt *nicht*

$\models_{\mathcal{D},0} AX \ \rightarrow \ (\forall x)Gx$

Zum Beweis wähle man ein Modell M von AX. Ist die Anfangshypothese h für G abgeschlossen, dann sind alle Dinge des Wertebereichs in der revidierten Extension für G im Modell M. Die Revisionsfolge $\delta^n_{\mathcal{D},M}(h)$ stabilisiert sich auf die Menge, die der gesamte Wertebereich ist. Ist allerdings die Anfangshypothese h für G nicht abgeschlossen, dann gibt es einen Gegenstand d der in M+h zwar $(\forall y)(y < x \ \rightarrow \ Gy)$ erfüllt, aber nicht Gx. Die Anwendung des Revisionsoperators $\delta_{\mathcal{D},M}$ ergibt als neue Extension für G eine Menge, in der alle Dinge kleiner als d und d selbst enthalten sind. Eine weitere Anwendung fügt den Nachfolger von d hinzu usw. Das bedeutet, daß es keine natürliche Zahl p gibt, so daß $(\forall x)Gx$ wahr in $M + \delta^p_{\mathcal{D},M}(h)$ ist. Hieraus folgt die Behauptung.

Was könnte die logische Schwäche zur Folge haben, wenn wir $S_0$ auf die Definition des Wahrheitsprädikats anwenden würden? Es könnte zur Folge haben, daß sich die Definition zu unrecht nicht als material adäquat herausstellt. Bestimmte Sätze, in denen das Wahrheitsprädikat vorkommt, würden z.B. nicht als kategorisch herausgestellt werden, auch wenn wir unseren sprachlichen Intuitionen folgend meinten, es müsse sich als kategorisch herausstellen, da es ohne Einschränkung bejaht (bzw. verneint) werden kann.

## 7.2 Revisionsfolgen

Für die Definition der beiden semantischen Systeme $S^\#$ und $S^*$ in Abschnitt D des 5. Kapitels stellen Gupta und Belnap im Abschnitt C einige Tatsachen zu Revisionsfolgen und dem zugehörigen Begiffsapparat zusammen, den ich kurz und auf einen spezialisierten Fall angewandt wiedergeben werde. Die in diese Systeme eingehende Überlegung ist grob gesagt die, den Revisionsprozeß über die natürlichen Zahlen hinaus zu erweitern auf beliebige Ordinalzahlen. Nehmen wir an, es liegt eine zirkuläre Definition D eines einstelligen Prädikats G vor und es sei ein Modell M gegeben. Wir beginnen mit einer Initialypothese X für die Extension des Prädikats G; diese wird im Revisionsschritt revidiert zu einer neuen Extension $\delta_{D,M}(X)$, im folgenden Revisionsschritt zu einer anderen Extension $\delta^2_{D,M}(X)$ usw. Nun können wir uns fragen, was wir machen, wenn wir diese Revision in Gedanken unendlichmal oft durchgespielt haben. Technisch gesprochen würden wir uns fragen, wie denn die neue Extension $\delta^\omega_{D,M}(X)$ bei der ersten Limeszahl $\omega$ aussehen soll.[98] Diese Frage läßt sich dann verallgemeinern auf beliebige Limeszahlen. Klar ist: Wir werden in $\delta^\omega_{D,M}(X)$ diejenigen Dinge aufnehmen, die sich ab einer Revisionsstufe immer innerhalb der Extension von G befunden haben; sie waren die gesamte Zeit vor $\omega$ in der Extension von G, also gehören sie auch hier an der Limeszahl $\omega$ in die Extension von G. Klar ist auch, daß wir alle Dinge, die ab einer Revisionstufe vor $\omega$ immer außerhalb der Extension von G waren, auch bei $\delta^\omega_{D,M}(X)$ außerhalb lassen. Was ist aber mit den restlichen Dingen? Die verschiedenen semantischen Syteme aus der Revisionstheorie unterscheiden sich alle im Prinzip in der Beantwortung dieser Frage. Manche besagen, es dürfe nichts mehr hineingenommen werden[99], andere besagen, es müßten diese oder jene Sätze hinzugenommen werden. Bei der Entwicklung einer revisionstheoretischen Semantik gehört also noch die Angabe hinzu, wie bei einer erweiterten Revisionsfolge mit den Limeszahlen zu verfahren ist. Man mag mit der einen Limeszahl $\alpha$ in einer gegebenen Revisionsfolge anders verfahren als man mit einer verschiedenen Limeszahl $\beta$ in derselben Revisionsfolge verfährt. Im Prinzip kann diese Angabe also von der betrachteten Limeszahl abhängen. Darüber hinaus kann es sein, daß man für eine Revisionsfolge beginnend mit der Initialhypothese $X_1$ bei der Limeszahl $\alpha$ angibt, daß noch die zusätzliche Menge an Dingen $H_1$ aus dem Gegenstandsbereichs von M zu $\delta^\alpha_{D,M}(X_1)$ aufzunehmen ist, daß aber für eine andere Revisionsfolge beginnend mit der Initialhypothese $X_2$ bei derselben Limeszahl $\alpha$ eine andere zusätzliche Menge $H_2$ von Dingen zu $\delta^\alpha_{D,M}(X_2)$ zu zählen ist.

Belnaps und Guptas Systeme zeichnen sich durch maximale Beliebigkeit aus: Hier darf sozusagen an jeder Limeszahl geraten werden. Daß es hier dann nicht mehr unbedingt Sinn macht zu sagen, die revidierte Extension sei als bessere Hypothese für die Extension des Definiendums anzusehen, wie es noch bei $S_0$ der

[98] Diese Notation ist ungenau und wird von Belnap und Gupta auch nicht verwandt. Ich habe sie lediglich zum Zwecke der Veranschaulichung und Motivation der Ideen eingeführt.

[99] Das war ursprünglich Herzbergers Idee ([20]).

Fall war, ist klar und wird auch von Belnap und Gupta angemerkt.[100]

Im folgenden beschränke ich mich auf den Fall, daß nur ein Definiendum G, welches ein einstelliges Prädikatsymbol ist, vorliegt. Hypothesen für G sollen der Einfachheit halber nun nicht wie im oben angedeuteten allgemeinen Fall Funktionen in einen Wertebereich von Wahrheitswerten sein, sondern einfach Teilmengen des Gegenstandsbereichs. Außerdem mache ich die Einschränkung, daß nur zwei Wahrheitswerte vorliegen, so daß man in vernünftiger Weise von der Extension von G sprechen kann. Ein gegenstand d ist dann entweder in der Extension von G oder in der Extension von nicht-G; eine dritte Möglichkeit gibt es nicht. Sei $\Sigma$ eine Folge von Hypothesen bzgl. der Extension des Definiendums G mit der Länge $lh(\Sigma)$. Dabei ist die Länge einer Folge definiert als $lh(\Sigma)$ = Definifionsbereich der Folge $\Sigma$. Folgen sind hierbei verallgemeinert auf Funktionen, deren Definitionsbereich Ordinalzahlen sind. Bei Belnap und Gupta ist auch die Klasse On aller Ordinalzahlen gestattet. Mit '$\Sigma_\gamma$' wird das $\gamma$-te Folgenglied von $\Sigma$ herausgegriffen, wobei $\gamma$ wieder eine Ordinalzahl ist. $\delta$ ist eine Funktion, die Hypothesen (hier: Extensionen des einstelligen Prädikats G) als Input und Output hat; er ist der Revisionsoperator. Damit läßt sich der Begriff der Stabilität definieren:

> Ein Ding d ist stabil G in $\Sigma$ $\quad \Leftrightarrow_{Df}$
> Es gibt eine Ordinalzahl $\beta < lh(\Sigma)$ so daß für alle $\gamma$ gilt: Wenn $\beta \leq \gamma < lh(\Sigma)$, dann ist $\mathrm{d} \in \Sigma_\gamma$.

Ebenso definiert man, was es heißt, daß ein Ding d stabil nicht-G ist. Für einen Gegenstand d, der stabil G ist, wird also sein Status als ein zur Extension von G zugehöriges Ding irgendwann in der Folge festgelegt. Allgemein heißt ein Gegenstand d stabil, wenn es stabil G ist oder stabil nicht-G ist. Der folgende Kohärenzbegriff greift diejenigen Hypothesen heraus, die sozusagen dem Ergebnis einer Folge $\Sigma$ von Hypothesen, welche Dinge d stabil sind, nicht widersprechen. Wie die nichtstabilen Gegnstände behandelt werden, ist freigestellt:

> Eine Hypothese $h$ kohäriert mit $\Sigma$ $\quad \Leftrightarrow_{Df}$ $\quad$ Für alle $d$ gilt: (Ist $d$ stabil G in $\Sigma$, dann ist $d \in h$) und (ist $d$ stabil nicht-G in $\Sigma$, dann ist $d \notin h$).

Wir interessieren uns nicht für beliebige Folgen $\Sigma$ von Hypothesen, sondern solche, die generiert werden durch den Revisionsoperator, welcher durch die Definition D für unser Definiendum G festgelegt wird. Der Schritt von einer Ordinalzahl $\beta$ zu der nachfolgenden Ordinalzahl $\beta + 1$[101] ist der, den wir für $S_0$ kennengelernt haben. Neu hinzu kommt die Erklärung des Verhaltens bei den Limeszahlen:

> $\Sigma$ ist eine Revisionsfolge für die Revisionsregel $\delta$ $\quad \Leftrightarrow_{Df}$
> Für alle $\alpha < lh(\Sigma)$ gilt: Ist $\alpha = \beta + 1$, dann ist $\Sigma_\alpha = \delta(\Sigma_\beta)$. Und ist $\alpha$ eine Limeszahl, dann kohäriert $\Sigma_\alpha$ mit der Restriktion $\Sigma \restriction \alpha$[102].

[100] [18], S.168, Fn 25

[101] Üblicherweise wird die Nachfolgerzahl $\beta + 1$ definiert als $\beta \cup \{\beta\}$.

[102] Die Restriktion $\Sigma \restriction \alpha$ ist diejenige Folge, deren Definitionsbereich $\alpha$ ist und die mit $\Sigma$ auf dieser Menge $\alpha$ übereinstimmt.

Für den Begriff der rekurrierenden Hypothese ist der Begriff der Kofinalität nötig.

> Eine Hypothese $h$ ist kofinal in $\Sigma$ für $\delta$ $\Leftrightarrow_{Df}$ Für alle Ordinalzahlen $\alpha < lh(\Sigma)$ gibt es ein $\beta$, so daß $\alpha \leq \beta < lh(\Sigma)$ und $\Sigma_\beta = h$.

Zwischen dem Begriff der Stabilität und der Kofinalität besteht eine enge Beziehung, welche im Theorem 5C. zum Ausdruck kommt ([18], S.171). Dieser besagt auf unseren Spezialfall einer Definition des einstelligen Prädikats G angewandt: Wenn ein Gegenstand d stabil G (bzw. stabil nicht-G) in einer Folge $\Sigma$ von Hypothesen für G ist, dann ist d in allen Hypothesen h (bzw. außerhalb aller Hypothesen h), die kofinal in $\Sigma$ sind. (Diese Richtung ist einfach zu beweisen.) Die umgekehrte Richtung gilt unter der Bedingung, daß der Definitionsbereich der Folge $\Sigma$ die Klasse aller Ordinalzahlen ist, also $lh(\Sigma) = \text{On}$ gilt.

Nun können wir den wichtigen Begriff der rekurrierenden Hypothese erklären:

> Eine Hypothese $h$ ist rekurrierend für die Revisionsregel $\delta$ $\Leftrightarrow_{Df}$ Es gibt eine Revisionsfolge der Länge On $\Sigma$ für $\delta$, so daß $h$ kofinal in $\Sigma$ ist.

Intuitiv sind rekurrierende Hypothesen solche, die den Revisionsprozess überstehen. Das Problematische an dieser Definition ist, daß wir hier eine Quantifikation über echte Klassen vornehmen. Diese Problematik – Belnap und Gupta wollen in ZFC arbeiten ! – läßt sich mit dem Begriff der reflexiven Hypothese übergehen; man kann nämlich zeigen, daß alle und nur die rekurrierenden Hypothesen reflexiv sind.[103]

> Eine Hypothese $h$ ist $\alpha$-reflexiv für $\delta$ $\Leftrightarrow_{Df}$ Es gibt eine Revisionsfolge $\Sigma$ für $\delta$, so daß $\alpha < lh(\Sigma)$ und $\Sigma_0 = \Sigma_\alpha = h$.
>
> $h$ ist reflexiv für $\delta$ $\Leftrightarrow_{Df}$ $h$ ist $\alpha$-reflexiv für ein $\alpha > 0$.

## 7.3 $S^\#$ und $S^*$

Das System $S^\#$ wird mit Hilfe des Begriffs der rekurrierenden Hypothese definiert.

> **Gültigkeit in $S^\#$**
> Ein Satz $A$ ist gültig in $M$ in $S^\#$, kurz: $M \models_{D,\#} A$, $\Leftrightarrow_{Df}$ Für alle Hypothesen $h$, die für $\delta_{D,M}$ rekurrierend sind, gibt es eine Zahl n, so daß für alle $p > n$ gilt: A ist wahr in $M + \delta^p_{D,M}(h)$.
>
> A ist gültig in $S^\#$, kurz: $\models_{D,\#} A$ $\Leftrightarrow_{Df}$ Für alle klassichen Modelle M von L gilt: $M \models_{D,\#} A$

[103] [18], S. 174-175, Theorem 5C.13

Ein Satz A ist also gültig in $S^\#$, wenn A sich, sobald der Revisionsprozeß mit einer rekurrierenden Hypothese beginnt, nach einigen Revisionsschritten $n$ irgendwann auf den Wahrheitswert Wahr stabilisiert.

$S^\#$ erfüllt tatsächlich die von ihm erwartete Eigenschaft der Nicht-Kreativität. Alle Sätze, die in $S_0$ gültig sind, sind es auch in $S^\#$ – nicht aber umgekehrt. alle Sätze, die in $S^\#$ gültig sind, sind es auch in $S_n$, für $n > 1$, aber nicht umgekehrt. In endlichen Situationen, d.h. wenn die Menge der Definitionen endlich ist, wie ich ja eh vorausgesetzt habe, und wenn der Gegenstandsbereich des Modells M endlich ist, sind $S^\#$ und $S_n$ $(0 \leq n)$ äquivalent, d.h. etwas ist gemäß des einen Systems genau dann gültig in M, wenn es gemäß des anderen gültig in M ist. Unter solchen Umständen kann man also auch den kalkül $C_0$ verwenden, um für die Gültigkeit von Sätzen aus $S^\#$ zu argumentieren. Die üblichen aus der klassischen Definitionstheorie bekannten Einführungs- und Beseitigungsregeln für das Definiendum ((DFE) und (DFB)) gelten auch in $S^\#$ nur in nicht-hypothetischen Kontexten. Wie Philip Kremer in [24] nachweist, gibt es aber keinen korrekten und vollständigen Kalkül für $S^\#$.[104] Außerdem ist $S^\#$ von der Komplexität $\Pi^1_2$.[105] Eine das System $S^\#$ in gewissem Sinne degradierende Eigenschaft ist, daß es $\omega$-inkonsistent ist, d.h. daß die Menge der in $S^\#$ gültigen Sätze $\omega$-inkonsistent ist; und das wiederum bedeutet, daß es einen in 'x' offenen Satz A(x) gibt, so daß alle Sätze $\neg A(1), \neg A(2), \ldots$ gültig in $S^\#$ sind, aber gleichzeitg auch ($\exists$ natürliche Zahl n $)A(n)$ gültig in $S^\#$ ist. Hier behauptet also der letzte Satz, daß es eine Zahl gibt, die eine Eigenschaft A(x) hat, obwohl es keine natürliche Zahl gibt, die das bezeugen könnte.

Das System $S^*$ wird ebenfalls unter Rückgriff auf die rekurrierenden Hypothesen erklärt.

> **Gültigkeit in $S^*$**
> A ist gültig in M in $S^*$, kurz: $M \models_{D,*} A$, $\Leftrightarrow_{Df}$ A ist wahr in allen Modellen $M + h$, wobei $h$ eine rekurrierende Hypothese von $\delta_{D,M}$ ist.
>
> A ist gültig in $S^*$, kurz $\models_{D,*} A$ $\Leftrightarrow_{Df}$ Für alle Modelle M gilt: $M \models_{D,*} A$.

$S^*$ ist ebenfalls nicht-kreativ, besitzt keinen vollständigen Kalkül und hat ebenfalls eine Komplexität von $\Pi^1_2$.[106] Im Unterschied zu $S^\#$ ist $S^*$ aber nicht $\omega$-inkonsistent. Leider ist aber $S^*$ auch nicht so stark wie $S^\#$ – d.h. in $S^*$ sind nicht so viele Sätze gültig wie $S^\#$ – als daß man $S^*$ dem System $S^\#$ vorziehen könnte. Dieses Dilemma macht sich unangenehm bemerkbar, wenn die beiden Systeme auf die Definition des Wahrheitsprädikats angewandt werden. Ein von

[104] Was Kremer tatsächlich zeigt ist, daß für eine endliche Menge $\mathcal{D}$ an Definitionen die Menge $\{A : A$ ist gültig für $\mathcal{D}$ in $S^\#\}$ nicht rekursiv aufzählbar ist. Gebe es einen vollständigen Kalkül für $S^\#$, dann ließe sich aber diese Menge rekursiv aufzählen.

[105] Siehe den Aufsatz [24] von Kremer und [2] von Antonelli.

[106] Zu letzterem Ergebnis siehe wieder [24] und [2]

McGee bewiesenes Theorem zeigt, daß unter bestimmten Umständen für eine Wahrheitsdefinition immer so ein Dilemma zu erwarten ist.[107]

[107]Siehe dazu auch den kritischen Teil meiner Arbeit.

# 8 Wahrheit

Im folgenden soll Guptas und Belnaps Revisionstheorie der Wahrheit (RTW) umrissen werden. Zweck dieses Abschnitts ist in erster Linie nicht, die für eine Kritik an Wahrheitstheorien typischen Fragestellungen zu verfolgen und zu schauen, ob (RTW) diesen genügen kann, sondern sich ein Bild davon zu machen, wie die Theorie zirkulärer Definitionen auf einen umgangssprachlichen Begriff angewandt werden kann. Bisher hatten wir es nur mit dem hohlen Prädikatbuchstaben 'G' zu tun, dem in verschiedenen Beispielen durch zirkuläre Definitionen stipulativ eine Bedeutung aufgedrückt wurde. Würden nicht neben solche künstlichen, konstruierten Gebilde „echte", bereits mit einer Bedeutung ausgestattete Ausdrücke gestellt und würde nicht gezeigt werden, daß ihre Bedeutung durch eine zirkuläre Definition erfaßt wird, dann bliebe eine durch die klassische Theorie der Definition genährte Skepsis gegenüber zirkulären Definition gerechtfertigt.

Da Gupta und Belnap bei der Entwicklung von (RTW) voraussetzen, daß der Leser mit Tarskis Wahrheitskonzeption vertraut ist, werde ich hier in den allergröbsten Zügen dessen semantische Wahrheitskonzeption und – was vielleicht noch wichtiger ist – seine Vorüberlegungen hierzu darstellen, deren Kern das Adäquatheitskriterium für eine jede Definition des Wahrheitsbegriffs bildet.

Da Gupta und Belnap ihre Wahrheitstheorie u.a. im Vergleich mit der Fixpunkttheorie bewerten, wäre eine Darstellung auch dieser Theorie angebracht. Ich möchte aber nicht auf die Kritik von Gupta und Belnap an der Fixpunkttheorie und den Vergleich von (RTW) mit anderen Wahrheitstheorien eingehen. Dieser vergleichende Teil, der in RTT im dritten Kapitel seinen Platz hat, wird vollständig ausgelassen. Die Darstellung von Tarskis Wahrheitskonzeption hat lediglich den Zweck, sich die Ideen Guptas und Belnaps leichter verständlich zu machen. Ich folge der Darstellung in [26].

## 8.1 Tarskis semantische Wahrheitskonzeption

Ehe Tarski in seinem Aufsatz „Der Wahrheitsbegriff in den formalisierten Sprachen" ([42]) seine eigene, semantische Wahrheitskonzeption entwickelt, stellt er Vorüberlegungen darüber an, was eine Theorie oder Definition der Wahrheit grundsätzlich zu leisten hat. Die Theorien bzw. Definitionen, die für Tarski relevant sind, sollen eine ganze Schar von Prädikaten, nämlich solche der Form 'ist wahr in L', (definitorisch) erläutern. Das Zeichen 'L' steht hierbei für eine Sprache, für deren Sätze das Prädikat 'wahr in L' erklärt wird. Insbesondere sind also Tarskische Wahrheitswertträger konkrete Sätze einer bestimmten Sprache.

Was für Gebilde Sprachen genau sind und wie sie angemessen charakterisiert werden können, steht hier nicht zur Debatte. Es muß hier jedoch erwähnt werden, woran Tarski nicht denkt, wenn er von einer Sprache L spricht. Tatsächlich ist es das naive Verständnis der Sprache und nicht eines, welches durch die mathematische Logik nahegelegt werden könnte. Die Sprachen L sind *nicht* bloß Mengen

von Zeichenketten, die durch bestimmte Konstruktionsregeln gebildet werden. Sie sind auch *nicht* Mengen von bestimmten Zeichenketten zuzüglich eines Modells aus der klassischen Quantorenlogik. Man hat sich stattdessen vorzustellen, daß es vernünftig ist, von der Sprache L zu sagen, über die in L geltenden syntaktischen und semantischen Regeln hinaus, welche sich in der Quantorenlogik beschreiben lassen, besäßen die Ausdrücke aus L eine Bedeutung. Der Grund ist folgender: Wenn für einen Ausdruck $A$ der Objektsprache nur die Extension bekannt wäre und er in die deutsche Sprache „übersetzt“ werden sollte, dann könnte es sein, daß die deutsche Sprache zwei bedeutungsverschiedene Ausdrücke $B_1$ und $B_2$ besitzt, die dieselbe Extension haben, nämlich gerade die, die ihnen als Übersetzung von $A$ zuzuordnen wäre. Aber welchen der Ausdrücke $B_1$ oder $B_2$ wählen wir? Die Übersetzung von der Objektsprache zur Metasprache wäre nicht eindeutig. Damit könnte man aber nicht die Bedeutung (Intension) des Prädikats ‘ist wahr’ für die Objektsprache festlegen, aber mindestens das soll eine Wahrheitsdefinition leisten: Die Intension von ‘ist wahr’ festlegen.

Ergebnis der Vorüberlegungen sind zwei Kriterien: Das Kriterium der formalen Korrektheit und Konvention W.

*Formale Korrektheit*
Die formale Korrektheit fordert, daß die Definition des Prädikats ‘ist wahr in L’ in einer von der sogenannten Objektsprache L verschiedenen Metasprache M zu erfolgen hat. Es wird nicht ausgeschlossen, daß die Objektsprache eine (dann aber echte) Teilmenge der Metasprache ist. Insbesondere gehört das Prädikat ‘ist wahr in L’ der Metasprache an. Weitere formale Anforderungen erwachsen aus der Konvention W:
(1) In M gibt es Bezeichnungen für die Ausdrücke der Objektsprache L. Tarski nennt zwei Möglichkeiten, auf die Ausdrücke von L Bezug zu nehmen: a) durch Anführungsnamen, bei denen der Ausdruck E in Anführungsstriche gesetzt wird und b) durch Beschreibung der Struktur des Ausdrucks E, bei der von links nach rechts fortschreitend die einzelnen Zeichen der linearen Zeichenkette E benannt werden. Bezeichnungen letzteren Typs werden strukturell-deskriptive Namen genannt.[108]
(2) M enthält logische Konstanten, insbesondere eines wie ‘genau dann, wenn’ oder ‘$\leftrightarrow$’ oder ähnlichem.
(3) Alle Sätze der Objektsprache L müssen in die Metasprache übersetzbar sein.

*Konvention W*

[108]Beispielsweise ist der folgende Ausdruck ein Anführungsname: ‘1+2’. Mit ihm wird der folgende Ausdruck bezeichnet: 1+2. Ein strukturell-deskriptiver Name für letzteren Ausdruck könnte etwa lauten: Diejenige lineare Zeichenkette, die aus dem Zeichen für die Zahl Eins, ‘1’, gefolgt von dem Zeichen für die Addition ‘+’ und dem Zeichen für die Zwei, ‘2’, (in dieser Reihenfolge von links nach rechts gelesen) besteht.

> Nach diesem Kriterium, das nun nicht den formalen sondern sachlichen oder inhaltlichen Aspekt von Wahrheitsdefinitionen betrifft, ist eine Definition von 'wahr in L' adäquat gdw aus ihr in der Metasprache M alle Sätze folgen, die Instanzen des folgenden Schemas sind:
>
> (W) S ist wahr genau dann, wenn p
>
> Dabei entstehen Instanzen des Schemas (W) durch Ersetzung von 'S' durch eine Bezeichnung eines Satzes der Objektsprache L und von 'p' durch die Übersetzung des so bezeichneten Satzes in die Metasprache. Diese Einsetzungsinstanzen sollen fortan als W-Äquivalenzen bezeichnet werden.
>
> Ist beispielsweise die Objektsprache L ein fragmentarisches Englisch L1, welches den Satz 'Snow ist white' enthält, und die Metasprache das Deutsche, dann müßte dieser Konvention gemäß aus einer adäquaten Definition für 'wahr in L1' die Instanz
>
> 'Snow is white' ist wahr in L1 genau dann, wenn Schnee ist weiß
>
> folgen.

Diese beiden Forderungen stellen nach Tarski also *Adäquatheitskriterien* für eine *jede* Theorie bzw. Definition der Wahrheit dar. Insbesondere ist mit der Angabe der beiden Kriterien keine *Definition* für den Wahrheitsbegriff gegeben. Zu diesem Irrglauben mag Tarskis Bemerkung beigetragen haben, daß man die W-Äquivalenzen als Teildefinitionen ansehen könne. (s.u.)

Tarskis eigene Wahrheitsdefinition, die er im Anschluß an die Formulierung dieser Kriterien entwickelt, erfüllt natürlich diese Kriterien. Seine Wahrheitsdefinition bedient sich der Rekursion: Die Wahrheit von komplexen Sätzen wird auf die Wahrheit von einfacheren Sätzen zurückgeführt. Für Sätze, die quantorenlogische Elemente beinhalten, kann dieses Verfahren nicht direkt angwandt werden: So ist z.B. ein Satz der Form $\ulcorner(\forall x)Fx\urcorner$ nicht wieder aus Sätzen aufgebaut, als daß man die Bestimmung des Wahrheitswertes eines solchen Satzes zurückführen könnte auf die Bestimmung des Wahrheitswertes der in ihm als Teilsätze enthaltenen Sätzen. Man denkt sich vielmehr solche Sätze zusammengesetzt aus derjenigen Komponete im Satz, die dem Quantorzeichen '$(\forall x)$' entspricht und derjenigen Komponente im Satz, die dem offenen Satzschema '$Fx$' entspricht. Erstere Komponente gehört der logisch-grammatischen Kategorie der Quantoren an – in diesem Fall speziell den Allquantoren. Die zweite Komponente gehört der Kategorie der offenen Sätze an. Z.B. wäre ein in 'er' offener Satz der Ausdruck 'er ist ein wundersamer und verrückter Philosoph', wenn hier 'er' nicht deiktisch oder anaphorisch gebraucht wird. Von solchen Gebilden kann man nicht mit Sinn fragen, ob sie wahr sind oder nicht. Was sich allerdings sagen läßt, wäre z.B. daß der Philosoph Diogenes die durch den offenen Satz gegebene Bedingung erfüllt; auf

ihn trifft zu oder von ihm gilt, was man mit diesem offenen Satz ausdrückt. Diese Intuition wird mit dem technischen Begriff der Erfüllung eingefangen. Tarski gebraucht ihn so, daß alle Einsetzungsinstanzen des Schemas

(E) Ein Gegenstand a erfüllt einen in einer Variablen offenen Satz O genau dann, wenn a $\phi$t

wahr werden, wenn man für 'O' einen Namen eines offenen Satzes der Objektsprache einsetzt und für '$\phi$t' dessen Übersetzung in die Metasprache. Auch hier läßt sich ein Adäquatheitskriterium formulieren, das für eine jede Definition des Begriffs der Erfüllung fordert, daß aus ihr sämtliche korrekte Instanzen des Schemas (E) folgen. Dieser Begriff der Erfüllung gibt Tarskis *semantischer* Wahrheitskonzeption ihren Namen: Er ist ein semantisches Prädikat im primären Sinne, und Wahrheit als ein durch diesen Begriff definiertes Prädikat ein semantisches im sekundären Sinne.[109]

Bei offenen Sätzen, die in mehr als einer Variabler frei sind bzw. die mehr als ein als Variable gebrauchtes Pronomen enthalten, sind es dann nicht mehr einzelne Gegenstände von Dingen, die diese sättigen, sondern Tupel von Gegenständen – bei n freien Variablen also n-Tupel. Da es keine Beschränkung der Anzahl der in einem offen Satz frei vorkommenden Variablen gibt, formuliert Tarski den Begriff der Erfüllung so, daß es nun unendliche Folgen von Dingen – sozusagen $\infty$-Tupel – sind, die offene Sätze erfüllen. Im folgenden Beispiel beschränken wir uns auf den Fall eines in einer Variablen offenen Satzes.

Zur Veranschaulichung betrachten wir folgendes kompliziertere Beispiel einer Objektsprache L. Das fragmentarische Englisch L3 bestehe aus folgenden Komponenten:

(I) *Vokabular*
Es sind an nichtlogischen Ausdrücken enthalten: Zwei Prädikate **is a town** und **has a university**; unter den logischen Ausdrücken finden sich eine Variable **it**, ein Quantor **at least one object is such that**, zwei Junktoren **and** und **it is not the case that**. Als zusätzliche Hilfszeichen stehen eine Linksklammer '(' und eine Rechtsklammer ')' zur Verfügung

(II) *Syntax*
Man erklärt zunächst rekursiv, was ein *offener Satz in L3* ist.
(i) **it is a town** und **it has a university** sind offene Sätze.
(ii) Die Verknüpfung zweier offener Sätze durch Einfügung von: **and** zwischen sie und von Außenklammern ist wieder ein offener Satz. (Kurznotation: Mit $O_1$ und $O_2$ ist auch $\ulcorner(O_1$ **and** $O_2)\urcorner$ ein offener Satz.)

[109] Ein zweistelliges Prädikat ist semantisch im primären Sinne gdw es eine Beziehung zwischen sprachlichen Zeichen und etwas, was (normalerweise) kein sprachliches Zeichen ist, ausdrückt. Ein Prädikat (beliebiger Stellenzahl) ist semantisch im sekundären Sinne gdw es mit Hilfe eines Prädikates definiert wird, welches semantisch im primären Sinne ist.

(iii) Mit $O$ ist auch ⌜ **it is not the case that** $O$⌝ ein offener Satz.
(iv) Nur solche Ausdrücke, die sich nach endlichmaliger Anwendung der Regeln (i)-(iv) bilden lassen, sind offene Sätze. („Endklausel“)
*Rekursive Definition von 'Satz in L3'*
(a) Ist O ein offener Satz, dann ist
⌜ **at least one object is such that** $O$⌝ ein Satz.
(b) Mit $O_1$ und $O_2$ ist auch ⌜($O_1$ **and** $O_2$)⌝ ein Satz.
(c) Mit $O$ ist auch ⌜ **It is not the case** $O$⌝ ein Satz.
(d) „Endklausel“

(III) *Semantik von L3*
Hier wird jetzt also zunächst das sematische Prädikat der Erfüllung erklärt, auf dessen Basis die Wahrheit definiert werden kann.
*Rekursive Definition von 'erfüllt in L3'*
(A1') a erfüllt '**it is a town**' gdw a eine Stadt ist.
(A2') a erfüllt '**it has a university**' gdw a eine Universität hat.
(B') a erfüllt einen komplexen offenen Satz der Form ⌜($O_1$ **and** $O_2$)⌝ gdw a erfüllt $O_1$, und a erfüllt $O_2$.
(C') a erfüllt einen komplexen offenen Satz der Form ⌜ **It is not the case that** $O$⌝ gdw es nicht der Fall ist, daß a O erfüllt.
*Rekursive Definition von 'wahr in L3'*
(A) Ein Satz der Form ⌜ **At least one object is such that** $O$⌝ ist wahr gdw mindestens ein Gegenstand so beschaffen ist, daß er O erfüllt.
(B) Ein Satz der Form ⌜($O_1$ **and** $O_2$)⌝ ist wahr gdw $O_1$ wahr ist und $O_2$ wahr ist.
(C) Ein Satz der Form ⌜ **it is not the case that** $O$⌝ ist wahr gdw es nicht der Fall ist, daß $O$ wahr ist.

Anhand eines Beispiels machen wir uns klar, daß folgende W-Äquivalenz ableitbar ist: '**At least one object is such that it is a town and it is not the case that it has a university**' ist wahr gdw es einen Gegenstand gibt, der eine Stadt ist und für den es nicht der Fall ist, daß er eine Universität hat.
(1) Nach Klausel (II),(i) ist '**it is a town**' ein offener Satz.
(2) Nach derselben Klausel auch '**it has a university**'.
(3) Nach Klausel (II), (iii) angewandt auf Zeile (2) ist auch '**it is not the case that it has a university**' ein offener Satz.
(4) Nach Klausel (II), (ii) angewandt auf (1) und (3) ist auch '**it is a town and it is not the case that it has a university**' ein offener Satz.
(5) Gemäß (II), (a) angewandt auf (4) ist '**At least one object is such that it is a town and it is not the case that it has university**' ein Satz.
(6) Aus (III), (A) ist '**At least one object is such that it is a town and it is not the case that it has university**' wahr gdw es einen Gegenstand gibt, der '**it is a town and it is not the case that it has university**' erfüllt.

(7) Nach (III), (B') angewandt auf den Ausdruck rechts vom 'gdw' in (6) und mit Hilfe der Aussagenlogik folgt: '**At least one object is such that it is a town and it is not the case that it has university**' ist wahr gdw es einen Gegenstand gibt, der '**it is a town**' erfüllt und der '**it is not the case that it has a university**' erfüllt.
(8) Nach (III), (C') angewandt auf das zweite Konjunkt hinter 'gdw' in (7) und mit der Aussagenlogik folgt: '**At least one object is such that it is a town and it is not the case that it has university**' ist wahr gdw es einen Gegenstand gibt, der '**it is a town**' erfüllt und für den es nicht der Fall, daß er '**it has a university**' erfüllt.
(9) Nach (III), (A1') angewandt auf das erste Konjunktionsglied hinter 'gdw' in (8) und nach (III), (A2') angewandt auf das zweite Konjunktionsglied hinter 'gdw' in (8) und mit der Aussagenlogik folgt schließlich die erwünschte W-Äquivalenz: '**At least one object is such that it is a town and it is not the case that it has university**' ist wahr gdw es einen Gegenstand gibt, der eine Stadt ist und für den es nicht der Fall ist, daß er eine Universität hat.

Entsprechend kann per Induktion über den Formelaufbau in L3 nachgewiesen werden, daß sämtliche W-Äquivalenzen ableitbar sind, die aus dem Schema (W) durch Einsetzen eines Namens für einen Satz aus L3 an der S-Position und durch Einsetzen der Übersetzung ins Deutsche an der p-Position entstanden sind.

Zum späteren Vergleich mit der Wahrheitstheorie von Gupta und Belnap führe ich hier die Hauptkritikpunkte an Tarskis semantischer Wahrheitskonzeption auf. Die Kritiken lassen sich grob in zwei Kategorien unterteilen. Die erste Kategorie beinhaltet kritische Fragen zur Anwendbarkeit von Tarskis Theorie auf natürliche Sprachen. Wir haben mit der Vorführung einer rekursiven Definition des Prädikats 'ist wahr in L3' ein relativ leichtes Projekt veranstaltet. Auch Tarskis Projekt, rekursive Definitionen der Wahrheit für bestimmte mathematische Theorien aufzustellen, ist – wiewohl nicht mit unserem mickrigen Beispiel vergleichbar, so doch – immer noch keine Definition des Wahrheitsprädikates einer natürlichen Sprache. Die zweite Kategorie hinterfragt den definitorischen Anspruch: Wird mit den rekursiven Definitionen der Sinn des Prädikats 'ist wahr' erfaßt?

Die im folgenden genannten Unterpunkte zu den beiden Kategorien bilden eine Auswahl von kritischen Anmerkungen, die ich im Vergleich mit (RTW) für wichtig hielt.

> *„Frage der Anwendbarkeit und Reichweite"*
> Tarski selbst glaubte nicht daran, daß sich seine semantische Wahrheitskonzeption auf die natürlichen Sprachen übertragen ließe. Die Hauptursache hierfür sah er darin, daß in solchen Sprachen das Wahrheitsprädikat enthalten ist, also keine Unterscheidung zwischen Objektsprache und Metasprache vorliegt, und sich aufgrund dessen fatale Paradoxa wie das Lügnerparadox ableiten lassen.

Natürliche Sprachen enthalten Indikatoren wie ‘ich’, ‘jetzt’ etc. Für Sätze, die Indikatoren enthalten, ist aber der Sachbezug nicht eindeutig durch die Gestalt des Satzes bestimmt: Indikatoren sind kontextsensitive Ausdrücke, für deren Bezug die äußeren Umstände relevant sind. Für solche Sätze sind die entsprechenden W-Äquivalenz je nach Äußerungskontext nicht immer wahr. Daher wäre es auch verfehlt, für solche Sprachen zu fordern, daß aus ihnen die W-Äquivalenzen folgen müssen.

Natürliche Sprachen sind sehr komplex und in manchen Hinsichten problematisch. In ihnen finden sich Junktoren wie z.B. ‘Helmut Kohl glaubt, daß p’, ‘Es ist notwendig wahr, daß p’ oder ‘p, weil q’. Der semantische Status von Sätze, die solche Operatoren enthalten, hängt nicht allein vom semantischen Status der in ihnen enthaltenen Teilsätze ab. So sind z.B. die beiden Sätze ‘Niemand ist sein eigener Vater’ und ‘Gerhard Schröder ist Bundeskanzler im Jahre 2000’ wahr, aber während der Satz ‘Es ist notwendig wahr, daß niemand sein eigener Vater ist’ wahr ist, ist der Satz ‘Es ist notwendig wahr, daß Gerhard Schröder Bundeskanzler im Jahre 2000 ist’ falsch. Für eine Sprache mit intensionalen Operatoren scheint daher eine rekursive Definition des Wahrheitsprädikats nicht möglich zu sein.[110]

*Definitorischer Anspruch*
Das erste Problem mit dem definitorischen Anspruch könnte Projektionsproblem genannt werden. Tarski hat gleich eine Schar von Prädikaten zum Gegenstand seiner semantischen Wahrheitskonzeption gemacht, nämlich solche von der Form ‘wahr in L’. Die rekursive Definition für ein Wahrheitsprädikat ‘wahr in L1’ sagt nichts über das Wahrheitsprädikat ‘wahr in L2’ einer anderen Sprache L2. Für unterschiedliche Sprachen L sind die Extensionen der Prädikate ‘wahr in L’ (normalerweise) unterschiedlich. Folglich können solche Prädikate auch nicht denselben Sinn haben.

Betrachten wir eine einfache Sprache L1, die lediglich die beiden Sätze ‘**The earth moves**’ und ‘**The moon is round**’ enthält. Dann könnte das Wahrheitsprädikat ‘wahr in L1’ in der Metasprache so erklärt werden:

(D1) x ist wahr in L1 gdw
(x = ‘**The earth moves**’ und die Erde bewegt sich) oder (x = ‘**The moon is round**’ und der Mond ist).

Diese Definition erfüllt das Adäquatheitskriterium. Würde man nun behaupten wollen, daß mit dieser Definition der Sinn von ‘wahr in L1’ angegeben wird, dann scheint es, als müßte man für folgende Definition des Prädikats ‘x ist eine Tochter von Laban’, welche der Definition ‘wahr in L1’ ähnelt, ebenfalls behaupten, es gebe den Sinn von ‘ist eine Tochter von Laban’ an.

[110] Davidson schlägt vor, Sätze mit intensionalen Operatoren parataktisch zu lesen.

(D2) x ist eine Tochter von Laban gdw (x = Lea) oder (x= Rahel).

Damit scheint aber doch nicht der Sinn angegeben zu sein. Dieser wäre erst durch

(D3) x ist eine Tochter Labans gdw (x ist weiblich, und x wurde von Laban gezeugt.)

festgelegt. Wie gelangt man zu diesem Urteil? Mit (D2) läßt sich absolut nichts darüber sagen, unter welchen Bedingungen jemand eine Tochter von Namik Özçep ist. Das Prädikat links vom 'gdw' in (D2) ist ein relationales Prädikat, bei dem in die vom Ausdruck 'Laban' eingenommene Position hineinquantifiziert werden kann. Das bedeutet, daß ein satzförmiges Gebilde, welches den Sinn eines solchen Prädikats festlegt, auf Fälle mit anderen singulären Termen in der Position des Ausdrucks 'Laban' übertragbar oder projizierbar zu sein hat. Für (D2) ist das nicht möglich, hier taucht im Definiens nicht einmal das Wort 'Laban' auf.[111] Anders verhält es sich mit (D3), welches 'Laban' im Definiens enthält und auch Projizierbarkeit gestattet.

Genau diese Problematik läßt sich auch auf die Definition des Wahrheitsprädikats 'wahr in L1' in (D1) übertragen. Mit dieser Definition erfahren wir nichts darüber, wann z.B. 'wahr in L3' auf einen Satz von 'L3' zutrifft.

Drei Anmerkungen zu dieser Problematik:
1.) Interpretiert man (D1) einfach als Bikonditional, dann wird mit (D1) natürlich nur die Extension von 'ist eine Tochter Labans' festgelegt. Aber selbst in der (STDL) legt eine Definition immer mindestens die Intension, d.h. die Anwendungsbedingungen in allen möglichen Welten, fest, sofern die möglichen Welten durch Modelle repräsentiert werden können. Das ändert aber nichts an der Tatsache, daß man mit (D1) nicht den Sinn von 'ist eine Tochter Labans' wiedergibt.
2.) Bleibt etwas von der Projektionsproblematik übrig, wenn man wie Gupta und Belnap ausdrücklich betont, es dürfe von den (partiellen) Definitionen der Wahrheitsprädikate nicht mehr als die intensionale Adäquatheit gefordert werden? Muß die Angabe der Anwendungsbedingungen für 'wahr in L1' etwas über die Anwendungsbedingungen von 'wahr in L3' abzuleiten gestatten? Diese Frage werde ich im kritischen Teil meiner Arbeit wieder aufnehmen.
3.) Auf die Projektionsproblematik könnte jemand einwenden, Tarskis Ziel sei doch nicht die Definition eines zweistelligen Prädikates 'wahr in L' gewesen, sondern von vielen einstelligen Prädikaten 'wahr in L1' , 'wahr in L2',

[111] Man beachte, daß das Vorkommen des Ausdrucks 'Laban' im Definiens nicht hinreichend wäre. Es muß zusätzlich gewährleistet sein, daß der Ausdruck 'Laban' potent bleibt. In einem Definiens wie '(x = Lea) oder (x= Rahel) oder (Laban ist Vater oder es ist nicht der Fall, daß Laban Vater ist.)' ist das Vorkommnis nicht potent.

'wahr in L3' etc. und er hätte zur Verdeutlichung besser 'wahr-in-L1', 'wahr-in-L2', 'wahr-in-L3' schreiben sollen. Mich interessiert jetzt nicht die Frage, ob das tatsächlich Tarskis Vorstellungen entsprochen hätte. Man wird den Einwendenden fragen wollen, weshalb in all diesen verschiedenen Prädikaten derselbe Ausdruck 'wahr' verwendet wird, wenn er doch im Kontext des Prädikats 'wahr-in-L' keinen zum Sinn des gesamten Prädikats beitragenden Sinngehalt hat. Könnte da der Einwendende erwidern, es solle darauf hinweisen, daß alle diese Prädikate eine bestimmte Funktion in den einzelnen Sprachen erfüllen würden, die sich zwar beschreiben läßt, aber nicht für die Definition eines zweistelligen Prädikats 'wahr in L' geeignet ist?

Das zweite Problem kann das Problem des epistemischen Status genannt werden. Betrachten wir hierzu wieder das obige Beispiel. Nach der Definition (D2) haben der Satz

(P1) Rahel ist eine Tochter von Laban

und der Satz

(P1) Rahel = Rahel oder Rahel = Lea

denselben Wahrheitswert. Würde (D2) tatsächlich den Sinn von 'ist eine Tochter von Laban' angeben, dann müßten diese beiden Sätze auch denselben epistemische Status haben, was offensichtlich nicht der Fall ist. Ähnlich verhält es sich mit den folgenden Sätzen, die denselben Wahrheitswert haben:

(Q1) 'Snow is white' ist wahr in L1 gdw Schnee weiß ist

(Q2) ('Snow is white'='Snow is white' und Schnee ist weiß) oder ('Snow is white'='Grass is green' und Gras ist grün)) gdw Schnee ist weiß

Um zu wissen, daß der letztere dieser beiden Sätze wahr ist, muß man lediglich ein wenig Kenntnisse in der Junktorenlogik besitzen. Für den ersteren dieser beiden Sätze muß man hierüber hinaus aber auch wissen, wie der Satz 'Snow is white' im Englischen verwendet wird, um die Wahrheit von (Q1) einzusehen.

## 8.2 Was (RTW) leisten soll

Nach 31 Seiten fassender einleitender Diskussion über den Gegenstand ihrer Untersuchung formulieren Gupta und Belnap die Aufgabe, welche (RTW) zu lösen hat:

> Given a first-order language L with a distinguished predicate **T** that means „true-in-L“, and given a classical model **M** of the **T**-free fragment of L, construct a systematic account of the signification of **T** that
>
> - yields a classification of the sentences of L into true/false/paradoxical/etc. – a classification that conforms to our ordinary intuitions and uses of ‘true’ and
> - yields an interpretation of the T-biconditionals that is in accord with the Signification Thesis. ([18], S.32)

Um diese Aufgabenstellung zu verstehen, müssen natürlich die Ausdrücke ‘Signifikation’, ‘Signifikationsthese’ und ‘T-Bikonditional’ erläutert werden. Diese Erläuterungen werde ich im folgenden geben.

Man wird aber auch ohne eine genaue Vorstellung von der Bedeutung dieser Worte einsehen können, was die (RTW), die Revisions**theorie der Wahrheit**, *nicht* ist: Eine Theorie der Wahrheit, die sämtliche philosophisch relevanten Aspekte des umgangssprachlichen Ausdrucks ‘ist wahr’ im Deutschen bzw. ‘is true’ im Englischen betrachtet und womöglich erklärt. Das Hauptaugenmerk der (RTW) liegt in der Angabe der Anwendungsbedingungen des Wahrheitsprädikats, welche eine korrekte Einteilung der Sätze in solche, die wahr, falsch oder paradox etc.[112] sind, ergibt, und in der Erklärung dieser Einteilung. Die von Gupta und Belnap mit (RTW) propagierte Theorie der Wahrheit ist eine Theorie, welche von der Beobachtung des merkwürdigen Verhaltens des Prädikats ‘ist wahr’ in Paradoxien wie der des Lügners ausgeht und versucht, diejenigen Eigenschaften des Wahrheitsbegriffs, die für dieses Verhalten verantwortlich sind, kenntlich zu machen.

Die Wahrheitswertträger, also diejenigen Gegenstände, auf die das zu erklärende Wahrheitsprädikat genuin anzuwenden ist, sind wie auch bei Tarski Sätze.[113] Daß es hiermit im Falle von Sprachen, die Indikatoren enthalten, Probleme gibt, haben wir bereits gesehen. Gupta und Belnap folgen Tarski, indem sie die Definition von Wahrheitsprädikaten ‘wahr in L’ für solche Sprachen L zu geben versuchen, die abgesehen vom Wahrheitsprädikat unproblematisch sind: Diese

[112] Hinter dem ‘etc.’, das sei hier schon vorweggenommen, verbergen sich keine wesentlich anderen Klassifikationskategorien wie z.B. Vagheit.

[113] Obwohl Gupta und Belnap die Prädikate ‘wahr in L’ auf Sätze der Sprache L angewandt wissen wollen, räumen sie ein, daß bei einer Weiterentwicklung der Theorie, die auch etwas über ‘wahr in L’ für eine natürliche Sprache L sagt, Propositionen die Rolle von Wahrheitswertträgern spielen müßten. Man müßte dann erklären, wann eine Proposition p, die in dem und dem Kontext K von einem Satz s der natürlichen Sprache ausgedrückt wird, wahr ist.

dürfen keine Indikatoren, vage oder ambige Ausdrücke enthalten, sie dürfen keine intensionalen Konstruktionen gestatten und die Sätze dürfen keine Wahrheitswertlücken aufweisen. Mit der Beschränkung auf Sprachen, die den Sprachen der klassischen Quantorenlogik ähneln, können Gupta und Belnap diesem Desiderat nachkommen.

Der Unterschied zu Tarskis Projekt besteht darin, daß jetzt zugelassen wird, daß das Prädikat 'wahr in L' in der Sprache L selbst vorkommen darf. Eine Unterscheidung zwischen Objektsprache und Metasprache gibt es immer noch, sie betrifft aber nicht das Wahrheitsprädikat T, sondern die von Gupta und Belnap für die semantische Erklärung von zirkulären Definitionen benutzten Begriffe „kategorisch", „paradox", „Gültigkeit" etc.[114] Obwohl das Sprachfragment von L, welches man durch Herausnahme des Wahrheitsprädikates erhält, klassisch ist, wollen Gupta und Belnap für das Wahrheitsprädikat nichts dergleichen Einschränkendes fordern. Es solle die Möglichkeit nicht ausgeschlossen werden, daß das Wahrheitsprädikat sich als drei-wertiges oder vier-wertiges ... oder n-wertiges Prädikat herausstellt. Das meint folgendes: Nachdem man das Wahrheitsprädikat zur klassischen Sprache hinzugenommen hat, soll die Möglichkeit nicht ausgeschlossen werden, daß man eventuell die Sätze, die das Wahrheitsprädikat enthalten, in einer drei-wertigen, vier-wertigen ... n-wertigen Logik semantisch bewertet. Es solle auch nicht vorher ausgeschlossen werden, daß das Wahrheitsprädikat vielleicht vage ist oder andere semantische Sonderheiten aufweist. Es sei die Aufgabe einer Wahrheitstheorie, festzulegen, welches die geeignete Semantik für das Wahrheitsprädikat ist.

(RTW) hat einen ganz bestimmten Wahrheitsbegriff der natürlichen Sprachen zum Gegenstand, nämlich den logischen, schwachen, absoluten Begriff der Wahrheit. In den drei folgenden Unterparagraphen soll erklärt werden, was der „(1) logische, (2) schwache, (3) absolute Begriff der Wahrheit" ist.

### 8.2.1 Logisch vs. nichtlogisch

Nehmen wir an, daß Sätze Wahrheitswertträger sind, und fragen wir uns, ob der Satz 'Schnee ist weiß oder Schnee ist nicht weiß' notwendigerweise wahr ist. Im Logikunterricht bekommt man eine bejahende Antwort. Also würde man dem Satz (S1) zustimmen wollen:

[114]Zwei Anmerkungen hierzu: 1.) Tatsächlich erwägen Belnap und Gupta auch die Möglichkeit, daß eine Sprache L das Prädikat 'ist kategorisch (in L)' enthält. Die für eine solche Sprache zu entwickelnde Semantik wird aber wieder in einer Metasprache vorgenommen, in der ein *anderes* Prädikat 'ist kategorisch' erklärt wird. Mit der so entwickelten Semantik läßt sich dann wieder einsehen, daß das in L befindliche Prädikat 'ist kategorisch (in L)' einen zirkulären Begriff ausdrückt. 2.) Die Beibehaltung einer Unterscheidung von Objekt- und Metasprache wirft natürlich die Frage auf, ob Guptas und Belnaps Zugang nicht im wesentlichen der Tarskische Zugang ist – lediglich um eine Ebene verschoben. Auf diese Kritik gehen Gupta und Belnap ein. ([18], S.256)

(S1) Der Satz ‘Schnee ist weiß oder Schnee ist nicht weiß’ ist notwendigerweise wahr.

Gleichzeitig wird man aber auch jemandem zustimmen wollen, der den folgenden Satz (S2) äußert:

(S2) Wenn ‘oder’ das bedeutet hätte, was ‘und’ bedeutet, dann wäre der Satz ‘Schnee ist weiß oder Schnee ist nicht weiß’ nicht wahr gewesen.

Allerdings ergibt sich ein Problem, wenn man weiterhin an den beiden Sätzen (S1) und (S2) festhalten möchte: Aus (S2) läßt sich folgern, daß es möglich ist, daß der Satz ‘Schnee ist weiß oder Schnee ist nicht weiß’ nicht wahr ist, was im Widerspruch zu (S1) besagt, daß der Satz ‘Schnee ist weiß oder Schnee ist nicht weiß’ nicht notwendigerweise wahr ist. Belnap und Gupta gemäß läßt sich der Widerspruch mit der Unterscheidung zwischen dem logischen[115] und nicht-logischen Wahrheitsbegriff auflösen: Das in (S1) verwandte Prädikat ‘ist wahr’ drückt den logischen Wahrheitsbegriff aus; dieses Prädikat trifft in einer möglichen Welt w auf einen Satz, i.e. eine bestimmte wohlgeformte Zeichenkette, zu, wenn dieser in der Bedeutung, die er in unserer Welt hat, korrekt in w ist. In (S2) hingegen drückt das Prädikat ‘ist wahr’ den nichtlogischen Wahrheitsbegriff aus; dieses Prädikat trifft genau dann auf einen Satz in einer möglichen Welt w zu, wenn dieser in der Bedeutung, die er in w hat, korrekt in w ist. Natürlich kann man sich fragen, ob sich dieser Widerspruch nicht auch anders lösen läßt, indem man z.B. anführt, daß eine bestimmte Lesart von ‘möglich’ bzw. ‘ist notwendig’ zugrunde zu legen ist, die es verbietet, aus (S2) den Satz zu folgern ‘Es ist möglich, daß der Satz ‘Schnee ist weiß oder Schnee ist nicht weiß’ nicht wahr ist’. Dann hätte man explizit eine Unterscheidung nicht für ‘ist wahr’, sondern für ‘ist möglich’ bzw. ‘ist notwendig’ getroffen.[116] Eine viel natürlichere Auflösung als die von Gupta und Belnap gegebene ist, (S1) als elliptisch formuliert anzusehen und es auszubuchstabieren mit

($S1^*$) Der Satz ‘Schnee ist weiß oder Schnee ist nicht weiß’, so wie er jetzt im Deutschen verstanden wird, ist notwendigerweise wahr.

### 8.2.2 Schwach vs. stark

Die hier intendierte Unterscheidung schwach vs. stark geht nach Belnap und Gupta auf Yablos Aufsatz [49] zurück. Für den schwachen Begriff der Wahrheit gilt, daß der semantische Status von ‘„P“ ist wahr’ mit dem von P identisch ist.

[115] Eine Anmerkung zur Terminologie: Die Ausdrücke ‘logisch’ und ‘nicht-logisch’ für die im folgenden zu gebende Unterscheidung ist denkbar ungünstig und wird von Belnap und Gupta nicht motiviert. Besser geeignet wäre eine Redeweise, die auf die Einbeziehung der Modallogik aufmerksam zu machen vermag – etwa aktual vs. nicht-aktual

[116] Wenn man sich die Realisierung dieser Unterscheidung im Detail überlegt, könnte sich allerdings herausstellen, daß damit eine Unterscheidung des Wahrheitsprädikats impliziert wird.

Für den starken Begriff der Wahrheit gilt das nicht mehr. Für den mag '„P“ ist wahr' ein falscher Satz sein, selbst wenn P nicht falsch, sondern vielleicht weder wahr noch falsch ist.

Hier muß man natürlich hellhörig werden, da Gupta und Belnap doch beanspruchen, den Wahrheitsbegriff der Umgangssprache zu erfassen. Für den gilt doch, daß man mit seiner Hilfe folgendes zum Ausdruck bringen kann:

(*) Der Satz 'Die Anzahl der Bäume auf der kannadisch-US-amerikanischen Grenze ist gerade' ist weder wahr noch falsch.

Hier wird man sagen wollen, daß man den starken Wahrheitsbegriff verwendet. Diesen Satz würde man bejahen wollen, da nicht genau festgelegt ist, wo die Grenze verläuft, wie breit sie ist, als daß man sagen könnte dieser oder jener Baum gehört zur Grenze; und wieviel vom Baum muß sich auf der Grenze befinden, damit es als auf der Grenze befindlich angesehen wird? Außerdem ist nicht klar, was denn nun als Baum zählt. Eine umfassende Wahrheitstheorie müßte diesem sprachlichen Faktum Rechnung tragen können: Wir haben die Möglichkeit, mit dem Ausdruck 'wahr' bzw. 'falsch' den obigen Satz zu formulieren, ohne dabei in Widersprüche zu geraten – eine Eigenschaft des umganssprachlichen Wahrheitsbegriffs, die so manche deflationäre Wahrheitstheorie nicht erklären bzw. mit ihrer Konzeption abbilden kann.[117] Die Revisionstheorie der Wahrheit wird solch ein Phänomen mit ihrer Wahrheitskonzeption ebenfalls nicht erklären können, wenn sie mit dem schwachen Wahrheitsbegriff operiert. Sie wird die Gründe anderswo suchen müssen.

Ebenso wie im Falle des obigen Satzes ist es naheliegend zu sagen, daß jemand den Wahrheitsbegriff im starken Sinne gebraucht, der behauptet, der Satz 'Der gegenwärtige König von Frankreich ist kahlköpfig' sei weder wahr noch falsch. In einer Fußnote[118] merken Belnap und Gupta aber sofort an, daß man nicht gezwungen ist, es so zu lesen. Wenn es eine plausible Begründung hierfür gibt, würde das ein Stückchen mehr für die Erklärungskraft von RTT sprechen. Denn wenn sich herausstellen sollte, daß mit dem schwachen Begriff der Wahrheit nicht das ausgedrückt werden kann, was mit dem starken ausdrückbar ist, dann könnte nicht ausgeschlossen werden, daß gerade diejenigen Dinge, die den Philosophen interessieren, mit dem schwachen Begriff der Wahrheit nicht einzufangen sind. Gupta und Belnap schreiben:

> The strong reading is not forced, however. We can defend the weak reading if we interpret 'neither P nor Q' so that its truth does not require the falsity of the components P and Q. In general, the weak notion of truth carries more information than any of the strong notions and can do their work if the language is rich in logical resources. ([18], S. 22, Fn 40)

[117] Die deflationäre Wahrheitstheorie von C.F.J. Williams ([47]) z.B. kann das nicht erklären.
[118] [18], S.22, Fußnote 40

Daß Belnap und Gupta erklären müssen, wie genau 'Weder P noch Q' zu lesen ist, wenn sie mit dem schwachen Wahrheitsbegriff arbeiten wollen, zeigt ein sonst drohender Widerspruch. Für die Zwecke des folgenden Arguments soll 'Sem(p)' den semantischen Status (z.B einen Wahrheitswert) des Satzes p bezeichnen.

1. Sem('p' ist wahr) = Sem(p)
2. Sem('p' ist falsch) = Sem(¬ p)
3. Weder ('p' ist wahr) noch ('p' ist falsch)
4. Also (?): $\neg p$ & $\neg(\neg p)$

Die erste Prämisse trägt der Tatsache Rechnung, daß Belnap und Gupta mit dem schwachen Begriff der Wahrheit arbeiten wollen. Es stellt sich nun die Frage, wie Belnap und Gupta den Begriff der Falschheit erklären wollen und wie 'Weder P noch Q' zu erläutern ist. Wenn sie keine vernünftige Lesart angeben, dann droht bei falscher Lesart ein Widerspruch. Problematisch ist die zweite Prämisse: Hier taucht das Zeichen '¬' auf. Dieses kann aber nicht das Zeichen für die klassische Negation sein, da wir jetzt mindestens in einer dreiwertigen Logik arbeiten, in der es neben den Wahrheitswerten w für wahr und f für falsch auch einen Wert n für „wahrheitswertlos" gibt, mit dem wir andeuten können, daß der semantische Status des Satzes p = 'Der König von Frangreich ist glatzköpfig' Sem(p) = n ist. Die Negation muß also auch für diesen Wert n erklärt werden. Und hier gibt es zwei echte Alternativen, die den Begriff der Negation nicht zu stark verzerren. In der Folge ergeben sich die externe Negation $\neg_{ext}$ und die interne Negation $\neg_{int}$. Ersteres angewandt auf einen Satz p, der den Wahrheitswert n hat, ergibt Sem($\neg_{ext}p$) = w. Letzters angewandt auf einen Satz p mit dem Wahrheitswert n ergibt hingegen Sem($\neg_{int}p$) = n. Belnap und Gupta wollen an dem schwachen Wahrheitsbegriff festhalten, also müssen sie konsequenterweise auch den schwachen Begriff der Falschheit wählen. Der würde dann angewandt auf einen Satz p, der den Wahrheitswert n hat, einen Satz ergeben, der ebenfalls den Wert n hat. D.h. für den schwachen Begriff der Falschheit gilt folgendes:
Sem('p' ist falsch) = Sem($\neg_{int}p$). Belnap und Gupta haben also in der zweiten Prämisse des obigen Arguments das ¬ als interne Negation $\neg_{int}$ zu lesen. Und dennoch haben sie die Möglichkeit, dem Widerspruch zu entkommen, indem sie nämlich das 'Weder P noch Q' in der dreiwertigen Logik entsprechend stark interpretieren. Würden sie die dritte Prämisse lesen als

(3a) Es ist falsch, daß ('p' ist wahr) & Es ist falsch, daß ('p' ist falsch)

dann würden sie dem Widerspruch nicht entkommen. Lesen sie (3) hingegen als

(3b) $\neg_{ext}$('p' ist wahr) & $\neg_{ext}$('p' ist falsch)

dann droht nicht mehr der Widerspruch. Würden Belnap und Gupta statt der externen Negation die interne Negation wählen, dann liefe das im Prinzip auf (3a) hinaus und sie gerieten wieder in einen Widerspruch. Das ist intuitiv auch klar: Sind die beiden Begriffe „wahr“ und „falsch“ schwach, dann müssen Belnap und Gupta die starke Negation, d.h. die externe Negation wählen, um dem Widerspruch zu entfliehen, und nicht die schwache, also interne Negation.

Die englische oder deutsche Sprache hat eine große logische Ausdruckskraft; damit dürfte es nach Meinung von Gupta und Belnap keinen wesentlichen Unterschied machen, ob das Wahrheitsprädikat der natürlichen Sprache nun in der schwachen oder starken Lesart interpretiert wird.[119]

Wie ist die Behauptung zu verstehen, der schwache Wahrheitsbegriff enthalte grundsätzlich mehr Information als der starke Wahrheitsbegriff? Belnap und Gupta geben nicht an, wie sie hier ‘Information’ verstehen wollen. Die Ausführungen an einer anderen Stelle in ihrem Buch, in denen ein Informationsbegriff eine Rolle spielt, führen hier nur in die Irre.[120] Ich glaube, man muß Belnap und Gupta etwa folgendermaßen verstehen: Wenn man annimmt, daß bei der Anwendung des starken Wahrheitsprädikats auf einen Satz sich immer ein Satz ergibt, der entweder den Wahrheitswert w (für wahr) oder den Wahrheitswert f (für falsch) enthält, dann würde das starke Wahrheitsprädikat eine Reduktion des semantischen Status des eingebetteten Satzes auf die beiden Wahrheitswerte w und f bewirken: Wenn man z.B. gesagt bekäme, daß der Output-Satz den Wahrheitswert f hat, dann könnte man mit dieser Information nicht herausfinden, ob der Input-Satz in ein Wahrheitswertloch fällt oder den Wahrheitswert w hat. Da beim schwachen Wahrheitsbegriff der semantische Status von Output- und Inputsatz derselbe ist, hätte man kein derartiges Problem. Man hat also beim schwachen Wahrheitsprädikat in dem Sinne mehr Information, daß man auf jeden Fall aus dem semantischen Status des komplexen Satzes auf den semantischen Status des eingebetteten Satzes schließen kann. Beim starken Wahrheitsbegriff ist das nicht mehr der Fall.

### 8.2.3 Absolut vs. modellrelativ

Neben dem „üblichen“ Wahrheitsprädikat gibt es auch auf Modelle M relativierte Wahrheitsprädikate ‘ist wahr in einem Modell M’. Der übliche Wahrheitsbegriff, i.e. der absolute Wahrheitsbegriff, ist sozusagen die Wahrheit in dem einzigen Modell, das die aktuale Welt repräsentiert. Gupta und Belnap machen darauf aufmerksam, daß man in keiner Richtung (auf offensichtliche Weise) die Defini-

[119] Allerdings sollte man Gupta und Belnap nicht die Meinung unterschieben, in der Umgangssprache genieße die schwache Lesart Priorität. Sie behaupten lediglich, mit dem schwachen Begriff der Wahrheit ließe sich alles das ausdrücken, was man mit dem starken Begriff ausdrücken kann.

[120] Siehe [18], S.34

tion des einen aus der Definition des anderen erhält.[121] Daß es nur der absolute Wahrheitsbegriff sein kann, der mit den partiellen Definitionen in Form der T-Bikonditionale definiert werden soll, sei bereits daraus zu ersehen, daß normalerweise auf der rechten Seite der T-Bikonditionale keine Variable vorhanden ist, die für die Modelle M frei ist. Mit dieser dreifachen Unterscheidung sehen Gupta und Belnap das Wahrheitsprädikat, welches Gegenstand ihrer Untersuchung und der hieraus entwickelten Revisionstheorie sein soll, hinreichend eingegrenzt.

### 8.2.4 Die Signifikationsthese

In der Terminologie von Gupta und Belnap muß eine Wahrheitstheorie die Signifikation des Wahrheitsprädikates angeben. Der Begriff der Signifikation wird dabei folgendermaßen erklärt:

> Let the *(extensional) signification* of an expression (or concept) in a world w be an abstract something that carries all the information about the expression's extensional relations in w. ([18], S.30)

Beispielsweise besteht für ein klassisches Prädikat die Angabe seiner Signifikation für eine mögliche Welt w in der Angabe seiner Extension in der Welt w. Die Angabe der Signifikation eines dreiwertigen Prädikates G besteht in der Angabe seiner Extension und Antiextension, wobei letzteres die Menge derjenigen Dinge a umfaßt, für die die Aussage 'Ga' den Wahrheitswert falsch erhält.[122] Zentral für die Entwicklung einer Wahrheitstheorie ist nach Gupta und Belnap die Signifikationsthese. Sie besagt folgendes:

> **Signifikationsthese**
> Die T-Bikonditionale fixieren die Signifikation der Wahrheit in jeder möglichen Welt.

Diese These sei eigentlich eher als Gleichung oder als ein Schema anzusehen, das erst durch eine korrekte Erklärung der beiden in ihm enthaltenen Ausdrücke 'T-Bikonditional' und 'Signifikation der Wahrheit' zu einer eindeutigen (und wahren) Aussage werde.

Motiviert ist die Signifikationsthese durch die Implikationsthese, die Gupta und Belnap als zu stark ansehen und daher ablehnen:[123]

> **Implikationsthese**
> Eine Definition der Wahrheit sollte sämtliche T-Bikonditionale implizieren.

[121] [18], S.23

[122] Diese Begrifflichkeit ist z.B. für das folgendermaßen definierte Prädikat G nützlich: Wenn $15 \leq x$, dann Gx und wenn $x \leq 13$, dann $\neg$ Gx. Hiernach wäre z.B. die Zahl 14 weder in der Extension noch in der Antiextension von G.

[123] Die Gründe für die Ablehnung der Implikationsthese und den Nachweis, daß die Signifikationsthese neutraler ist als die Implikationsthese, will ich nicht besprechen, obwohl Belnap und Gupta tatsächlich eine Begründung geben ([18], S.25-29).

Es liegt nahe, beim Ausdruck 'T-Bikonditional' an die Tarskischen W-Äquivalenzen zu denken und die Implikationsthese als eine aus dem Tarskischen Adäquatheitskriterium folgende These anzusehen. Den Ausdruck 'T-Bikonditional' wollen Gupta und Belnap aber so verwenden, daß nicht ausgeschlossen wird, daß auf der linken Seite der Bikonditionale Namen für Sätze der Sprache L vorkommen, die das im Bikonditional vorkommende Wahrheitsprädikat enthalten. Mit dieser liberalen Lesart droht natürlich wieder das fatale Lügnerparadox. Daher müsse man sich überlegen, wie genau man die „T-Bikonditionale" interpretiert, damit man nicht in Widersprüche gerät. Gupta und Belnap wollen an der Intuition festhalten, daß die T-Bikonditionale, was für Gebilde das auch sein mögen, die Signifikation des Wahrheitsprädikates festlegen. Um der drohenden Antinomie aus dem Wege zu gehen, beschreiten sie nicht den Weg Tarskis, das Vorkommen des Wahrheitsprädikates in der Sprache, für die es erklärt werden soll, zu verbieten, sondern der Intuition durch eine Explikation der Ausdrücke 'T-Bikonditional' und 'Signifikation' auf die Beine zu helfen.

In der Signifikationsthese kommt der modallogische Terminus 'mögliche Welt' vor. Dies hat jedoch nicht unbedingt zur Folge, daß man in einem modallogischen System arbeiten muß. Da das wahrheitsfreie Sprachfragment L' der Sprache L, für die das Wahrheitsprädikat erklärt werden soll, vollkommen extensional (im oben erläuterten Sinne) ist, ist der semantische Status eines jeden Satzes aus L' vollständig dadurch festgelegt, daß angegeben wird, a) welche Dinge die Namen aus L' bezeichen, b) welche Extensionen die Prädikate aus L' haben und welche Funktionen den Funktionsbuchstaben aus L' zugeordnet sind. Das rechtfertigt zumindest für das Sprachfragment L', die möglichen Welten (bzw. möglichen Situationen) mit klassischen Interpretationen (Modellen/Strukturen) für L' zu identifizieren.

Zur Veranschaulichung ein Vergleich aus der Logik: In der Logik beschäftigt man sich mit Argumenten und versucht systematisch die „guten" Argumente zu erfassen. Die erste informale Erläuterung des Schlüssigkeitsbegriffs ist ein erster Schritt auf dem Weg zu diesem Ziel. Hiernach ist ein Argument genau dann schlüssig, wenn es unmöglich ist, daß alle Prämissen wahr sind, die Konklusion aber falsch ist. Diesen Schlüssigkeitsbegriff versucht man in der klassischen Aussagen- und Quantorenlogik dadurch zu fassen, daß man genau die Argumente für formal schlüssig erklärt, deren zugehörige Argumentschemata korrekt sind. Korrekheit wiederum wird unter Rückgriff auf den Interpretations- bzw. Modellbegriff erläutert: Ein Argumentschema ist korrekt gdw es kein Modell gibt, in dem alle Prämissen wahr sind, die Konklusion aber falsch ist. Es ist hier keine Rede mehr davon, daß etwas möglich oder unmöglich ist etc. Diese formale Explikation des Schlüssigkeitsbegriffs eignet sich für „extensionale Verhältnisse", wie man sie meist in den semiformalen Sprachfragmenten der Mathematiker findet. Sobald man aber Sprachen vorliegen hat, in denen intensionale Konstruktionen gestattet sind, versagt die Explikation der klassischen Logik. Z.B. kann man dann nicht mehr nachvollziehen, daß das Argument 'Es ist notwendigerweise wahr, daß

Schnee weiß ist. Also: Schnee ist weiß' ein (formal) schlüssiges Argument ist.

Fragen wir uns nun, ob diese Identifikation von Modellen mit möglichen Situationen auch für die gesamte, das Wahrheitsprädikat beinhaltende Sprache L gewährleistet ist. Daß es tatsächlich möglich ist, führen Belnap und Gupta auf eine Eigenschaft des Wahrheitsbegriffs zurück, die sie die „Supervenienz der Signifikation der Wahrheit" („supervenience of the signification of truth") nennen.[124] Daß Wahrheit diese Eigenschaft hat, ist eine Art metaphysische Annahme, die Belnap und Gupta treffen. Diese beinhaltet, daß es zur Bewertung des Status von Sätzen einer Sprache erster Stufe, welches das eigene Wahrheitsprädikat enthält, nicht nötig ist, eine semantische Interpretation des Wahrheitsprädikates zu geben. Der Status eines jeden Satzes (einer Sprache erster Stufe mit eigenem Wahrheitsprädikat) – und das hieß hier, ob der Satz problematisch ist und wenn unproblematisch, ob er wahr oder falsch ist – ist durch Fakten bestimmt, die grob gesagt nicht den Wahrheitsbegriff beinhalten. Solche Fakten heißen nichtsemantische Fakten. Z.B. ist die Tatsache, daß Schnee weiß ist, ein nichtsemantisches Faktum. Ein Gegensbespiel zu nichtsemantischen Fakten ist z.B. die Tatsache, daß der Satz 'Zwei und Zwei ist Vier' wahr ist. In dieses Faktum „geht ein" der Wahrheitsbegriff. Modelle des wahrheitsfreien Fragmentes enthalten sämtliche nichtsemantische Information.[125] Folglich gilt, daß die möglichen Welten mit den Modellen des wahrheitsfreien Fragmentes identifiziert werden können. Damit besagt die Signifikationsthese für Sprachen erster Stufe L mit eigenem Wahrheitsprädikat, daß die T-Bikonditionale für jedes Modell des wahrheitsfreien Fragmentes die Signifikation des Wahrheitsprädikates von L festlegt.

## 8.3 Die Revisionstheorie der Wahrheit (RTW)

In diesem Abschnitt möchte ich darstellen, wie Belnap und Gupta die (RTD) auf den Begriff der Wahrheit anwenden, um so die oben dargelegte Aufgabe zu lösen. Wieder werde ich dabei von dem einfachen semantischen System $S_0$ und dem zugehörigen Kalkül $C_0$ ausgehen. Für eine korrekte Beurteilung von Guptas und Belnaps Wahrheitstheorie müssen aber die komplizierteren semantischen Systeme $S^*$ und $S^\#$ zugrunde gelegt werden. Das wird bei der Besprechung der Einwände gegen (RTW) und (RTD) im letzten Kapitel meiner Arbeit nachgeholt werden; ich werde die hier relevanten Ergebnisse dann schlichtweg referieren.

Im folgenden Abschnitt will ich dann das Argument von Belnap und Gupta dafür, daß der logische, schwache, absolute Begriff der Wahrheit zirkulär ist, genauer betrachten.

---

[124] [18], S.18; 87f.

[125] Ein Beispiel: Der Status des Satzes "Sokrates ist weise' ist wahr' ist vollkommen dadurch festgelegt, ob Sokrates weise ist oder nicht – und diese Information ist in dem Modell des wahrheitsfreien Fragmentes enthalten, das uns sagt, welchen Gegenstand 'Sokrates' denotiert und was die Extension von 'ist weise' ist.

### 8.3.1 Syntax

Die formalen Sprachen, welche Belnap und Gupta betrachten, beinhalten im Unterschied zu den von Tarski behandelten Sprachen gewisse universelle Züge: Jedes dieser Sprachen enthält ein Wahrheitsprädikat und Anführungsnamen für alle in ihr enthaltenen Sätze. Ansonsten sind die Verhältnisse wie für klassische Sprachen der Quantorenlogik. Das Vokabular der betrachteten Sprachen enthält also als Komponenten:

*Technische Zeichen*
Es liegen übliche Klammern und Variablen (für Quantifikation über den Gegenstandsbereich) vor.

*Logische Zeichen*
Es stehen Zeichen (Ausdrücke) für die üblichen Junktoren der Aussagenlogik zur Verfügung, also Zeichen für das 'und', 'oder', 'wenn-dann', 'Es ist nicht der Fall, daß', 'genau dann, wenn'. Außerdem liegen Zeichen für die beiden Quantoren 'Alle' und 'Es existiert' vor; meist auch ein Zeichen für die Identität und schließlich ein Zeichen für das Wahrheitsprädikat – meist 'T' oder 'ist wahr'.

*Nichtlogische Zeichen*
Es stehe eine (möglicherweise leere) Menge an Prädikat- und Namenbuchstaben zur Verfügung. Darüber hinaus gibt es eine abzählbare Menge an Anführungsnamen, die links mit einem linken Hochkomma beginnen, gefolgt von einem Satz dieser Sprache, und mit einem rechten Hochkomma enden.[126] (Funktionszeichen soll es der Bequemlichkeit halber nicht geben).

Mit diesem Vokabular wird schließlich in bekannter Weise erklärt, was ein Term, eine wohlgeformte Formel und ein Satz (= wohlgeformte Formel ohne freies Variablenvorkommnis) ist.

### 8.3.2 Semantik

Für die Semantik von Sprachen L, deren Syntax gerade vorgestellt wurde, ist der Begriff eines Basismodells („ground model") wichtig.

**Basismodell**
$M$ ist ein Basismodell einer Sprache L $\Leftrightarrow_{Df}$ $M =< D, I >$, wobei $D$ der Gegenstandsbereich und $I$ eine partielle Interpretation von L ist, und M erfüllt die folgenden Bedingungen:

- Der Wertebereich $D$ enthält sämtliche Sätze von L.

[126] Bei exakter Formulierung muß natürlich angegeben werden, wie man solche Namen erhält. Siehe hierzu z.B. [50], S. 44-45

- Jeder Namenbuchstabe $m$ von L wird durch ein Objekt $I(m)$ aus $D$ interpretiert.
- Für Anführungsnamen '$s$' (mit einem Satz s aus L) gilt $I(\text{'}s\text{'})=s$.
- Jeder n-stellige Prädikatbuchstabe $R$, der nicht das Wahrheitsprädikat 'T' ist, wird durch eine Menge $I(R)$ von n-Tupeln von Gegenständen aus $D$ interpretiert.
- Jeder T-freie Satz aus L erhält in M einen klassischen Wahrheitswert, für dessen Festlegung Tarskis semantischer Begriff der Erfüllung zum Einsatz kommt.[127]

Bei Basismodellen handelt es sich nicht um Modelle (Strukturen) im klassischen Sinne, da sie das Wahrheitsprädikat 'T' uninterpretiert lassen. Die Aufgabe besteht nun darin, in einem Basismodell zu erklären, wie 'T' zu interpretieren ist, damit man korrekterweise von einem Wahrheitsprädikat sprechen kann.

Hier kommen nun die mehrfach erwähnten T-Bikonditionale ins Spiel. Als T-Bikonditionale wollen Belnap und Gupta Sätze der folgenden Form verstanden wissen:

> s ist wahr $\leftrightarrow$ A

's' steht dabei für einen Namen eines Satzes einer Sprache L, dessen Übersetzung in *dieselbe* Sprache L durch 'A' gegeben ist. Eine korrekte Instanz dieses Schemas für ein geeignetes Fragment des Deutschen wäre z.B. der folgende Satz:

> "Schnee ist weiß' ist wahr' ist wahr $\leftrightarrow$ 'Schnee ist weiß' ist wahr.

Die Idee ist nun, diese T-Bikonditionale als partielle Definitionen aufzufassen.[128] Die Idee, die T-Bikonditionale als partielle Definitionen aufzufassen, habe bereits Tarski gehabt. Dieser schreibt in seinem Aufsatz [43], §4:

> Es sollte hervorgehoben werden, daß weder der Ausdruck der Form (W) [i.e.: X ist wahr genau dann, wenn p] selbst (der keine Aussage, sondern das Schema einer Aussage ist) noch irgendein besonderer Fall der Form (W) als Definition der Wahrheit angesehen werden kann. *Wir können nur sagen, daß jede Äquivalenz der Form (W), die wir nach Ersetzung von 'p' durch eine bestimmte Aussage und von 'X' durch den Namen dieser Aussage erhalten, als eine partielle Definition der Wahrheit betrachtet werden*

[127] Wie das genau geht, wurde im Abschnitt über Tarskis Wahrheitstheorie dargestellt.

[128] Gupta und Belnap halten partielle Definitionen für (vielleicht nicht unbedingt interessant aber doch für) legitim ([18], S.131-132; 197). Die partielle Definition eines einstelligen Prädikats 'G' legt für einen einzelnen Gegenstand b die Bedingungen fest, unter denen b in der Extension von G liegt:

$$\text{G(b)} \quad =_{Df} \quad \text{A(b,G)}$$

Diese Definition sagt also, daß b genau dann in der Extension von G liegt, wenn es den in einer Variablen x offenen Satz A(x, G), der eventuell das Prädikat 'G' enthält, erfüllt.

> *kann, die erklärt, worin die Wahrheit dieser einen individuellen Aussage besteht. Die allgemeine Definition muß in einem gewissen Sinne die logische Konjunktion all dieser partiellen Definitionen sein.* (Mit einer kleinen Änderung, meiner Hervorhebung und meinem Zusatz in eckigen Klammern zitiert nach [39], S.146)

Die formale Korrektheit, welche Tarski von der Definition der Wahrheit gefordert habe, schließe aber aus, daß die T-Bikonditionale als partielle Definitionen aufgefaßt werden können. Gupta und Belnap erkennen hier einen Konflikt zwischen der Forderung der formalen Korrektheit und dem Adäquatheitskriterium. Dieser Konflikt sei stets zu Gunsten der formalen Korrektheit entschieden worden. Sie selbst, behaupten Gupta und Belnap, würden den entgegengesetzten Weg gehen und das Adäquatheitskriterium (mit einer gewissen Modifikation) akzeptieren. Tatsächlich hat aber Tarski sein Kriterium, welches Belnap und Gupta unter dem Titel 'Convention T' anführen[129], für W-Äquivalenzen aufgeführt; bei diesen ist schon eine Unterscheidung zwischen Objekt- und Metasprache für das Wahrheitsprädikat involviert. Hier kann es nicht zu Problemen wie dem der Lügnerparadoxie kommen. Insofern ist nicht zu sehen, weshalb es bei Tarski einen Konflikt zwischen seiner Konvention W und der formalen Korrektheit für die W-Äquivalenzen gibt. Für die als echte Bikonditionale aufgefaßten T-Bikonditionale von Belnap und Gupta besteht eingestandenermaßen diese Problematik noch. In ihrer Konvention T ist auch tatsächlich die Rede von T-Bikonditionalen und nicht von den Tarskischen W-Äquivalenzen. Ihr Gedanke ist wahrscheinlich, daß Tarski für seine W-Äquivalenzen zunächst keine Unterscheidung von Objekt- und Metasprache gemacht hatte, daß er aber, da er gesehen hat, daß sich dann Probleme mit seiner Forderung nach inhaltlicher Adäquatheit ergeben würden, diese Unterscheidung nachgeholt und damit die drohende Zirkularität ausgeschlossen hat.

Die T-Bikonditionale werden als Definitionen aufgefaßt, d.h. im Unterschied zu Tarskis W-Äquivalenzen sind sie keine echten Bikonditionale[130]: Es wird also ein Unterschied gemacht werden müssen zwischen den Sätzen aus einer Sprache L, die von der Form

s ist wahr $\leftrightarrow$ A

sind, und Sätzen, die von der Form

s ist wahr $=_{Df}$ A

sind.

Denken wir uns also für eine Sprache L, wie sie oben charakterisiert wurde, die folgenden T- Bikonditionale gegeben:

[129] [18], S.2

[130] Dann ist aber nicht zu verstehen, weshalb Belnap und Gupta noch von Bikonditionalen sprechen und außerdem statt des näherliegenden Zeichens $\Leftrightarrow_{Df}$ das Zeichen $=_{Df}$ wählen. Eine Begründung geben sie nicht.

$s_1$ ist wahr $=_{Df}$ $A_1$

$s_2$ ist wahr $=_{Df}$ $A_2$

$s_3$ ist wahr $=_{Df}$ $A_3$

$\vdots$

Die '$A_i$' können nun Vorkommnisse des Ausdrucks 'ist wahr' enthalten, so daß einige der partiellen Definitionen zirkulär sein würden. In Belnaps und Guptas Theorie zirkulärer Definitionen kann man aber auch solche Definitionen behandeln.

Sei M ein Basismodell. Dann wird durch diese partiellen Definitionen in M ein Operator $\tau_M$ festgelegt, der als „Tarski jump" bezeichnet wird. Ihn könnte man so definieren:

> Sei M die Gesamtheit der relevanten Fakten und U eine willkürlich gewählte Menge. Dann ist $s_i \in \tau_M(U) \quad \Leftrightarrow_{Df} \quad A_i$ ist wahr in M unter der Hypothese, daß die Extension von 'ist wahr' U ist, wobei '$A_i$' das Definiens in der partiellen Definition von '$s_i$ ist wahr' ist.[131]

Gupta und Belnap schreiben, daß man den Operator $\tau_M$ auch anders charakterisieren könne:

> Another characterization of $\tau_M$: It is the rule of revision yielded by the infinitistic circular definition
>
> x ist wahr $=_{Df}$
> (x=$s_1$ und $A_1$) oder (x=$s_2$ und $A_2$) oder (x=$s_3$ und $A_3$) ... ([18], S. 133)

[131] Diese Definition ist zwar eingängig, aber nicht hinreichend präzise formuliert. Die korrekte Definition geben Gupta und Belnap im Abschnitt 2B. Hier wird allgemein für ein Bewertungsschema $\varrho$ der Sprungoperator in einem Basismodell M, $\varrho_M$, erklärt. (Unter einem Bewertungsschema $\varrho$ verstehen Gupta und Belnap dabei die Spezifikation der Bedeutung der logischen Konstanten einer Sprache - etwa durch Angabe von Wahrheitstabellen. $\tau$ z.B. ist das klassische Bewertungsschema, das den logischen Konstanten die in der klassischen Quantorenlogik übliche Bedeutung gibt. Ein anderes Schema würde z.B. einen zusätzlichen Wahrheitswert **n** für 'weder wahr noch falsch' einbeziehen und die Bewertung der logischen Konstanten unter Berücksichtigung dieses zusätzlichen Wertes vornehmen.) Damit kann der Tarski Sprungoperator korrekt so erklärt werden: Sei M $= < D, I >$ ein Basismodell für L. $\mathcal{L}_g$ sei die interpretierte Sprache $< L, M + g, \tau >$, die über die Interpretation durch M hinaus den Prädikatbuchstaben G mit dem Prädikat g interpretiert. Dann ist der Tarski Sprungoperator $\tau_M$ eine Operation auf der Klasse der möglichen Interpretationen g eines einstelligen Prädikatbuchstabens G in M, der für alle d $\in$ D folgende Bedingung erfüllen muß:

$$\tau_M(g)(d) = \begin{cases} Val_{\mathcal{L}_g}(A) & \text{falls } d \text{ (der Code von } A\text{) ist} \\ \mathbf{f} & \text{sonst} \end{cases}$$

Dabei ist $Val_{\mathcal{L}_g}(A)$ der unter dem Modell M+g der Formel $A$ zugewiesene Wahrheitswert – entweder **t** oder **f**.

Probleme mit der Tatsache, daß Gupta und Belnap die Wahrheitsprädikate für klassische Sprachen der Quantorenlogik erklären, in denen keine unendlichen Disjunktionen vorkommen, ergeben sich mit dieser Charakterisierung nicht. Denn die Definitionen für eine Sprache L sind Guptas und Belnaps Konzeption gemäß keine Bestandteile dieser Sprache. Im Gegensatz zur klassischen Definitionstheorie kommen Definitionen nicht als eine besondere Sorte von Sätzen oder Formeln unter den Formeln von L vor. Auch wenn sie bestimmte Zeichen aus dem Vokabular von L enthalten, so ist doch das in ihnen enthaltene Definitionszeichen '$=_{Df}$' keine Komponente des (logischen) Vokabulars von L.

Allerdings muß man mit dieser Definition vorsichtig umgehen, da nicht klar ist, (1) ob mit ihr eine Quantifikation über x vorgenommen wird, die nicht explizit aufgeführt ist, (2) wie die Quantifikation zu verstehen ist, wenn denn eine vorliegt, und (3) worüber die Quantifikation verläuft, was also der Gegenstandsbereich bzw. der Substitutionsbereich sein sollte.

Nun da man die partiellen Definitionen für das Wahrheitsprädikat T in L aufgestellt hat, lassen sich sämtliche Sätze der Sprache L mit der für zirkuläre Definitionen entwickelten Revisionssemantik behandeln.

### 8.3.3 Ein Beispiel

Betrachten wir mit Gupta und Belnap an einem Beispiel, wie der Operator $\tau_M$ arbeitet.[132] Sei L ein Fragment der deutschen Sprache, das folgende logische Operatoren enthält: Das **ist** der Identität und der Prädikation, **nicht**, **und**, **oder**, **Wenn...dann...**, **Alles**, **Etwas**. Neben dem einstelligen Prädikat **ist wahr** bilden folgende Elemente das nichtlogische Vokabular.

> *Namen:* **Schnee**, **Jones**, und Anführungsnamen für alle Sätze der Sprache L.
>
> *Einstellige Prädikate:* **ist weiß**.
>
> *Zweistellige Prädikate:* **äußert**.

Es wird nun ein Modell M betrachtet, das durch folgende Angaben festgelegt wird:

> *Wertebereich:* D = {Schnee, Jones} vereinigt mit der Menge aller Sätze aus L.
>
> *Interpretation der Namen:* I(**Schnee**) = Schnee; I(**Jones**) = Jones; für die Anführungsnamen gilt, daß sie die mit ihnen intendierte Interpretationen haben, also I('**S**') = **S**
>
> *Interpretation der 1-stelligen Prädikate:* I(**weiß**) = {Schnee }.

[132] [18], S.133f.

*Interpretation der 2-stelligen Prädikate:* I(**äußert**) = { < Jones, **Schnee ist weiß** >, < Jones, **'Schnee ist weiß' ist wahr**>, < Jones, **Etwas, was Jones äußert, ist wahr** > }.

Mit Hilfe der Funktion $\tau_M$ läßt sich jetzt feststellen, welcher der Sätze aus dem Wertebereich D als wahr herausgegeben werden, wenn man als Anfangshypothese für das Prädikat **ist wahr** irgendeine Teilmenge von D wählt. Wählt man z.B. als Anfangshypothese für die Extension von **ist wahr** die leere Menge, setzt man also I'(**ist wahr**) = $\emptyset$ (d.h. nimmt man an, keines der Sätze in D ist wahr), dann gibt $\tau_M$ an, daß unter dieser Hypothese die folgenden Sätze als wahr anzusehen sind:

$\tau_M(\emptyset)$ = { **Schnee ist weiß, 'Schnee ist weiß' ist nicht wahr, "Schnee ist weiß' ist wahr' ist nicht wahr,...., Alles ist nicht wahr, Alles, was Jones äußert, ist nicht wahr, ...** }

Machen wir uns klar, weshalb das der Fall ist. Für den Satz **Schnee ist weiß** liegt das T-Bikonditional

**'Schnee ist weiß' ist wahr** $=_{Df}$ **Schnee ist weiß**

vor. Unter den obigen Fakten M ist das Definiens wahr, da I(**weiß**) = {Schnee} und I(**Schnee**) = Schnee. Da im Definiens das Prädikat **ist wahr** nicht vorkommt, spielt es auch keine Rolle für die Wahrheit dieses Definiens, welche Hypothese wir bzgl. **ist wahr** getroffen haben. Die Auswertung des Satzes läuft wie üblich.

Für die folgenden Sätze in der obigen Aufzählung bemerkt man, daß jeder Satz der Form ⌜**s ist wahr** ⌝ falsch ist, da ja keiner der Sätze in der Extension des Prädikates **ist wahr** sein kann: Denn die Extension ist per hypothesin die leere Menge - und die enthält nichts. Damit ist aber nach der Wahrheitstafel für das **nicht** jeder Satz der Form ⌜ **s ist nicht wahr** ⌝ wahr.

Auf ähnliche Weise kann man für andere Teilmengen U des Gegenstandsbereichs ausrechnen, was $\tau_M$ ergibt. Ganz genau so wie im darstellenden Teil zu (RTD) vorgeführt, lassen sich jetzt auch für den Revisionsoperator $\tau_M$ mehrfache Anwendungen und ganz ähnlich mittels der Definition $\tau_M^0(U) = \tau_M(U)$, $\tau_M^{n+1}(U) = \tau_M(\tau_M^n(U)$ Revisionsfolgen betrachten.

Diese Betrachtung von mehrfacher Anwendung des Operators $\tau_M$ gestattet schließlich, einige Merkwürdigkeiten zu beheben. Für die Hypothese I'(**ist wahr)** = $\emptyset$ ergab sich z.B., daß der Satz **'Schnee ist weiß' ist nicht wahr** in der revidierten Extension von **ist wahr** lag und (somit) daß der Satz **'Schnee ist weiß' ist wahr** nicht in ihm lag, obwohl wir doch diesen Satz bejahen würden. Eine weitere Anwendung des Revisionsoperators hätte aber bereits diesen Satz in die Extension von **ist wahr** aufgenommen. Allgemein kann man zeigen, daß für jede beliebige Anfangshypothese U für die Extension von **ist wahr** eine natürliche Zahl $n > 1$ existiert, so daß der Satz **'Schnee ist weiß' ist wahr** $\in \tau_M^n(U)$. In

der bereits erläuterten Terminologie wäre dieser Satz gültig (kategorisch wahr): Wir könnten ihm ohne wenn und aber beipflichten.

Für dieses Beispiel gilt sogar noch ein bißchen mehr: Für alle Sätze der Sprache gibt es eine endliche Zahl an Revisionschritten, nach der sein semantischer Status sich auf einen Wert stabilisiert; entweder gilt, daß dieser Satz nach der endlichen Anzahl von Schritten immer wahr ist, d.h. für alle folgenden Anwendungen von $\tau_M$ in den ausgegebenen revidierten Extensionen für das Prädikat **ist wahr** liegt, oder es ist immer falsch. Das Beispiel ist durch besondere Verhältnisse ausgezeichnet, die so beschrieben werden können: Wiederholte Anwendungen von $\tau_M$ ergeben eine Menge, die irgendwann unverändert bleibt: man erhält einen Fixpunkt V von $\tau_M$; darüber hinaus gilt aber noch, daß mit welcher Hypothese auch man den Revisionsprozeß beginnt, man immer *denselben* Fixpunkt erhält. Gupta und Belnap nennen die Bedingungen, unter denen sich der Revisionsoperator $\tau_M$ so verhält, „Thomason conditions“.[133] Den Grund für die klaren Verhältnisse im Falle dieses Beispiels sehen Gupta und Belnap in der Abwesenheit einer fatalen Referenz („vicious reference“), wie sie z.B. bei der Einführung des Lügnersatzes vorliegen würde. In einer Fußnote merken Gupta und Belnap zum Begriff der fatalen Referenz an:

> Since no precise definition of „vicious reference“ is available, we cannot *prove* the claim that the Thomason conditions consist of all and only those in which there is no vicious reference in the language. Nonetheless, we can evaluate the claim by relying on our intuitive judgements concerning the presence or absence of vicious reference. As many conditions that are intuitively free of vicious reference turn out to be Thomason conditions (and conversely) we identify the two in the informal remarks below. But the reader should note that this is one place where the theory's descriptive adequacy can be questioned. If there are conditions that are intuitively free of vicious reference but which are not Thomason (or the other way around), then that counst against the theory. ([18] S.135, Fn 17)

Könnte man das eine oder das andere von den Gegeninstanzen beweisen, dann würde das nach Gupta und Belnap gegen die RTT sprechen.[134]

[133] So benannt, da eine Frage von Richmond Thomason die Autoren von RTT zur Beschäftigung mit den Modellen, in denen soche Bedingungen vorliegen, animiert hat.

[134] Die Verhältnisse mit dem Begriff der fatalen Referenz sind ähnlich wie mit dem Begriff der Berechenbarkeit. Auf der einen Seite stehen die Turingberechenbaren (oder rekursiven oder Abakus-berechenbaren oder ... Funktionen), die eine eindeutige technische Definition haben, auf der anderen Seite die Intuition darüber, wann eine Funktion auf mechanischen Wegen, auf der Grundlage eines Algorithmus berechenbar ist. Beweisen im strengen Sinne läßt sich die Churchsche These, die die Äquivalenz der technische Begriffe auf der einen Seite und des intuitiven Berechenbarkeitsbegriffs auf der anderen Seite behauptet, nicht, aber sie bestätigt sich aufs neue, wenn ein neuer technischer Begriff der Berechenbarkeit eingeführt und seine Äquivalenz mit den anderen bereits zur Verfügung stehenden Berechenbarkeitsbegriffen gezeigt wird.

Die hier ausgedrückte Intuition über fatale Referenz und den Wahrheitsbegriff wird nicht näher erläutert. Sie wird aber sozusagen gestützt durch eine empirische Basis.

Betrachten wir an einem Beispiel einer Sprache mit fataler Referenz, ob sich die Intuition bestätigen läßt.[135]

Nehmen wir an, die Situation M' ist genauso wie im obigen Beispiel M, mit der einzigen Ausnahme, daß Jones nur den einen folgenden Satz äußert:

**Etwas von dem, was Jones äußert, ist wahr.**

Intuitiv ist dieser Satz genau so problematisch wie der Satz 'Dieser Satz ist wahr'. Der zugehörige Revisionsoperator $\tau_{M'}$ liefert zwar für jede Anfangshypothese Revisionsfolgen, die sich nach bestimmten Revisionsstufen stabilisieren; allerdings sind diese Fixpunkte verschieden. $\tau_{M'}$ hat nämlich zwei Fixpunkte. Je nachdem, ob in den Initialhypothesen der Revisionsfolgen dieser Satz vorkommt oder nicht, ergeben sich Extensionen für $\tau_{M'}$, die diesen Satz enthalten oder nicht. Folglich liegen hier dank der fatalen Referenz (?) keine Thomason-Bedingungen vor. Im Modell M' wäre der Satz **Etwas von dem, was Jones äußert, ist wahr** pathologisch.

Auf andere Weise verletzt das folgende Beispiel die Thomason Bedingungen. Man betrachte hierzu ein Modell $M^*$, das sich von obigem M nur darin unterscheidet, daß das, was Jones sagt, lediglich der Satz ist:

**Etwas von dem, was Jones äußert, ist nicht wahr.**

Intuitiv ist dieser Satz genau so problematisch wie der Satz 'Dieser Satz ist nicht wahr'. Alle Revisionsfolgen für den Revisionsoperator $\tau_{M^*}$ liefern zwar unabhängig von der Anfangshypothese dasselbe Verhaltensmuster, das aber besteht nicht in einer Stabilisierung auf einen Fixpunkt, sondern in einer Oszillation: Enthält eine beliebige Anfangshypothese diesen Satz, dann ist dieser in der nächsten Revisionsstufe nicht in der Extension für das Wahrheitsprädikat. Nach einer weiteren Anwendung des Revisionsoperators dann wieder schon usw. Folglich liegen wieder – dank der fatalen Referenz (?) – keine Thomason-Bedingungen vor. Im Modell $M^*$ wäre der Satz **Etwas von dem, was Jones äußert, ist nicht wahr** in dem von Belnap und Gupta diesem Wort verliehenen Sinne paradox (in $S_0$).

Schließlich könnte man noch einen dritten Typ von Beispielen betrachten, der die Thomason-Bedingungen auf doppelte Weise verletzt: Weder ergibt sich bei solchen Beispielen unabhängig von den Anfangshypothesen dasselbe Verhaltensmuster für die Revisionsfolgen noch stabiliseren sich die Revisionsfolgen auf bestimmte Extensionen.

Nach Vorführung dieser Beispiele gelangen Belnap und Gupta zu folgendem Urteil:

[135] Man beachte, daß hier mit 'Sprache' ein Zweiertupel bestehend aus einer Menge von wohlgeformten Zeichenketten (syntaktisch konstruierte Sprache) und einer Interpretation/einem Modell für diese Menge von Zeichenketten gemeint ist.

> In summary: If the Tarski biconditionals are read as partial definitions, then (i) truth must be a circular concept, (ii) its signification is a rule of revision and is completely determined by the biconditionals (the Signification Thesis is therefore preserved), and (iii) much of the behavior of the concept of truth, both ordinary and pathological, can be explained. ([18], S.137)

Thesen (i) und (ii) werde ich im Unterabschnitt 'These von der Zirkularität der Wahrheit' ein wenig eingehender betrachten. These (iii) wird mit den folgenden Anmerkungen und dem Unterabschnitt 'Die Lügnerparadoxie erläutert in RTT' eine Erläuterung finden.

### 8.3.4 Systematische Einordnung von (RTW)

Gemäß Belnaps und Guptas Wahrheitstheorie sind die Anwendungsregeln für das natürlichsprachliche Wahrheitsprädikat nicht inkonsistent, wie es etwa Tarski behauptet hat. Die Paradoxa und andere paradoxähnliche Probleme mit dem Wahrheitsprädikat rühren daher, daß Wahrheit ein zirkulärer Begriff ist. Für zirkuläre Begriffe sind solche Verhaltensmuster typisch und können in der Revisionstheorie konsistent beschrieben werden. Auch wenn Belnap und Gupta nicht explizit von einer „property" sprechen, darf man annehmen, daß (RTW) gemäß hinter dem Wahrheitsprädikat eine Eigenschaft Wahr-zu-sein steckt und es sich bei ihm nicht etwa um eine bloße sprachliche Operation handelt.[136] Aus der (RTW) folgen also auch ontologische Thesen: Nämlich daß es eine besondere Sorte von Eigenschaften gibt, auf die man mit zirkulär definierten Prädikaten Bezug nehmen kann. Insofern muß man vorsichtig sein, wenn man zur systematischen Eingliederung von (RTW) das Prädikat 'ist deflationär' verwenden möchte. In einem gewissen Sinne ist (RTW) deflationär: Sie setzt im Prinzip nichts mehr über den Wahrheitsbegriff voraus als die in der Signifikationsthese zum Ausdruck kommende fundamentale Intuition. Bei anderen Wahrheitstheorien wird eine lange Liste von Kriterien zum Wahrheitsbegriff aufgestellt, an der die Theorie gemessen wird. Andererseits kommt die (RTW) nicht zum Ergebnis, daß das Prädikat 'ist wahr' eine bloße sprachliche Operation ist. Wahrheit ist (RTW) gemäß ein zirkulärer Begriff, dessen Rolle von keinem Begriff der bivalenten Sprache, in die das Wahrheitsprädikat eingeführt wird, übernommen wird. In diesem Sinne ist (RTW) nicht deflationär.

Zur weiteren systematischen Einordnung der Revisionstheorie von Belnap und Gupta kann man sich noch folgende Fragen stellen:

- Ist (RTW) eine epistemische Wahrheitstheorie?
- Ist Wahrheit eine Eigenschaft von Sätzen?
- Ist Wahrheit eine einfache Eigenschaft von Sätzen (Propositionen)?

[136] Es dürfte sich als schwierig erweisen, einerseits zu behaupten, das Wahrheitsprädikat drücke einen Begriff aus, andererseits aber sich der Frage gegenüber, ob es eine Eigenschaft Wahr-zu-sein gebe, indifferent zu zeigen oder diese gar zu verneinen. Wenn man aber eine starke Lesart von 'Eigenschaft' zugrunde legt, wäre auch das vielleicht möglich.

- Kann man auf endliche Weise angeben, was Satzwahrheit ist?

Eine Wahrheitstheorie soll genau dann epistemisch genannt werden, wenn sie Wahrheit in der Begrifflichkeit von Glauben oder Wissen expliziert. In der Tat verwenden Belnap und Gupta nirgends in RTT eines der beiden genannten epistemischen Begriffe. Auch ist in den partiellen Definitionen, die die (hypothetischen) Anwendungsbedingungen von 'ist wahr' festlegen, keines dieser Begriffe enthalten. Also ist (RTW) eine nicht-epistemische Wahrheitstheorie.

In der (RTW) ist Wahrheit eine Eigenschaft von Sätzen. Belnap und Gupta sehen sich damit aber nicht endgültig auf Sätze als die eigentlichen Wahrheitswertträger festgelegt. Diese Frage werde von der (RTW) nicht entschieden. Allerdings glauben sie, daß auch in dem Falle, daß sich Propositionen als die echten Wahrheitswertträger herausstellen sollten, die These aufrechterhalten werden kann, daß Wahrheit ein zirkulärer Begriff ist.

Wahrheit soll genau dann eine einfache Eigenschaft sein, wenn sie in einem ontologischen Sinne irreduzibel ist. Zu dieser Frage beziehen Gupta und Belnap keine Stellung. Eine andere Frage ist, ob der Begriff der Wahrheit analyseresistent ist. Mit der Theorie zirkulärer Definitionen und den zirkulären Begriffen gewinnt die Frage, ob Wahrheit analyseresistent ist, eine neue Dimension. Zunächst einmal ergeben sich aus der Tatsache Konsequenzen, daß in (RTW) unendlich viele partielle Definitionen die (hypothetischen) Anwendungsbedingungen festlegen. Wenn man meint, eine Analyse sei ein endlicher Satz bestimmter Form, und man außerdem glaubt, (RTW) würde eine Analyse des Wahrheitsbegriffs angeben wollen, dann würde die (RTW) das Projekt auf den ersten Blick verfehlen. Nun geben Belnap und Gupta auch an, daß man statt der partiellen Defininition von einer unendlichen Disjunktion ausgehen kann. Wiederum ist eine unendliche Disjunktion nach Voraussetzung keine endliche Zeichenkette, also läge keine Analyse vor. Der Überlegung, mit einer substitutionellen Quantifikation den Wahrheitsbegriff zu definieren, gehen Belnap und Gupta nicht nach.

Nehmen wir diesen Endlichkeitsaspekt heraus und fragen uns, ob (RTW) gemäß Wahrheit immer noch analyseresistent ist. Das übliche Verständnis einer Analyse schließt schon von vornherein aus, daß im Analysans ein Ausdruck verwendet wird, das mit dem Analysandum sinngleich ist. Dieses Zirkularitätsverbot wird natürlich durch die Definition des Wahrheitsprädikats in (RTW) verletzt. Die Wahrheitsdefinition gemäß (RTW) wäre schon dann keine Analyse, weil sie eines der logischen Regeln verletzt. Nun haben wir gelernt, daß das Zirkularitätsverbot nicht unbedingt immer gut motiviert zu sein braucht. Vielleicht läßt sich das, was mit einer Analyse bezweckt wird, auch über ein analysierendes Schema erreichen, welches das Zirkularitätsverbot verletzt? Was wird aber mit einer Analyse bezweckt? Ich habe nur die vage Idee davon, daß man sich mit einer Begriffsanalyse verspricht, bestimmte Beziehungen zwischen verschiedenen Begriffen zu beleuchten. Genauer versucht man sich ein Bild davon zu machen, wie der durch das Analysandum ausgedrückte Begriff mit anderen Begriffen zusammenhängt.

Nun wird in der klassischen Theorie der Analyse als Minimalbedingung für diese Erhellung folgendes gefordert: Analysans A und Analysandum B müssen sich ko-implizieren, d.h. alles was unter den Analysandumbegriff fällt, fällt auch unter den Analysansbegriff und umgekehrt; das läßt sich dann wiedergeben durch

$$\models A \leftrightarrow B$$

Diese Forderung ist für den klassischen Begriff eines Begriffs und des klassischen Begriffs der intensionalen Äquivalenz gleichwertig mit der Forderung, daß Analysans und Analysandum intensional äquivalent sein müssen. Die erste Forderung wird von manchen zirkulären Definitionen in (RTD) nicht erfüllt. Z.B. ist für die Definition

$$Gx =_{Df} \neg Gx$$

das zugehörige (allquantifizierte) Bikonditional

$$(\forall x)(Gx \leftrightarrow \neg Gx)$$

nicht gültig (kategorisch wahr). Wenn man aber die zur ersten Forderung äquivalente Forderung hinreichend schwach liest, dann gilt, daß Definiendum und Definiens in (RTD) intensional äquivalent sind gemäß des schwachen Standards – und das für jede Definition. D.h. ist G(x) das Definiendum einer Definition in (RTD) und A(G,x) das zugehörige Definiens, dann ist ein Satz G(a) in jedem Modell genau dann kategorisch wahr (bzw. kategorisch falsch bzw. paradox bzw. pathologisch aber nicht paradox), wenn A(G,a) kategorisch wahr (bzw. kategorisch falsch bzw. paradox bzw. pathologisch aber nicht paradox) ist. Wir sind dann also genötigt, die erste Forderung zu verwerfen und stattdessen die hierzu klassisch gesehen äquivalente zweite Forderung mit dem Ausdruck 'intensional äquivalent' im schwachen von (RTD) vorgezeichneten Sinn zu lesen, wenn wir sagen wollen, daß zirkuläre Definitionen aus (RTD) die Minimalbedingung für eine Begriffsanalyse erfüllen. In diesem Sinne scheiden zirkuläre Definitionen aus (RTD) nicht von vornherein als Kandidaten für eine Begriffsanalyse aus. Und wenn also nicht mehr für eine Analyse nötig ist als eine erhellende Anwort auf die Frage, wie der Begriff der Wahrheit mit den anderen Begriffen zusammenhängt, dann ist der Begriff der Wahrheit (RTW) gemäß analysierbar. Ob zirkuläre Definitionen in (RTD) aber tatsächlich zu einer Analyse in einem anspruchsvolleren Sinn taugen, das hängt von der hinreichenden Bedingung für anspruchsvolle Analysen ab – und wie die genau aussieht, dafür scheint es noch keine vernünftige Antwort zu geben. Daß es nicht Sinnidentität zwischen Analysans und Analysandum sein kann, belegt die berühmte Aporie der Analyse von Langford.[137] Diese läßt sich einleuchtend

[137][38], S.323

lösen, wenn man Sinnidentität aufgibt.[138] Die notwendige Bedingung auch für eine anspruchsvollere Analyse erfüllen prinzipiell die zirkulären Definition, wenn man denn zirkuläre Begriffe akzeptiert und 'intensional äquivalent' schwach liest. Eine Analyse des Wahrheisbegriffs mit einem Analysans, in das nur nichtzirkuläre Begriffe eingehen, gibt es nach (RTW) nicht.

Da die Signifikation des Wahrheitsprädikats in einem Modell als eine Revisionsregel $\tau_M$ angegeben wird, diese aber unter anderem durch die unendliche Zahl an partiellen Definitionen festgelegt wird, läßt sich in diesem Sinne nicht auf endliche Weise angeben, was Satzwahrheit ist. In einem anderen Sinne ließe sich aber auch (RTW) gemäß auf endliche Weise angeben, was Satzwahrheit ist. Wenn man nämlich die unendliche Disjunktion der partiellen Definitionen betrachtet und den Begriff der substitutionellen Quantifikation zur Verfügung hat, dann kann man in einem endlichen Satz sagen, was Satzwahrheit ist.

## 8.4 Die Lügnerparadoxie erläutert in RTT

Belnaps und Guptas Erläuterung der Lügnerparadoxie läuft nicht darauf hinaus, das durch die Lügnerparadoxie gestellte normative Problem zu lösen, nämlich ein künstliches Wahrheitsprädikat in einer künstlichen, hinreichend ausdrucksschwachen Sprache zu konstruieren, für die sich keine Widersprüchlichkeiten ergeben. Statt dessen versuchen sie das durch die Lügnerparadoxie gestellte deskriptive Problem zu lösen, d.h. eine Erklärung des Wahrheitsprädikats und der Paradoxie für die natürliche Sprache Englisch (Deutsch) zu geben. Indem sie zulassen, daß das Wahrheitsprädikat für L in L selbst vorkommen darf und daß L gewisse universelle Züge trägt, werden sie dieser Aufgabe zu einem gewissen Grade gerecht. Belnap und Gupta halten an dem Faktum fest, daß sich mit dem natürlichen Wahrheitsprädikat bestimmte Probleme wie die Lügnerparadoxie ergeben. Das sehen sie aber nicht als einen hinreichend Grund dafür an, die These zu verwerfen, das natürlichsprachliche Wahrheitsprädikat drücke einen Begriff aus, für das es konsistente Anwendungsregeln gibt. Wenn man wie Belnap und Gupta annimmt, das Wahrheitsprädikat drücke einen zirkulären Begriff aus, dann kann man an der These festhalten, das Wahrheitsprädikat der natürlichen Sprachen sei konsistent und drücke einen Begriff aus, ohne damit selbst in Widersprüche zu geraten.

Die fatale Intuition, an dem Zitattilgungsprinzip (ZTP) festzuhalten, wird auch von Belnap und Gupta als verfehlte Intuition angesehen. Allerdings versuchen sie möglichst viel von dieser fundamentalen Intuition herrüberzuretten, indem sie nicht sämtliche echten Bikonditionale der Form

[138] Zur Auflösung der Aporie siehe z.B. [27], S.33-38. Was die (RTD) Interessantes zur Aporie der Analyse und der hinreichenden Bedingung für Analysen sagen könnte, behandelt Orilia in einem kurzen Abschnitt seines kürzlich erschienen Aufsatzes [36]. Ich glaube aber, daß nicht mehr heraus kommt als etwas, was wir schon wissen dürften, nämlich daß für eine Begriffsanalyse nicht Sinnidentität von Analysans und Analysandum gefordert werden darf.

$s_i$ ist wahr $\leftrightarrow A_i$

akzeptieren – was einer Akzeptanz von (ZTP) gleichkäme – sondern sämtliche als partielle Definitionen verstandenen „Bikonditionale“

$s_i$ ist wahr $=_{Df} A_i$

Mit ihrer Revisionstheorie bieten sie eine Möglichkeit, wie man solche partiellen Definitionen, unter denen ja auch echt zirkuläre sind, behandeln und auffassen kann. Für die so verstandenen T-Bikonditionale läßt sich aber kein Widerspruch mehr ableiten: Aus der partiellen Definition

‘Dieser Satz ist nicht wahr’ ist wahr $=_{Df}$ Dieser Satz ist nicht wahr

läßt sich im revisionstheoretischen Kalkül $C_0$ nicht mehr aus der Annahme, daß der Satz ‘Dieser Satz ist nicht wahr’ wahr ist, ableiten, daß der Satz ‘Dieser Satz ist wahr’ dann nicht wahr ist. Der Grund ist, daß hier eine Änderung des Index stattfindet, die durch die beiden modifizierten Regeln ($DFE_r$) und ($DFB_r$) bedingt ist.

Hierdurch wird aber das pathologische Verhalten, welches das Wahrheitsprädikat in der Paradoxie an den Tag legt, nicht wegerklärt – es wird lediglich ausgeschlossen, daß die Erklärung, die man gibt, selbst wieder Paradoxien provoziert. Das paradoxe Verhalten findet sich in der Revisionssemantik abgebildet; ein Satz wie ‘Dieser Satz ist nicht wahr’ ist paradox in dem in der Revisionssemantik zugrunde gelegten Sinne: Für keine Anfangshypothese X für die Extension des Wahrheitsprädikats gibt es eine natürliche Zahl $p$, so daß der Satz ‘Dieser Satz ist nicht wahr’ für alle $n$ mit $p \leq n$ in den revidierten Extensionen $\tau^n_M(X)$ wäre. Der Satz ‘Dieser Satz ist nicht wahr’ oszilliert in sämtlichen Revisionsfolgen.

Als ein besonderes Verdienst ihrer Erklärung der Lügnerparadoxie sehen Belnap und Gupta die Tatsache an, daß in ihr die unterliegende Logik, in deren Rahmen das Wahrheitsprädikat eingeführt wird, nicht geändert wird: Es werden keine neue Wahrheitswerte eingeführt oder die logischen Konnektoren uminterpretiert, um eine Erklärung der Lügnerparadoxie zu erwirken.

## 8.5 These von der Zirkularität der Wahrheit

Die Motivation im Abschnitt 5.2 ‘Wahrheit ein erster Vergleich’ ließ bereits Guptas und Belnaps These erahnen, daß der Wahrheitsbegriff der natürlichen Sprache ein zirkulärer Begriff ist. Ein in sich geschlossenes, übersichtliches Argument für diese These findet sich nicht in der RTT. Lediglich in [19], S.634-635, findet man ein kurzes Argument.[139] Dort ist es in etwa folgendermaßen formuliert:

[139]Es dient dort allein dem Zweck, ihren Rezensenten Koons, der in [23] behauptet, Belnap und Gupta könnten nicht die Zirkularität des Wahrheitsbegriffs nachweisen, vom Gegenteil zu überzeugen. Koons akzeptiert die zweite These des Arguments, nicht aber die erste. Allerdings zeugt seine Begründung von einem Mißverständnis oder einem anderen Verständnis des Ausdrucks ‘intensional adäquat’. Leider gibt er auch nach der Korrespondenz mit Belnap und Gupta nicht an, wie er ‘intensional adäquat’ lesen möchte.

1. Enthält das Definiens ‘A(x,G)’ einer Definition ‘$x$ ist $G =_{Df} A(x, G)$’ den Ausdruck „G“ wesentlich und ist es intensional adäquat bzgl. des Definiendums ‘x ist G’, dann ist G ein zirkulärer Begriff.

2. Man kann in einer wesentlich zirkulären, stipulativen Definition einen einstelligen Begriff T definieren, für welchen gilt: T und der logische, schwache, absolute Begriff der Wahrheit sind intensional äquivalent.

3. Also: Der logische, schwache, absolute Begriff der Wahrheit ist zirkulär.

Für die Beurteilung des Arguments ist Guptas und Belnaps äquivalente Formulierung der ersten Prämisse relevant. Der Satz in (1) ist mit folgendem Satz äquivalent:[140]

(1’) Wenn G ein zum Begriff H intensional äquivalenter Begriff ist, wobei H über eine essentiell zirkuläre stipulative Definition definiert wird, dann ist G ein zirkulärer Begriff.

Hiermit läßt sich einsehen, daß die zweite Prämisse mehr oder weniger eine Instanz des Antezedens von (1’) darstellt, während die Konklusion eine Instanz des Konsequens ist. Die Schlüssigkeit des Arguments ließe sich also durch Hinzunahme der von Gupta und Belnap behaupteten Äquivalenz im wesentlichen mit der Anwendung der Beseitigungsregeln für den Allquantor und den Pfeil beweisen.[141]

In Satz (1’) und (2) ist von „*der* intensionalen Äquivalenz “ die Rede. Für die Revisionstheorie ist diese Rede problematisch, da – wie wir von Gupta und Belnap erfahren haben – im Rahmen der Revisionstheorie nicht von *der einen* intensionalen Äquivalenz gesprochen werden kann; zwei zirkulär definierte Prädikate F und G mögen nach dem einen Standard intensional äquivalent sein, nach dem anderen aber nicht. An einem Ende steht der schwächste Standard für die intensionale Äquivalenz, wonach zwei zirkulär definierte Prädikate F und G (auch dann) intensional äquivalent sind, wenn in jeder möglichen Welt ein G enthaltender Satz genau dann kategorisch (bzw. paradox bzw. pathologisch aber nicht paradox) ist, wenn der entsprechende F enthaltende Satz kategorisch (bzw. paradox bzw. pathologisch aber nicht paradox) ist. Am anderen Ende steht der stärkste Standard, wonach zwei zirkulär definierte Prädikate genau dann intensional äquivalent sind, wenn sie in jeder möglichen Welt dieselbe Revisionsregel haben. Es ist aus dem Kontext nicht ersichtlich, ob Belnap und Gupta die Ausdrücke ‘intensional äquivalent’ und ‘intensional adäquat’ in (1) bzw. (1’) vor dem Hintergrund ihrer (RTD) gelesen haben wollen oder ein intuitives Verständnis voraussetzen. Was zu einer Entscheidungshilfe bzgl. der Alternativen beitragen

[140] Man erinnere sich daran, daß für die Zirkularität eines Begriffs hinreichend ist, daß es in mindestens einem Modell pathologische Fälle für den Begriff gibt. Vgl. Kap.6

[141] Man beachte, daß hier über Prädikate quantifiziert wird. Eine Ableitung des zugehörigen Argumentschemas hätte daher in einer höherstufigen Logik zu erfolgen.

könnte, ist die Tatsache, daß Belnap und Gupta sich vermutlich nicht den Vorwurf der Äquivokation zu Schulden kommen lassen wollen; an allen Stellen im Argument wird folglich 'intensional äquivalent' in derselben Bedeutung zu lesen sein.

Definitionen werden für bestimmte Sprachen L aufgestellt. Die in den Definitionen definierten Prädikate gehören zur Sprache L. Im obigen Argument ist aber immer nur die Rede von den Begriffen G und T bzw. den Prädikaten 'ist G' und 'ist T'.[142] Tatsächlich „erklärt" (RTW) – wie auch Tarskis Wahrheitskonzeption – immer „nur" ein Prädikat T in L oder 'wahr in L': Die partiellen Definitionen haben auf der linken Seite vom Definitionszeichen '$=_{Df}$' immer nur Standardbezeichnungen für Sätze einer Sprache L.[143] Vergegenwärtigen wir uns noch einmal, wie diese Sprachen L, für die das Wahrheitsprädikat erklärt wird, beschaffen sind: Wir haben uns vorzustellen, daß die Sprache L in einer Sprachgemeinschaft einer möglichen Welt w verwendet wird. Die Ausdrücke der Sprache L haben alle Bedeutungen, doch ist L eine sehr idealisierte Sprache. Die Syntax dieser Sprachen kann mit Mitteln der klassischen Quantorenlogik dargestellt werden. Die in ihnen enthaltenen logischen Konstanten gehorchen den Regeln der klassischen Quantorenlogik. Die Extensionen der nichtlogischen Konstanten können in einer klassischen interpretierten Sprache[144] wiedergegeben werden. In diesen Sprachen soll es neben einem ausgezeichneten Prädikat, welches den Wahrheitsbegriff ausdrücken soll, weiterhin Standardbezeichnungen für sämtliche Sätze aus L geben, so daß man also in L über die Sätze von L reden kann. Für solche Sprachen L erklären Gupta und Belnap das Wahrheitsprädikat und auf solche Sprachen kann das Argument angewandt werden. Nun ist das Deutsche (bzw. Englische) aber keine Sprache, die sämtliche oben aufgelisteten Eigenschaften hat, aber gerade für den Wahrheitsbegriff dieser Sprache wollen Gupta und Belnap die Zirkularität nachweisen. Wie kommen sie dazu? Hierzu sagen Gupta und Belnap nichts, aber man kann sich leicht vorstellen, wie sie zu ihrer Behauptung gelangen: Das Deutsche enthält als Fragmente bestimmte Sprachen L, die den oben aufgestellten Bedingungen gehorchen. Belnaps und Guptas Intuition scheint zu sein: Wenn überhaupt etwas zum logischen, schwachen, absoluten Wahrheitsbegriff des Deutschen gesagt werden kann, dann ist es dieses, daß der semantische Status von Sätzen eines geeigneten Teilfragments L, die das logische schwache absolute Wahrheitsprädikat für L – T in L – enthalten, derselbe ist wie der von den entsprechenden Sätzen des Deutschen, die statt T in L das Wahrheitsprädikat des Deutschen enthalten. Da T in L aber einen zirkulären Begriff ausdrückt,

[142] Ich werde im folgenden von den Prädikaten 'G', 'H' und 'T' reden und nicht von den Begriffen G, H und T.

[143] Das ist natürlich auch dann noch wahr, wenn wie im obigen „Jones-Beispiel" das Anhängsel 'in L' einfach weggelassen wird.

[144] Zur Erinnerung: Eine interpretierte Sprache $\mathcal{L}$ ist ein Paar $(L, M)$, deren erste Komponente $L$ eine syntaktisch konstruierte Sprache und deren zweite Komponente $M$ ein quantorenlogisches Modell für diese Sprache (eine Interpretation der Sprache) ist.

ist somit der Wahrheitsbegriff des Deutschen ebenfalls zirkulär. Nennen wir diese Intuition Guptas und Belnaps die „These von der konservativen Erweiterung der natürlichen Sprachen“. Somit erhalten wir mit den oben angedeuteten Ergänzungen folgende Ableitung:

1. Für alle Sprachen L mit obigen Eigenschaften, für alle Prädikate G und alle Prädikate H aus L gilt: Wenn H und G intensional äquivalent sind und H über eine wesentlich zirkuläre stipulative Definition in L definiert wird, dann drückt G einen zirkulären Begriff aus.

2. Für jede Sprache L mit obigen Eigenschaften gilt: Man kann in einer wesentlich zirkulären, stipulativen Definition ein einstelliges Prädikat T in L definieren, für welches gilt: T in L und das Prädikat ‘ist wahr in L’, welches den logischen, schwachen, absoluten Begriff der Wahrheit für L ausdrückt, sind intensional äquivalent.

3. Also: Für alle Sprache L mit obigen Eigenschaften gilt: Der logische, schwache, absolute Begriff der Wahrheit für L ist zirkulär.

4. „These von der konservativen Erweiterung der natürlichen Sprachen“.

5. Also: Der logische, schwache, absolute Wahrheitsbegriff der deutschen Sprache ist zirkulär.

In diesem Argument haben wir es mit zwei problematischen Prämissen zu tun, einer allgemein formulierten Aussage in 1. und einer Aussage in 2., die speziell den Wahrheitsbegriff betrifft; die dritte, von Belnap und Gupta nicht explizit erwähnte Prämisse in 4., halte ich für plausibel. Beginnen wir daher mit der Prämisse in 4. Es ist eine Tatsache, daß Gupta und Belnap keine Wahrheitstheorie für die natürliche Sprache Deutsch oder Englisch entwickeln. Das Englische und das Deutsche sind höchst komplexe Sprachen, die sich mit der Zeit ändern. Selbst wenn man sich darauf verständigen würde, eine Wahrheitstheorie zu entwickeln, die sich nur auf das Deutsche zu dem und dem Zeitpunkt oder zusätzlich sogar nur auf den einen Idiolekt, praktiziert von der und der Person zu dem und dem Zeitpunkt, bezieht, ist nicht sichergestellt, ob es möglich ist, die Revisionstheorie hierauf anzuwenden.[145] Die 4. Prämisse besagt aber nicht, daß sich die Revisi-

[145]Einen Einwand hiergegen macht Vann McGee in [34], S. 401-402: In der Form, wie die Revisionstheorie vorgestellt wurde, gingen Belnap und Gupta immer davon aus, daß der Gegenstandsbereich des zugrunde liegenden Basismodells für die Objektsprache eine Menge ist. Mit dieser Einschränkung können Belnap und Gupta aber z.B. nicht die Wahrheitsbedingungen eines Antwortsatzes auf die folgende Frage geben: „Gibt es eine nichtabzählbare Kardinalzahl, deren Mächtigkeit kleiner als die der reellen Zahlengerade ist?“ Gupta gibt aber in einer persönlichen Korrespondenz mit Vann McGee eine plausible Modifikation der Revisionstheorie an, mit der derartige Probleme umgangen und also auch solche Objektsprachen, deren Gegenstandsbereiche keine Mengen sind, in dieser Hinsicht zumindest revisionstheoretisch behandelt werden können. Die revisionstheoretische Konzeption wird durch diese Modifikation nicht zerstört. ([34], S.404-405)

onstheorie auf das Deutsche oder das Englische anwenden läßt, sondern lediglich, daß die „Einschränkung“ des natürlichsprachlichen Wahrheitsbegriffs auf geeignete Fragmente durch die Revisionstheorie der Wahrheit beschrieben werden kann. Solange man nicht nachweisen kann, daß diese Einschränkung eine Verzerrung des natürlichsprachlichen Wahrheitsbegriffs mit sich bringt, sollte man an dieser intuitiv plausiblen Prämisse festhalten dürfen.

Was ist an der ersten Prämisse problematisch? Wenn man die These akzeptiert, daß es zirkuläre Begriffe in dem Sinne gibt, wie es in der Revisionstheorie erläutert wird, dann dürfte es keine Probleme geben, die erste Prämisse für korrekt zu erklären. Da spielt es dann auch keine Rolle, ob man die Ausdrücke ‘intensional äquivalent’ und ‘intensional adäquat’ gemäß des schwachen oder starken Standards liest.[146]

Hat man ein schlagendes ontologisches Argument dafür, daß es keine Eigenschaft und keinen Begriff zu zirkulär definierten Prädikaten geben kann, daß also solche Prädikate keine „echten“ Prädikate sind, dann wird man auch bestreiten wollen, daß man in vernünftiger Weise von der Intension zirkulär definierter Prädikate sprechen kann. Folglich würde es auch keinen Sinn mehr machen, von der intensionalen Äquivalenz zwischen einem Prädikat G und einem zirkulär definierten Prädikat H zu sprechen. Die erste Prämisse würde also dann mit dem Hinweis verworfen werden, man wende auf einen Gegenstand (das zirkulär definierte Prädikat) eine Funktion an (die Funktion, welche einem Ausdruck (Prädikat) ihre Intension zuordnet), die diesen Gegenstand nicht in ihrem Definitionsbereich hat. Genauso fälschlich wäre es, auf die Zahl 3 die Funktion anzuwenden, welche durch den Ausdruck ‘Die Farbe von ()’ ausgedrückt wird, und dann Aussagen über die Farbe der Zahl 3 zu treffen. Ich kenne aber kein derartiges ontologisches Argument, welches Kriterien dafür aufstellt, was eine Eigenschaft und was ein Begriff ist, die implizieren, daß es keine zirkulären Begriffe und die ihnen entsprechenden Eigenschaften geben kann.

Betrachten wir die zweite Prämisse. Wodurch wird diese Prämisse gestützt? Für das Prädikat T in der extensionalen Sprache L wurden die T-Bikonditionale als partielle Definitionen aufgestellt. In diesen Bikonditionalen kommt das metasprachliche Zeichen ‘$=_{Df}$’ vor. Können wir hier irgendwelche sprachlichen oder sonstigen Intuitionen mobilisieren, die uns zum Urteil führen könnten, sämtliche T-Bikonditionale für eine Sprache L würden den Wahrheitsbegriff für L definieren? Ich habe keine derlei sprachliche oder sonstige Intuitionen. Unsere Intuitionen reichen doch nur soweit, daß wir sagen möchten, daß einige Sätze der deutschen Sprache, die die Form haben:

[146] Wenn Belnap und Gupta auch np-zirkuläre Begriffe akzeptieren, dann könnte es sein, daß sich mit der schwächsten Lesart Probleme für die erste Prämisse ergeben. Belnap und Gupta wollen aber eh nachweisen, daß der Begriff der Wahrheit zirkulär und pathologisch – und nicht etwa np-zirkulär – ist. Daher ist die np-Zirkularität hier Fehl am Platze.

‘p’ ist wahr genau dann, wenn p[147]

zu akzeptieren sind. Den Satz “Schnee ist weiß’ ist wahr genau dann, wenn Schnee weiß ist’ z.B. werden wir akzeptieren. Die aus dem Lügnerparadoxon gezogene Lehre wird uns daran hindern, für alle Sätze der Sprache L, in der auch der Lügnersatz formuliert werden kann, die entsprechenden Bikonditionale anzunehmen. Ottonormalverbraucher wird eh keinen Gedanken daran verschwendet haben, ob nun sämtliche Bikonditionale wahr sind oder nicht. Wenn er eine Intuition äußert, dann bezieht die sich auf die Menge aller Sätze, mit welchen er normalerweise konfrontiert wird – und unter denen kommt kein Satz wie ‘Dieser Satz ist falsch’ vor.[148] Nun zwingt uns auch die Revisionstheorie nicht, sämtliche Sätze der Form:

‘p’ ist wahr genau dann, wenn p

für korrekt zu erklären. Stattdessen soll hingenommen werden, daß sämtliche metasprachlichen Sätze der Form

‘p’ ist wahr $=_{Df}$ p

den Wahrheitsbegriff definieren. Wir sollen also alle T-Bikonditionale mit ‘genau dann wenn’ verstanden als definitorisches $=_{Df}$ für korrekt erklären. Hat Ottonormalverbraucher bestimmte Intuitionen, alle T-Bikonditionale zu akzeptieren, wobei er ‘genau dann wenn’ jetzt definitorisch versteht? Ich glaube nicht. Es wäre auch nicht im Sinne von Belnap und Gupta, die das Zeichen ‘$=_{Df}$’ nicht als ein Teil der Umgangssprache Deutsch, sondern als Komponente einer Metasprache, in der wir über das Deutsche sprechen, verstanden wissen wollen. Welches Recht haben wir zu sagen, daß alle als partielle Definitionen verstandenen T-Bikonditionale tatsächlich den üblichen Wahrheitsbegriff definieren würden? Steckt hier auch eine Intuition dahinter? In der Aufgabenstellung für (RTW) formulieren Gupta und Belnap die Signifikationsthese als eine „fundamentale Intuition“([18], S.25). Ich habe keine dahingehenden Intuitionen, die Signifikationsthese als korrekt zu akzeptieren. Die Ausdrücke ‘Intension’ bzw. ‘Signifikation’ und ‘T-Bikonditional’ so gelesen, wie es mir bekannt ist, führen sogar dazu, daß ich Signifikationsthese verwerfen würde. Ich habe auch keine „Schema-Intuition“, die die Signifikationsthese, welche ja eigentlich ein Schema mit auszufüllenden Wortkapseln ‘Signifikation’ und ‘T-Bikonditional’ ist, für „korrekt“ erklären würde.[149] Nun bieten Belnap und Gupta eine Lesart der Ausdrücke ‘T-Bikonditional’ und ‘Signifikation’ an, mit denen die Signifikationsthese ihrer Meinung nach wahr würden. Die

[147]Hier ist der Buchstabe ‘p’ wieder Platzhalter für einen Satz der deutschen Sprache.

[148]Es gibt aber Sätze, die ganz gewöhnlich ausschauen und dennoch paradox sind, weil besondere Verhältnisse vorliegen, z.B. kontingent paradoxe Sätze. Auch für solche Sätze wird Ottonormalverbraucher möglicherweise die T-Bikonditionale akzeptieren, ohne zu wissen, daß er damit in Widersprüche gerät.

[149]Vielleicht würden hier Belnap und Gupta erwidern, meine Intuition, an (ZTP) zu glauben, beziehe sich eigentlich auf die definitorisch verstandenen T-Bikonditionale.(?)

zweite Prämisse, so müssen wir annehmen, sehen Belnap und Gupta durch die Signifikationsthese in ihrer Lesart gestützt. Das Prädikat T in L wird definiert über partielle Definitionen der Form

$$\text{T}(\text{`p'}) =_{Df} \text{p}$$

worunter sich partielle Definitionen finden, die wesentlich zirkulär sind. Was rechtfertigt nun die Aussage, das Prädikat 'ist wahr in L', welches den Wahrheitsbegriff für die Sprache L ausdrückt, und T seien intensional äquivalent? Im strengen Sinne beweisen läßt sich die Aussage nicht. Anhand von Beispielen, wie dem obigen Jones-Beispiel – wo übrigens statt 'T' das Prädikat '**ist wahr**' verwandt wurde – können wir lediglich feststellen, daß die intuitiv problematischen Fälle mit dem Wahrheitsbegriff für L durch T wiedergespiegelt werden: Im obigen Modell $M'$ wird sozusagen eine mögliche Welt repräsentiert, in der ein gewisser Jones lediglich den Satz äußert: **Etwas von dem, was Jones äußert, ist wahr.** Würde das in diesem Satz vorkommende Prädikat '**ist wahr**' tatsächlich den Wahrheitsbegriff für L ausdrücken, dann müßte er die Problematik wiederspiegeln, die man mit einem solchen Satz hätte: Man wird ihm weder beipflichten, noch ihn verneinen wollen. Gleichzeitig aber wird man diesen Satz nicht als widersprüchlich empfinden; wir können in konsistenter Weise das eine oder andere annehmen. Wie wir gesehen haben, ergibt sich tatsächlich für diesen Satz auch, daß er in der Revisionstheorie in dem erläuterten Sinne als pathologisch ausgegeben wird. Ähnlich läßt sich die Intuition bestätigen, daß wenn Jones nur den Satz geäußert hätte: **Etwas von dem, was Jones äußert, ist nicht wahr**, ein paradoxer Satz vorliegen würde. Alle anderen Sätze aus L, die mit der Interpretation, welche sie in den Modellen erhalten, intuitiv nicht problematisch erscheinen, stellen sich auch als kategorisch heraus. Die Hoffnung ist, daß immer dann, wenn sich etwas intuitiv Problematisches in einem Modell ergibt – was Belnap und Gupta unter dem Titel 'fatale Referenz' führen – dieses auch durch den Revisionsoperator $\tau_M$ angezeigt wird. Die am Jones-Beispiel gemachte Beobachtung könnte uns folglich darin rechtfertigen zu sagen, daß für alle Sätze a der semantische Status von 'Ta' (bzw. **a ist wahr**) und 'a ist wahr in L' in folgendem Sinne derselbe ist: Ein Satz 'a ist wahr' ist genau dann intuitiv paradox, wenn 'Ta' paradox in dem technischen Sinne aus (RTD) ist. Ein Satz 'a ist wahr' ist genau dann intuitiv problematisch, aber nicht paradox, wenn 'Ta' nicht kategorisch und nicht paradox in dem technischen Sinne aus (RTD) ist. Ein Satz 'a ist wahr' ist genau dann intuitiv problematisch, wenn 'Ta' nicht kategorisch in dem technischen Sinn aus (RTD) ist. Was zwingt uns nun dazu, wie Belnap und Gupta zu behaupten, die Signifikation des Prädikats 'ist wahr in L' und T sei für jede mögliche Welt dieselbe, d.h. die Signifikation des Prädikats 'ist wahr in L' sei in jeder möglichen Welt nichts anderes als eine Revisionsregel – und damit im obigen Argument zu folgern, daß der logische, schwache Begriff der Wahrheit zirkulär ist? Eine Notwendigkeit kann ich hier nicht erkennen, es zeigt aber eine Möglichkeit, wie man das von der Lügnerparadoxie gesetzte deskriptive Problem lösen kann, ohne

daß man die These aufgibt, daß es eine Eigenschaft und einen Begriff Wahrheit gibt und daß das natürlichsprachliche Wahrheitsprädikat konsistenten Regeln gehorcht. Solange man nicht die Unmöglichkeit anderer Theorien der Wahrheit, die diese Desiderate erfüllen, aufzeigen kann, ist nicht im strengen Sinne bewiesen, daß Wahrheit ein zirkulärer Begriff ist. Der Fixpunkttheoretiker wird behaupten, er könne eine Theorie der Wahrheit geben, die ebenfalls die Desiderate erfüllt. Belnap und Gupta versuchen zu zeigen, daß das nicht der Fall ist. Und wenn jemand eine Theorie aufstellen würde, die den Desideraten genügt, dann müßte das Für und Wider auf einer anderen Ebene geschehen. Das ist ähnlich wie bei den Theorien zur Ontologie der Eigenschaften: Viele der Theorien vermögen die Rollen, die den Eigenschaften de facto zukommen, gleichermaßen gut zu erklären. Nichtsdestotrotz ist der einen Theorie gemäß eine Eigenschaft eine Menge von Einzeldingen, der anderen Theorie gemäß aber eine Universale. Die Abwegung für oder wider die eine oder andere Theorie kann dann nicht dadurch geschehen, daß man sich anschaut, welche der beiden besser erklärt, daß Eigenschaften die und die Rollen haben; denn nach Voraussetzung sollen beide die Rolle von Eigenschaften gleichermaßen gut erklären.

So einfach, wie es der erste Blick auf das obige kurze Argument vermuten lassen mag, ist die Zirkularität des Wahrheitsbegriffs nicht nachzuweisen. Da gehen metaphysische Annahmen (in Form der Signifikationsthese) und auch erklärungstheoretische Überlegungen mit ein: Mit der Annahme, daß der Begriff der Wahrheit zirkulär ist, kann man die Lügnerparadoxie gut erklären. Ein von diesen Überlegungen unabhängiges Argument für die These, der Wahrheitsbegriff sei zirkulär, gibt es nicht.

Sämtliche Einwände, die gegen die (RTW) gemacht worden sind, zielen im Prinzip darauf ab, nachzuweisen, daß das zirkuläre Prädikat T nicht die Rolle spielen kann, welche das natürlichsprachliche Wahrheitsprädikat hat. Ein ontologisches Argument gegen zirkuläre Begriffe gibt es nicht.

# 9 Andere zirkuläre Begriffe

Die Frage, welcher der natürlichsprachlichen Begriffe außer dem der Wahrheit zirkulär sind, beantworten Gupta und Belnap mit im Vergleich zum Wahrheitsbegriff knapp ausfallenden Ausführungen.[150] Eine Ausnahme bildet die Behandlung des Notwendigkeitsbegriffs.[151] Die Bandbreite der Prädikate, die Gupta und Belnap für zirkulär halten, ist recht erstaunlich: Neben semantischen Begriffen (Referenz, Erfüllung, etc.), mengentheoretischen und beweistheoretischen Begriffen (Elementbeziehung, Exemplifikation) finden sich auch modale und doxastische Begriffe (Notwendigkeit, Glauben, Wissen); auf eher spekulativer Basis argumentieren Gupta und Belnap unter Berufung auf Strawsons Ideen dafür, daß auch die Begriffe „ist derselbe Körper wie“ und „ist derselbe Ort wie“ zirkulär sind – was insofern bemerkenswert ist, als diese Begriffe im Unterschied zu den übrigen zirkulären Begriffen mit konkreten raumzeitlichen Gegenständen zu tun hat.

Ich werde mich im folgenden mit kurzen darstellenden Anmerkungen begnügen und nur für den Referenzbegriff etwas ausführlicher darstellen, wie dessen Zirkularität gefolgert wird.

Die Zirkularität des Notwendigkeitsbegriffs könnte man als eine Folge aus der Zirkularität des Wahrheitsbegriffs ansehen: Daß ein Satz notwendigerweise gilt, heißt doch nichts anderes als, daß es wahr in allen möglichen Welten ist. Belnap und Gupta geben aber auch eine andere Lesart von ‘ist notwendigerweise wahr’ an, die auf eine andere Zirkularität als jene, die durch den Wahrheitsbegriff bedingt ist, zu laufen scheint.[152] Die Erklärung modaler Lügnerparadoxien läuft dann wie für die gewöhnliche Lügnerparadoxie auf die Aussage hinaus, daß der Notwendigkeitsbegriff zirkulär ist.

In der naiven Mengenlehre gibt es Paradoxa, die eine verblüffende Ähnlichkeit mit der Lügnerparadoxie aufweisen. Man denke z.B. an die berühmte Russellsche Antinomie: Wir bilden die folgende Menge $R = \{x : x \notin x\}$, die aus Mengen x besteht, welche sich selbst nicht als Element enthalten. Enthält sich nun $R$ selbst als Element oder nicht? Wenn $R \in R$, dann erfüllt $R$ die Eigenschaft, die die Element aus $R$ charakterisiert, nämlich sich selbst nicht zu enthalten, also $R \notin R$. Dann müssen wir annehmen, daß also doch $R \notin R$. Damit ist $R$ aber eines der x, die in $R$ enthalten sind, folglich $R \in R$!? Belnap und Gupta führen die Problematik darauf zurück, daß die Elementbeziehung zirkuläre Definitionen hat. Eine genaue Ausarbeitung dieser Überlegung und eine revisionstheoretische Analyse sogenannter nicht wohlfundierter Mengen, welche hier eine Rolle spielen, hat z.B. Aldo Antonelli in seinem Aufsatz [1] gegeben.

---

[150] [18], S. 263-277
[151] [18], S. 235-252
[152] [18], S.270

## 9.1 Referenz

Schauen wir uns einmal genauer an, wie Belnap und Gupta die Zirkularität für den Referenzbegriff folgern.[153] Die zweistellige Relation 'denotiert' zwischen einem singulären Terminus und dem Gegenstand, welcher durch den singulären Terminus bezeichnet wird, ist nach Gupta und Belnap durch partielle Definitionen der folgenden Form gegeben:

(d) $(\forall x)$('t' denotiert x $=_{Df}$ t ist (identisch) mit x.)

Einige der Instanzen von der Form (d) enthalten den Ausdruck 'denotiert' im Definiens, z.B. die von Gupta und Belnap angeführten beiden folgende Sätze

> $(\forall x)$('Der Gegenstand, welcher durch 'Sokrates' denotiert wird' denotiert x $=_{Df}$
> der Gegenstand, der durch 'Sokrates' denotiert wird, ist (identisch mit) x.)
>
> $(\forall x)$ ('Die kleinste Zahl nicht benennbar mit weniger als neunzehn Silben' denotiert x $=_{Df}$
> Die kleinste Zahl nicht benennbar mit weniger als neunzehn Silben ist (identisch) mit x.)

Das letzte Beispiel übernimmt eine prominente Rolle im sogenannten Berry-Paradoxon, welches sich folgendermaßen liest: Es gibt eine Zahl n, die durch den Berryschen Ausdruck 'Die kleinste Zahl nicht benennbar mit weniger als neunzehn Silben' denotiert wird. Das folge daraus, daß es ja lediglich eine endliche Anzahl von singulären Termen gibt, die mit einer gegebenen endlichen Anzahl von Silben formuliert werden kann. Der Berrysche Ausdruck enthält aber achtzehn Silben. Wenn wir folglich annehmen, daß der Satz

> Die kleinste Zahl nicht benennbar mit weniger als neunzehn Silben ist identisch mit n.

wahr ist, dann ist sie falsch, da n ja durch den nur achtzehn Silben zählenden Berryschen Ausdruck denotiert wird. Ist dieser Satz hingegen falsch, dann folgt daraus seine Wahrheit.[154]

Eine kritische Auseinandersetzung mit der These, daß der Referenzbegriff zirkulär ist, wird nicht mit den kritischen Anmerkungen getan sein, die für den Wahrheitsbegriff gemacht wurden. Allgemeine Einwände gegen die Existenz zirkulärer Begriffe würden natürlich auch hier einschlägig sein. Aber für die Beurteilung der materialen Adäquatheit der partiellen Definitionen für den Referenzbegriff sind unsere sprachlichen Intuitionen gefragt, die die Ausdrücke 'Referenz',

[153] [18], S.264

[154] Der letzte Satz ist problematisch, aber man könne in einem ähnlichen, etwas komplexeren Argument eine Paradoxie ableiten ([18], S.264, Fn 23).

‘denotiert’ bzw. ‘bezeichnet’ betreffen. Diese mögen für den Referenzbegriff was die Beurteilung von paradoxen Sätzen betrifft, die eines der Ausdrücke ‘denotiert’, ‘referiert’ oder ‘bezeichnet’ enthalten, ähnlich sein wie für den Wahrheitsbegriff. Dennoch ist für jeden einzelnen Begriff, der Kanditat für Zirkularität ist, zu überprüfen, ob materiale Adäquatheit der zirkulären Definition gegeben ist.

# 10 Kritik an (RTW) und (RTD)

In seiner Antwort auf die Kritiken von Martin ([31]) und McGee ([34]) faßt Gupta die Hauptthesen der Revisionstheorie zusammen:[155]

(A) Zirkuläre Definitionen sind legitim. Man kann solchen Definitionen einen semantischen Sinn abgewinnen.

(B) Wahrheit ist ein zirkulärer Begriff.

(C) Zirkuläre Definitionen erteilen dem Definiendum eine Bedeutung, die einen hypothetischen Charakter hat.

(D) Die semantische Signifikanz [significance] eines zirkulären Prädikats wird durch die Revisionsregel eingefangen. In der Terminologie von RTT *ist* die Signifikation [signification] eines zirkulären Prädikats seine Revisionsregel.

(E) Kategorische Information kann aus der Revisionsregel extrahiert werden. Dafür sollte die Wirkung wiederholter Anwendung der Revisionsregel auf beliebige Hypothesen betrachtet werden.

(F) Diejenigen Objekte, die in allen Revisionsfolgen für G nach bestimmten Revisionsstufen immer in der Extension des definierten Prädikats G sind (bzw. die nach bestimmten Revisionsstufen immer außerhalb der Extension von G sind), erfüllen intuitiv ohne Einschränkung G (bzw. nicht-G).

Die Kritiken an der (RTW) können sinnvollerweise eingeteilt werden in solche, die sich auf A oder B beziehen und solche, die den Rest betreffen. Die zweite Sorte von Kritiken, und hierzu zählen die Kritiken von McGee ([34]), betreffen hauptsächlich die spezielle Semantik, die (RTD) zirkulären Definitionen gibt. Sie können weiter unterschieden werden als Kritik
a) an der Behandlung der Limespunkte in Revisionsprozessen oder
b) an der (RTD) im Vergleich zu den Rivalen von (RTD) in Bezug auf die Frage, wie aus den Revisionsregeln kategorische Aussagen zu schöpfen sind oder
c) an der Anwendbarkeit der Revisionstheorie auf natürliche Sprachen.

Die einzige direkte Kritik an der These (B), die mir bekannt ist, gibt Koons in seiner Rezension [23]. Leider läßt sich nur schwer rekonstruieren, was der Kern der Koonsschen Kritik ist, da er auf einem offensichtlichen Mißverständnis oder anderem Verständnis der Ausdrücke ‘intensional äquivalent’ und ‘intensional adäquat’ beruht.[156] Ich werde diese Kritik daher übergehen.

[155] [17], 419-422

[156] Koons leugnet, daß man aus der intensionalen Äquivalenz eines Prädikats H und eines Prädikats G, welches zirkulär definiert wird, darauf schließen kann, daß die Definition von G eine intensional adäquate Definition für H ist. Erst dann aber, wenn letzteres gesichert sei, könne man behaupten, daß H ein zirkulärer Begriff ist.

Weitere indirekte Kritik an der These, daß der natürlichsprachliche Begriff der Wahrheit ein zirkulärer ist, findet man bereits in der RTT von Belnap und Gupta vorweggenommen und besprochen. Diese werde ich im folgenden Unterabschnitt darstellen. In dem darauf folgenden Abschnitt bespreche ich die Einwände von McGee, sofern diese nicht bereits berührt wurden. Zunächst will ich aber untersuchen, ob sich einige der oben dargestellten Einwände gegen die Tarskische Wahrheitskonzeption auch gegen die (RTW) anwenden lassen.

## 10.1 Kritik an Tarski angewandt auf (RTW)

Betrachten wir noch einmal die kritischen Anmerkungen, die ich im Anschluß an die grobe Skizze von Tarskis semantischer Wahrheitskonzeption aufgelistet habe. Welche dieser Kritiken macht noch Sinn für (RTW)? Welche von den kritischen Punkten verliert an Brisanz?

Die gesamte Kritik am definitorischen Anspruch der Tarskischen Wahrheitskonzeption verliert für die Revisionstheorie natürlich ihren Gegenstand: Gupta und Belnap schreiben explizit, daß sie die Definitionen der Wahrheitsprädikate mit einem nicht stärkeren Standard als der intensionalen Adäquatheit bewertet wissen wollen; die Bikonditionale müssen also nicht den Sinn der Wahrheitsprädikate festlegen – und tun das auch nicht, wie man sich durch das Projektionsproblem und das Problem des epistemischen Status hat klarmachen können; sie sollen lediglich, so sagt es die Signifikationsthese, die Signifikation der Prädikate in allen möglichen Welten festlegen. Es war bereits im kritischen Teil die Frage angeklungen, ob die Projektionsproblematik auch dann noch besteht, wenn mit der Definition der Wahrheitsprädikate nicht mehr der Anspruch verbunden wird, den Sinn anzugeben. Vielleicht läßt sich das folgende Beispiel unter dem Titel Projektionsproblematik subsumieren. Es ist keines, das die (RTW) allein betrifft. Es operiert wesentlich mit der Tatsache, daß in (RTW) Wahrheit relativ zu einer Sprache L definiert wird, und läßt sich daher auch auf fast alle anderen Wahrheitstheorien, die eine derartige Relativierung vornehmen, übertragen.[157]

Für jede klassische unproblematische Sprache L liege also ein Wahrheitsprädikat L vor, welches – im Gegensatz zu Tarskis Konzeption – in der Sprache L selbst vorkommt. In solchen Sprachen wird keine Unterscheidung gemacht zwischen Wahrheitsprädikaten für die eine oder andere Sprache. Es gibt lediglich einen besonders gekennzeichneten Prädikatbuchstaben, etwa T, für den die T-Bikonditionale aufgestellt werden. In solchen Sprachen scheint die Möglichkeit verwehrt zu sein, über die Wahrheitsprädikate anderer Sprachen zu reden. (RTW) scheint nicht erklären zu können, weshalb wir einen Satz wie

(1) Der Satz ‘Serdar hat einiges Wahres auf Türkisch gesagt’ ist wahr

[157]Insbesondere für die Zitattilgungstheorie ist die Problematik einschlägig. Aus diesem Kontext habe ich auch das Argument.

für wahr halten. Ich glaube nicht, daß die Problematik darin besteht, daß (RTW) nichts zum Sinn der Wahrheit sagen möchte. Das Problem ist, daß sich so ein Satz in derart (immer noch) ausdrucksschwachen Sprachen, von denen (RTW) ausgeht, nicht einmal formulieren zu lassen scheint.

Unsere Intuitionen zum Wahrheitsbegriff werden darin übereinstimmen, den Satz (1) für korrekt zu erklären. Dieser Satz ist ein Satz des Deutschen, der zweimal ein Vorkommnis des Ausdrucks(typs) 'wahr' enthält. Beide Vorkommnisse enthalten keine Relativierung auf die deutsche oder türkische Sprache. Für (RTW) (wie aber auch für die Fixpunkttheorrie) allerdings gibt es für eine Sprache L immer nur das eine Wahrheitsprädikat 'wahr in L'. Angenommen wir relativierten die beiden Vorkommnisse von 'wahr' auf die deutsche Sprache bzw. auf ein hinreichend unproblematisches Fragment $L_D$ des Deutschen. Dann erhielten wir, wenn wir zusätzlich die im ersten Vorkommnis gemachte verdeckte Quantifizierung explizierten, den folgenden Satz:

(2) Der Satz 'Es gibt ein x mit: x ist ein Satz des Türkischen & x ist wahr in $L_D$ & Serdar hat x geäußert' ist wahr in $L_D$.

Das Wahrheitsprädikat 'wahr in $L_D$' wird für Basismodelle erklärt. Das heißt u.a., daß der Quantor über einen Wertebereich läuft, der sämtliche (Kodierungen für) Sätze von $L_D$ enthält. Es wird allerdings nicht ausgeschlossen, daß sich im Wertebereich des (übrigens ontisch gelesenen) Existenzquantors auch türkische Sätze finden. Damit wäre das erste Konjunkt aus der Matrix in (2) erfüllbar. Allerdings wird schon das zweite Konjunkt nicht durch einen Gegenstand erfüllt, der das erste Konjunkt erfüllt: Die T-Bikonditionale handeln lediglich von Sätzen der Sprachen $L_D$. Mit den semantischen Begriffen der Revisionstheorie ausgedrückt bedeutet das, daß sich Satz (2) nicht als kategorisch wahr herausstellen läßt. Das wäre aber nötig, wenn mit der (RTW) die Intuition bestätigt werden sollte, daß (1), welches wir durch (2) analysiert haben, korrekt ist.

Ähnliche Probleme erhalten wir, wenn wir die beiden Vorkommnisse des Wahrheitsprädikats in (1) auf die türkische Sprache bzw. ein geeignetes Fragment $L_T$ des Türkischen beschränken würden. In dem Fall ergeben sich dann nämlich Probleme mit dem zweiten Vorkommnis des auf $L_T$ eingeschränkten Prädikats.

Die Lösung, das erste Vorkommnis auf das Fragment des Türkischen $L_T$ zu relativieren und das zweite Vorkommnis auf das Fragment des Deutschen $L_D$, ergibt zwar den korrekt explizierten Satz

(3) Der Satz 'Es gibt ein x mit: x ist ein Satz des Türkischen & x ist wahr in $L_T$ & Serdar hat x geäußert' ist wahr in $L_D$.

Aber dieser ist für die Formulierung von (RTW), die Belnap und Gupta in RTT geben, nicht ausdrückbar, denn es wird immmer nur ein Wahrheitsprädikat 'wahr in L' für eine konkrete Sprache L erklärt.

Zu dieser Frage beziehen Gupta und Belnap indirekt Stellung:[158] Sie geben an, wie die Revisionstheorie auch auf solche Sprachen anzuwenden ist, die ein Prädikat 'ist wahr in **L**' enthalten, wobei der Buchstabe **L** eine Variable ist, welche über Sprachen läuft. Der Ausdruck 'ist wahr in **L**' wäre also ein zweistelliges Prädikat, dessen Extension Paare enthält mit Sätzen an erster und Sprachen an zweiter Stelle. Für eine derartige erweiterte Anwendung der (RTW) müßte aber Belnap und Gupta gemäß[159] u.a. folgendes geklärt werden

- Was für abstrakte Entitäten sind Sprachen?
- Wie hat man sich die Gesamtheit aller Sprachen vorzustellen? Eine Menge kann sie nicht darstellen, auch keine echte Klasse[160], dennoch braucht man solch eine Gesamtheit, um eine semantische Erläuterung von 'ist wahr in **L**' angeben zu können.
- Wie ist mit den Paradoxa umzugehen, die durch diese Begriffe hervorgerufen werden?

Gupta und Belnap überspringen die ersten beiden Fragen und besprechen einen idealisierten Fall, bei dem die Menge X, über die die Variable **L** läuft, lediglich klassische Sprachen erster Stufe beinhaltet; jede dieser Sprache enthält außerdem als einzigen „problematischen Begriff" „wahr in **L**". Der Wertebereich einer jeden Sprache aus X enthält einen Code für jede Sprache **L** und Sätze von **L**. Die Signifikation von 'ist wahr in **L**' ist für jede Sprache dieselbe, nämlich eine Revisionsregel, die als Input eine hypothetische Extension, deren Elemente Paare (s, L) bestehend aus einem Satz s und einem Codenamen L einer Sprache aus X sind, für 'ist wahr in **L**' aufnimmt und als Output eine neue Extension herausgibt.[161] Belnap und Gupta gehen auch der letzten Frage nach.[162] Mich interessiert aber nur, wie das obige Problem zu lösen ist: Man kann einsehen, daß mit der auf diese Weise erklärten Revisionstheorie auch komplizierte Aussagen, die Wahrheitsprädikate aus anderen Sprachen beinhalten, behandelt werden können. Die Revisonsregel für 'ist wahr in **L**' müßte herausgeben, daß der obige Satz (3)

(3) Der Satz 'Es gibt ein x mit: x ist ein Satz des Türkischen & x ist wahr in $L_T$ & Serdar hat x geäußert' ist wahr in $L_D$.

[158] [18], S.265-266

[159] [18], S.265

[160] Warum kann sie letzteres nicht darstellen? Daß es keine Menge sein kann, läßt sich vielleicht noch mit Kardinalitätserwägungen einsehen.

[161] Es kann sich durchaus herausstellen, daß sich in keiner der Sprachen aus X die T-Bikonditionale für 'ist wahr in **L**' ausdrücken lassen, da es eine Sprache geben kann, in der sich nicht all das ausdrücken läßt, was in den anderen Sprachen aus X ausdrückbar ist. Das läßt sich Belnap und Gupta gemäß z.B. von den natürlichsprachlichen Wahrheitsbegriffen einsehen, wenn man bedenkt, daß nicht alles das, was sich mit dem Türkischen sagen läßt, sich unbedingt auch mit der deutschen Sprache muß sagen lassen können, et vice versa. Folglich könne es sein, daß man im Türkischen nicht die Anwendungsbedingungen von 'ist wahr in der deutschen Sprache' wiedergeben kann – und umgekehrt. ([18], S. 266)

[162] [18], S.266

kategorisch wahr in $L_D$ ist, wenn mein Bruder Serdar tatsächlich einiges Wahres in türkischer Sprache gesagt hat. Wie wir mitgeteilt bekommen haben, kann es aber durchaus sein, daß keine der Sprachen, über die die Variable läuft, die Tarski-Bikonditionale ausdrücken kann, welche die Revisionsregel festlegen. Dann könnte ich z.B. im Deutschen nicht angeben, wie die (hypothetischen) Anwendungsbedingungen für 'wahr in **L**' sind. Ich vermute Belnap und Gupta denken an eine Metasprache zu allen Sprachen aus X, in der die partiellen Definitionen, welche die Revisionsregel festsetzen, formuliert werden können.

Übrigens sagt die (RTW) nichts darüber aus, wie ein des Türkischen nicht Mächtiger auf der Grundlage der als partielle Definitionen aufgefaßten Tarski-Bikonditionale dazu kommen kann, Satz (3) bzw. den eingebetten Satz in (3) für wahr (in der deutschen Sprache) zu erklären. Dafür sind die partiellen Definitionen in (RTW) nicht konzipiert.

Eine andere (von Belnap und Gupta nicht erwogenene) Möglichkeit, die (RTW) auch auf solche Sätze wie (1) anwendbar zu machen, besteht in einem Rekurs auf den Begriff der Übersetzung oder Synonymie. Der Revisionstheoretiker könnte mit diesem Begriff (1) semiformal folgendermaßen lesen:

(4) Der Satz '$(\exists x)(\exists y)(x$ ist ein Satz von $L_T$ & Serdar hat $x$ geäußert & y ist eine Übersetzung von $x$ in $L_D$ & $y$ ist wahr in $L_D)$' ist wahr in $L_D$.

Der in (4) eingehende Übersetzungsbegriff muß aber geklärt werden. Probleme können sich ergeben, wenn man für die Erläuterung des Übersetzungsbegriffs auf den Begriff der Wahrheit rekurrieren muß.

Die Problematik der Anwendbarkeit auf natürliche Sprachen im Falle von Tarskis Theorie besteht auch für (RTW) – bis auf eine Ausnahme: Man kommt den natürlichen Sprachen insofern näher, als jetzt die Sprachen, für die das Wahrheitsprädikat erklärt wird, das Wahrheitsprädikat bereits enthalten. Es gibt keine Trennung zwischen Objektsprache und Metasprache bezüglich des Wahrheitsprädikats. Außerdem hat man in diesen Sprachen die Möglichkeit, auf die Sätze dieser Sprache Bezug zu nehmen.

Ein anderes Problem betrifft die vermeintliche Universalität der natürlichen Sprachen. Es wird im folgenden Abschnitt zu den von Belnap und Gupta besprochenen Einwänden behandelt.

## 10.2 Von Belnap und Gupta besprochene Einwände

Die von Belnap und Gupta besprochenen Einwände betreffen in der Hauptsache ihre Theorie der Wahrheit (RTW) und nicht die darunter liegende Definitionstheorie (RTD). Die vier Einwände, welche diskutiert werden, sind:

1. Das Problem mit einer stärkeren Version der Lügnerparadoxie: „Dieser Satz ist nicht kategorisch oder nicht wahr."

2. Der Einwand, daß Belnaps und Guptas Vorschlag zur Lösung des Problems unter 1. im wesentlichen auf den Aufbau einer Tarskischen Hierarchie hinausläuft und nichts beiträgt zu der Entwicklung einer universellen Sprache.
3. Die vermeintliche Komplexität ihrer Erklärung des naiven Wahrheitsbegriffs.
4. Die Beibehaltung der klassischen logischen Gesetze in (RTD): Der Satz "Dieser Satz ist falsch' ist wahr oder es ist nicht der Fall, daß der Satz 'Dieser Satz ist falsch' wahr ist' stellt sich als gültig gemäß (RTD) heraus.

### 10.2.1 Eine stärkere Version des Lügnersatzes

Das erste der möglichen Einwände gegen (RTW), welche Belnap und Gupta im 7. Kapitel besprechen, betrifft den sogenannten „Strengthened Liar".[163] Auf die Revisionstheorie bezogen ist z.B. der folgende Satz ein solcher „Strengthened Liar":

(1) Enweder dieser Satz ist nicht kategorisch oder er ist nicht wahr.

Der Einwand gegen die Revisionstheorie auf der Grundlage von (1) lautet: Auch wenn die Revisionstheorie vielleicht eine erhellende Erläuterung der einfachen Lügnerparadoxie geben kann, verstrickt sie sich doch gerade mit den semantischen Mitteln, die sie zur Erläuterung der einfachen Lügnerparadoxie benutzt, in neue Paradoxien, für welche der „Strengthened Liar" ein Beispiel ist. Die entscheidende Frage, welche die Paradoxie zu Tage fördert, ist, ob Satz (1) kategorisch ist oder nicht. Angenommen (1) ist nicht kategorisch; dann ist (1) wahr, da dessen erstes Disjunktionsglied damit wahr ist. Folglich wäre (1) im Widerspruch zur Annahme kategorisch. Also müssen wir annehmen, daß (1) doch kategorisch ist. Damit aber ist das erste Disjunktionsglied falsch. Folglich hängt die Wahrheit von (1) vom zweiten Disjunktionsglied ab. (1) verhält sich daher wie der Lügnersatz 'Dieser Satz ist nicht wahr', welcher aber nicht kategorisch ist. Damit wäre also auch (1) im Widerspruch zur Annahme nicht kategorisch.

In ihrer Antwort auf diesen Einwand weisen Belnap und Gupta darauf hin, daß in die Ableitung der Widersprüche zwei Annahmen eingehen. Die erste Annahme, welche in der Ableitung des ersten Widerspruchs verwandt wird, kann durch Satz (2) ausgedrückt werden:

(2) Alle wahren Sätze sind kategorisch.

Satz (2) ist aber problematisch, da sich aus ihm ein paradoxer Satz ableiten ist: Der Lügnersatz ist nicht kategorisch. Hieraus und aus (2) folgt durch Kontraposition, daß der Lügnersatz nicht wahr ist. Dieser so abgeleitete Satz 'Der Lügnersatz

[163] [18], S.253f.

ist nicht wahr' ist aber Gupta und Belnap gemäß paradox.[164] Wenn man mit Belnap und Gupta die Aussage akzeptiert, daß der Lügnersatz nicht kategorisch ist, dann muß Satz (2) für die Problematik verantwortlich gemacht werden. Worin genau besteht nun nach Belnap und Gupta die Problematik mit (2)? Diese sehen sie in einer weiteren Annahme. Die Annahme ist, daß in der Sprache L ein Prädikat 'ist kategorisch' zur Verfügung steht, welches den Begriff „kategorisch" ausdrückt. Diese Annahme dürfe zwar gemacht werden, man könne aber nicht erwarten, daß für ein solches Prädikat die üblichen logischen Regeln gelten würden und die übliche Semantik angewandt werden könne. Stattdessen müsse man annehmen, daß dieses Prädikat in L einen zirkulären Begriff ausdrückt – ganz genau so wie auch das Wahrheitsprädikat 'ist wahr in L' in L einen zirkulären Begriff ausdrückt. Für solche Begriffe müssen natürlich auch andere logische Regeln zugrunde gelegt werden, wie es z.B. in dem dargestellten Kalkül $C_0$ der Fall ist. Legt man einen solchen Kalkül zugrunde, dann läßt nicht mehr wie oben die Aussage ableiten, daß der Satz (1) kategorisch ist. Was sich zeigen läßt ist, daß wenn Satz (1) nicht in die Extension von 'ist kategorisch' in einer Revisionsstufe gehört, es in der nächsthöheren dann doch zu ihr gehört. Die Sätze 'Satz (1) ist kategorisch' und 'Satz (1) ist nicht kategorisch' sind damit in der Terminologie der Revisionssemantik pathologisch, genauer: paradox in dem Sinne, wie er in der (RTD) erklärt wird. Damit sind die Sätze 'Satz (1) ist kategorisch' und 'Satz (1) ist nicht kategorisch' nicht L-kategorisch. 'L-kategorisch' ist dabei ein Ausdruck, der verschieden ist von dem, der in den Sätzen 'Satz (1) ist kategorisch' bzw. 'Satz (1) ist nicht kategorisch' vorkommt: Er ist ist ein höherstufiger Begriff des Kategorischen.

Für die Auflösung der Paradoxie ist also die Entwicklung einer Semantik nötig, deren Begriffe nicht der Sprache angehören, für die die Semantik entwickelt wird. Z.B. erfolgt bei der obigen stärkeren Version der Lügnerparadoxie die Erläuterung Belnap und Gupta gemäß durch einen Begriff „ist L-kategorisch", der verschieden ist von dem Begriff, welcher ausgedrückt wird durch das Prädikat, welches in (1) vorkommt. Die Diagnose ist, daß das Prädikat 'ist kategorisch', welches in (1) vorkommt, einen zirkulären Begriff ausdrückt. Für zirkuläre Begriffe wurde eine Semantik entwickelt. Mit deren Hilfe läßt sich sagen, daß Satz (1) nicht L-kategorisch ist. Diese Aussage selbst ist nicht problematisch, da 'kategorsich' und 'L-kategorisch' andere Begriffe ausdrücken.

Nun könnte man eine noch stärkere Version der Lügnerparadoxie formulieren. Die würde dann z.B. durch den folgenden Satz (3) ins Rollen gebracht werden.

[164]In einer Fußnote ([18], S.255, Fn 5) weisen Belnap und Gupta darauf hin, daß Skyrms ([40]) und Gaifmann ([14]) 'ist wahr' so lesen, daß der Satz 'Der Lügnersatz ist nicht wahr' durchaus nicht paradox ist. Die Autoren von RTT glauben aber, daß das nicht die gewöhnliche Bedeutung des Ausdrucks 'ist wahr' wiedergibt. Man sei doch geneigt, auf die Behauptung, der Lügnersatz sei nicht wahr, zu erwidern, dann sei der Lügnersatz ja doch wahr, weil es gerade das Behauptete bestätige. Damit stellt sich auch die Behauptung, der Lügnersatz sei nicht wahr, als paradox heraus.

(3) Dieser Satz ist entweder nicht L-kategorisch oder nicht wahr.

Die Ableitung der Widersprüche würde ähnlich wie oben erfolgen, und ähnlich wie oben würden Gupta und Belnap diese noch stärkere Paradoxie erklären, indem sie auf einen Begriff „*-kategorisch“ zurückgreifen würden, der verschieden ist von dem Begriff „L-kategorisch“. Belnap und Gupta geben also nicht nur eine Lösung der Lügnerparadoxie, sondern ein Anweisungsschema, mit dem man alle ähnlichen Paradoxa lösen kann. Dieser Schachzug könnte, so erwägen Belnap und Gupta, wieder Gegenstand für Kritik sein: Ist ihr Zugang zu den Paradoxien nicht im wesentlichen der Tarskische Zugang?[165] Zwar gibt es keine Unterscheidung zwischen Objekt- und Metasprache, die den Wahrheitsbegriff betrifft – und entsprechend gibt es keine hierarchische Sprachstruktur mit einem eigenen Wahrheitsbegriff für jede Sprachstufe; aber dafür bedarf es einer Unterscheidung zwischen Objekt- und Metasprache bzgl. des Begriffs des Kategorischen – und infolge dieser Unterscheidung hat man eine hierarchische Sprachstruktur mit einem Prädikat ‘ist kategorisch in L’ für jede Sprachstufe.

Zur Klarstellung muß betont werden, daß Guptas und Belnaps Theorie keine hierarchische Sprachstruktur bezüglich des Wahrheitsprädikats der natürlichen Sprache Deutsch zur Folge hat: In der gibt es (RTW) gemäß nur das eine (logische, schwache, absolute) Wahrheitsprädikat und nicht unendlich viele (logische, schwache, absolute) Wahrheitsprädikate für jede innerhalb des Deutschen zu unterscheidende Sprachstufe. Für die deutsche Sprache postuliert (RTW) keine hierarchische Sprachstruktur bzgl. des Wahrheitsbegriffs. Insofern geraten Belnap und Gupta nicht in einen augenscheinlichen Widerspruch, wenn sie einerseits sämtliche Wahrheitstheorien, die zum Inhalt haben, daß die natürliche Sprache bzgl. der Wahrheit hierarchisch aufgebaut ist, verwerfen, andererseits ihre Theorie (RTW) propagieren, die scheinbar eben eines dieser verworfenen Theorien ist. Belnap und Gupta würden aber auch die These verwerfen, daß aus der (RTW) folgt, daß die normale deutsche Sprache, die Ottonormalverbraucher beherrscht, bzgl. des Begriffs des Kategorischen hierarchisch aufgebaut sei: Die verschiedenen Prädikate ‘ist kategorisch’ kommen zur gewöhnlichen deutschen Sprache hinzu – und sind nicht bereits in ihr enthalten. Tatsächlich ist in der natürlichen Sprache Deutsch bzw. Englisch bis zur Einführung des technischen Begriffs des Kategorischen durch die (RTD) kein Pendant enthalten.

McGee hält die von der (RTW) gegebene Lösung der Paradoxie für nicht befriedigend. Seine Kritik beinhaltet, daß mit der Revisionstheorie die eigentliche Aufgabe, welches für eine erhellende Erläuterung der Lügnerparadoxie nötig sei, nicht gelöst werde:

> ... if we can only give the semantics of our simplified language within an essentially richer metalanguage, the fundamental and difficult problem of how to give the semantics for a language within the language itself will still remain before us. ([33], S.147)

[165] [18], S.256

McGees Vorstellung ist, daß die natürliche Sprache Englisch bzw. Deutsch hinreichend universelle Züge tragen dürfte, um mit den in ihr vorliegenden Begriffen die Semantik dieser Sprachen zu geben. Und selbst wenn das nicht der Fall sein sollte, dann, so McGees Hoffnung, wird sich das Deutsche/Englische durch Hinzunahme bestimmter Begriffe soweit entwickelt haben, daß sie ihre eigene Semantik in groben Zügen skizzieren kann. Belnap und Gupta gestehen ein, daß ihre Revisionstheorie überhaupt nichts zu diesem Problem beiträgt. Mehr noch glauben sie, daß nichts dazu beitragen könnte; eine derartige Aufgabe zu stellen und zu lösen zu versuchen, könne durch ein auf Kant gehendes Bild illustriert werden könne: Da versucht jemand, einen Ochsen zu melken, und ein anderer hält ein Sieb drunter. Ob die Verhältnisse tatsächlich so hoffnungslos sind, wie es dieses Bild nahelegt, vermag ich nicht zu beurteilen. Tarskis Diktum, natürliche Sprachen seien universell, versuchen sich die Autoren anhand von zwei Lesarten verständlich zu machen.[166] Nach der ersten Lesart ist eine natürliche Sprache universell, wenn in ihr alles – so wie sie zum gegebenen Zeitpunkt beschaffen ist – ausgedrückt werden kann, was überhaupt ausdrückbar ist. In dem Sinne ist die natürliche Sprache sicher nicht universell: Die Unmengen an Ergebnissen, die in den Natur-und Geisteswissenschaften von Tag zu Tag hervorgebracht werden, sind sicher nicht vernünftig darstellbar, ohne die Einbeziehung neuen Vokabulars. Vielleicht ist aber gemeint, daß in ihnen alles ausdrückbar ist, sofern hinreichend viele sprachliche Ressourcen noch hinzugenommen werden? In diesem Falle müßte man aber auch von einfachen künstlichen Sprachen sagen, sie seien universell – was anscheinend nicht intendiert ist. In der anderen Lesart von „universell“ stellt sich die Aussage als harmlos heraus: Zu jedem semantischen Begriff C gibt es eine Sprache L, die ihren eigenen C-Begriff enthält. Umgekehrt sind Belnap und Gupta nicht bereit, die These zu akzeptieren, daß es eine Sprache gibt, die für jeden semantischen Begriff C den eigenen C-Begriff enthält. Eine echte Begründung hierfür geben sie nicht an. Es gibt allerdings bisher auch kein Argument dafür, daß es solch eine Sprache gibt.

Da es neben der (RTW) andere Wahrheitstheorien gibt, die ebenfalls nicht das Problem der semantischen Selbstsuffizienz zu lösen vermögen, ist die (RTW) damit nicht als Wahrheitstheorie aus dem Rennen.

### 10.2.2 Die Komplexität von (RTW)

Einen weiteren Einwand, den man gegen (RTW) erheben könnte, betrifft die Komplexität dieser Theorie: Ihre Idee, so leiten Belnap und Gupta diesen Einwand ein[167], sei es gewesen, den natürlichsprachlichen Wahrheitsbegriff zu be-

[166]Ich glaube, daß keine der Lesarten das wiedergibt, was Tarski meint, wenn er sagt, die natürliche Sprache sei universell. Das meint bei ihm nur, daß die natürliche Sprache prinzipiell die Möglichkeit hat, auf alle ihre Sätze Bezug zu nehmen. Die natürliche Sprache ist eine ausdrucksstarke Sprache, in diesem Sinne technischen Sinne ist sie universell.

[167][18], S. 259-261

schreiben. Die Komplexität, welche sich durch die Einbeziehung des Revisionsprozesses ergibt und schließlich in den beiden mit transfiniten Ordinalzahlen arbeitenden semantischen Systemen $S^*$ und $S^\#$ gipfelt, mache es unwahrscheinlich, daß hier jener natürlichsprachliche Begriff adäquat beschrieben werde, über den man schon im Kindesalter verfügt: Als Kind weiß man nichts von transfiniten Ordinalzahlen; und wer gelangt schon, selbst nachdem er dem Kindesalter entwachsen ist, zu irgendwelchen Wahr-falsch-Urteilen, indem er in Gedanken irgendeinen Revisionsprozeß in Gang setzt, der letztendlich darüber entscheidet, ob ein Satz kategorisch wahr ist oder nicht? In welchem Sinn also beschreibt die Revisionstheorie der Wahrheit den gewöhnlichen Wahrheitsbegriff? Belnaps und Guptas Antwort ist dreierlei:

1.) Die Kernkonzeption ihrer Theorie sei überhaupt nicht komplex. Sie gingen davon aus, daß die T-Bikonditionale den Wahrheitsbegriff definieren. Infolge dieser Überlegung stelle sich heraus, daß der Wahrheitsbegriff zirkulär sei und daß dessen Bedeutung durch eine Regel mit hypothetischem Charakter gegeben sei. Es gebe nichts an dieser Konzeption, was man nicht problemlos den Benutzern von 'wahr' – ob nun Kind oder Erwachsener – zuschreiben könne. Man könne nicht leugnen, daß die T-Bikonditionale integraler Bestandteil der Sprachpraxis sind. Sie benutzten wir, wenn wir eine These beurteilten, die die Wahrheit oder Falschheit einer Aussage behauptet. Ich glaube, was den ersten Punkt betrifft, nämlich, daß ihre Kernkonzeption nicht komplex ist, dem läßt sich zustimmen. Die schwierigen Probleme ergeben sich erst dann, wenn man versucht, die Frage zu beantworten, wie aus der hypothetischen Regel kategorische Information zu schöpfen ist. Was den zweiten Punkt betrifft, nämlich, daß es nichts an der Konzeption gebe, was man den Benutzern von 'wahr' nicht problemlos zuschreiben könne, so bin ich mir nicht sicher, wie das genau zu verstehen ist. Ist etwa gemeint, daß wenn man einen Benutzer von 'ist wahr' danach fragte, ob die Konzeption von Belnap und Gupta überhaupt zu verstehen ist, dann ein Großteil vermutlich ja antworten wird? Oder ist gemeint, daß wenn man sie fragte, ob die Konzeption von Belnap und Gupta mit ihrem Gebrauch von 'ist wahr' kompatibel sei, sie dann alle ja antworten würden? Oder ist gemeint, daß wenn man Belnaps und Guptas Konzeption akzeptierte, dann hieraus noch kein Komplexitätsargument geschmiedet werden kann, dessen Konklusion ist: Aus der Komplexität der Konzeption folgt, daß Belnaps und Guptas Wahrheitstheorie nicht zur Beschreibung des natürlichsprachlichen Wahrheitsbegriffs verwandt werden kann. Die letzte Lesart scheint mir die wohlwollendste zu sein. Tasächlich kann der Einwand, der die Komplexität der Wahrheitstheorie (RTW) auf die Komplexität der Konzeption zurückführt, nicht fruchten, da die Konzeption schlichtweg nicht komplex ist. Ob die Konzeption richtig oder falsch ist, ist mit diesem Urteil natürlich nicht entschieden: Es gibt viele Aussagen oder Konzeptionen, die nicht besonders komplex sind, obwohl sie dennoch falsch sind. Was den dritten Unterpunkt betrifft, nämlich, daß die Bikonditionale einen integralen Bestandteil unser Sprachpraxis bilden, sobald irgendwie der Wahrheitsbegriff eine Rolle spielt, so

ist nicht zu leugnen, daß sie bei bestimmten Argumentationen explizit oder implizit verwendet werden. Aber verstehen wir die Bikonditionale auch als partielle Definitionen? Auch wenn Tarski einmal diese Überlegung geäußert hat, so wird er nicht die Konzeption im Hinterkopf gehabt haben, die Belnap und Gupta hatten; zum zweiten aber ist damit nicht gesagt, daß der Großteil der Benutzer von 'ist wahr' dieses Verständnis haben. Dieser letzte Unterpunkt von Belnap und Gupta betrifft allerdings schon eine Plausibilisierung der Grundkonzeption und nicht die Frage ihrer Komplexität – auch wenn manchmal eine Plausibilisierung schon zum Nachweis der Nichtkomplexität dienen kann.

2.) Die Komplexität ihrer Wahrheitstheorie, so Belnaps und Guptas zweiter Punkt, resultiere nicht aus dem Grundgedanken ihrer Theorie, sondern dem Versuch, kategorische Information aus der hypothetischen Regel, die dem Wort 'ist wahr' unterliege, zu gewinnen. Und die Komplexität hier resultiere daraus, daß in der Theorie sämtliche Situationen in Betracht gezogen würden, insbesondere die, in denen es einen unendlichen Querbezug gebe.[168] Wenn man sich auf endliche Situationen beschränke, dann sei es nicht nötig, transfinite Revisionsfolgen zu betrachten. Was heißt es, sich auf endliche Situationen zu beschränken? Die durch den Gebrauch an anderer Stelle ihres Buches nahegelegte Interpretation ist: Es liegt ein Modell mit endlichem Gegenstandsbereich und ein eine endliche Menge an Definitionen vor.[169] Unter diesen Umständen, wie schon im darstellenden Teil zu $S^{\#}$ erwähnt, sind alle Sätze gültig gemäß $S^{\#}$ genau dann, wenn sie es gemäß $S_0$ sind. Für letzteres aber brauchten wir keinen unendlichen Revisionsprozeß zu betrachten. Was aber soll das zeigen? Was ist der Wert dieser Aussage? Gegen was für einen Einwand wenden sich Belnap und Gupta mit dieser Erklärung? Dieser Einwand kann doch nur der sein, daß die Revisionstheorie der Wahrheit nicht den üblichen Begriff der Wahrheit beschreiben kann, weil kein gewöhnlicher Mensch transfinite Revisionsprozesse in Gedanken durchgeht, um kategorische Information zu enthalten. Belnaps und Guptas Anwort hierauf ist doch, daß Ottonormalverbaucher für seine Zwecke gar nicht transfinite Revision braucht: Er hat es mit endlichen Situationen zu tun; bei diesen Situationen liefert $S_0$ die kategorische Informationen, die man auch intuitiv für korrekt erklären würde; und $S_0$ ist so einfach, da kann man es Ottonormalverbraucher zutrauen, daß er es irgendwie beherrscht. Wenn das der Punkt sein sollte – was ich gehörig bezweifle – dann würde der folgende dritte Punkt keinen Sinn mehr machen: In dem behaupten Belnap und Gupta nämlich, mit ihrer Theorie würden sie nicht beanspruchen,

[168] Man bedenke, daß Belnap und Gupta die Theorie für den allgemeinen Fall einer möglicherweise unendlichen Menge $\mathcal{D}$ von vorliegenden Definitionen entwickeln. Sind $G$ und $G_1$ zwei Definienda aus der Menge $\mathcal{D}$, so sagt man, $G$ hängt von $G_1$ ab, wenn im Definiens für $G$ $G_1$ vorkommt. (Zur Terminologie siehe [18], S.149, Definition 5A.7) Nun kann es durchaus eine unendliche Kette von Abhängigkeiten zwischen den Definienda $G_i$ geben (mit $G_i \neq G_j$ für $i \neq j$): $G_1$ hängt von $G_2$ ab, $G_2$ hängt von $G_3$ ab, $G_3$ hängt von $G_4$ ab usw. Insofern kann es also einen unendlichen Querbezug geben, der in der Revisionstheorie behandelt wird.

[169] [18], S.184

die psychologischen Vorgänge bei denjenigen zu beschreiben versuchen, die den Wahrheitsbegriff gebrauchen. Wenn es aber nicht der Punkt ist, den Belnap und Gupta machen, dann glaube ich, muß man es als prophylaktische Maßnahme verstehen. Sie stellen sich vermutlich jemanden vor, der einen Komplexitätseinwand darüber macht, daß er a) vernünftige Lesarten von 'Komplexität', 'Beschreibung eines Begriffs' und 'Verfügen über ein Begriff' gibt und b) zeigt, daß zwischen dem Komplexitätsgrad einer Beschreibung eines Begriffs C und den kognitiven Fähigkeiten, die jemand besitzt, der über den Begriff C verfügt, ein proportionales Verhältnis bestehen muß. Wenn tatsächlich so etwas nachgewisen würde, dann könnten auch Belnap und Gupta erklären, weshalb auch ein Kind über den Wahrheitsbegriff verfügen kann: $S_0$ ist nicht komplex.

3.) Der letzte Punkt scheint mir der wichtigste zu sein. Belnap und Gupta geben an, daß sie unter dem Projekt, den Wahrheitsbegriff zu beschreiben, etwas anderes verstehen, als die psychologischen Vorgänge zu beschreiben, die mit dem Erwerb des Wahrheitsbegriffs oder seinem Gebrauch verbunden sind. Diese Projekte seien sehr unterschiedlich: Bei letzterem müsse man sich nicht Gedanken über den Gebrauch des Wahrheitsbegriffs machen, der die menschlichen Fähigkeiten übersteige – z.B. die die Anwendung von 'ist wahr' auf Sätze, die sehr lang und/oder komplex sind. Dafür müsse man hier aber versuchen, systematische Mißverständnisse beim Gebrauch eines Begriffs zu erklären.[170] Beim ersten Projekt seien die Verhältnisse genau umgekehrt. Hier müsse der Gebrauch des Begriffes auch in den Fällen behandelt werden, die die menschliche Fähigkeit übersteigen. Wie ist das genau zu verstehen? Das obige Beispiel von den sehr langen und komplexen Sätzen weist den Weg: Es muß auch für solche Sätze a, die wir tatsächlich nie werden erfassen können, geklärt werden, was die Anwendungsbedingungen von 'a ist wahr' sind. Belnap und Gupta sagen explizit, daß sie mit ihrer Revisionstheorie nicht beanspruchen, die Wege zu klären, auf denen Benutzer des Wahrheitsbegriffs dazu kommen zu behaupten, dieser Satz sei wahr und dieser falsch; noch will die Revisionstheorie die Wege klären, auf denen sie den Wahrheitsbegriff erwerben. Jemand, der das Urteil fällt, dieser oder jener Satz sei wahr, der gelangt zu diesem Urteil nicht dadurch, daß er irgendwelche Hypothesen revidiert. Worin genau besteht nun aber die Beschreibung eine Begriffs? Belnap und Gupta schreiben: „It aims instead to illuminate the concept of truth by giving a systematic account of the judgements to which these ways lead."([18], S.261) Was ist eine systematische Erläuterung der Urteile, zu denen wir gelangen? Leider sagen Belnap und Gupta nichts Genaueres hierüber. Wir haben bei der Aufgabenstellung, die Belnap und Gupta für (RTW) gestellt haben, gesehen, daß es eines ihrer Ziele war, eine systematische Erklärung der Signifikation des Wahrheitsbegriffs zu geben, die eine Einteilung der Sätze in wahr, falsch,

[170]Ein Beispiel, was das Konditional betrifft: Ein sehr beliebter und immer wieder gemachter Fehlschluß ist, aus einem Satz der Form $\ulcorner A \rightarrow B \urcorner$ und dem Konsequens B A zu folgern. Die Logik hat sich nicht darum zu kümmern, weshalb dieser Fehlschluß so häufig gemacht wird.

paradox etc. liefert, welche unseren üblichen Intuitionen entspricht. Was Belnap und Gupta nicht werden sagen können ist, daß die gemäß der Revisionssemantik gültigen (kategorisch wahren) Sätze genau die sind, die man üblicherweise für wahr hält: Denn es gibt sehr viele Sätze, über die wir noch keinen Gedanken verschwendet haben. Dann meint wohl die Einteilung unseren Intuitionen gemäß, daß die kategorisch wahren Sätze genau die sind, denen wir intuitiv zustimmen würden, sobald wir über sämtliche nichtsemantische Fakten verfügten, keine Fehler machen würden und unbegrenzte Rechenfähigkeiten besäßen.[171] Aber selbst diese unmenschlichen Fähigkeiten würden uns nicht helfen, sämtliche kategorisch wahren Sätze der Wahrheitstheorie für die Sprache der Arithmetik zu produzieren.[172]

Für die meisten Sätze aus unserer Praxis, die wir für wahr halten, können wir einsehen, daß sie gemäß der Revisionssemantik als kategorisch wahr herausgestellt werden. Was die vielen anderen als kategorisch wahr herausgestellten Sätze betrifft, müssen wir dergleichen sagen wie, daß wenn wir die Fähigkeit hätten, diese Sätze aufzufassen, wenn wir über die nichtsemantischen Fakten verfügten und wenn wir keinen Fehler machen würden, wir dann die Disposition hätten, einen kategorisch wahren Satz uneingeschränkt zu bejahen. Ich glaube in dieser dispositionellen Lesart muß man Belnap und Gupta verstehen. Wer sagt uns aber, daß wir unter den hoch künstlichen Umständen tatsächlich die Disposition haben, kategorisch wahre Sätze gemäß $S^{\#}$ bzw. $S^{*}$ intuitiv für korrekt zu erklären? Wie mir scheint nur die gemachte Erfahrung, daß die Revisionstheorie die gewöhnlichen Sätze unser Sprachpraxis, die wir tatsächlich intuitiv uneingeschränkt bejahen würden, als kategorisch wahr ausgibt und die Beobachtung, daß die T-Bikonditionale in den meisten unser Urteile eine wichtige Rolle zu spielen scheinen. Welche anderen ersthaften Gründe es gibt, das scheint mir durch Belnap und Gupta nicht hinreichend geklärt zu sein.

Belnap und Gupta beanspruchen, mit der (RTW) den natürlichsprachlichen Wahrheitsbegriff zu beschreiben. Wird in ihrem Projekt damit auch etwas zum Sinn des Wahrheitsprädikats gesagt? Stellen wir hierzu die Frage, was aus der (RTW) für eine Person S folgt, die beansprucht, 'wahr' in einem anspruchsvolleren Sinn zu definieren, nämlich nicht etwas nur über die Intension oder Signifikation von 'ist wahr' in allen möglichen Welten zu sagen, sondern etwas über seinen Sinn. Wenn S eine Definition angibt, die den Sinn von 'ist wahr' festlegt, dann folgt aus dieser Sinnfestlegung auch die Festlegung der Signifikation in allen möglichen Welten. Hier geht die klassische Vorstellung davon ein, daß der Sinn eines Ausdrucks dessen Extension (Signifikation) in allen möglichen Welten festsetzt. Ist also die mit der Sinnfestlegung von 'ist wahr' bedingte Festlegung der Exten-

[171] Das ist die Lesart, welche McGee vorschlägt [34], S.395.

[172] Das folgt aus einem von Burgess in [8] bewiesenen Satz, der besagt, daß die Menge aller gültigen Sätze eine vollständige $\Pi^1_2$-Menge ist.

sion (Signifikation) des Ausdrucks ‘ist wahr’ in allen möglichen Welten nicht mit der von der (RTW) gegeben Signifikationsfestlegung in allen möglichen Welten kompatibel, dann wird (RTW) gemäß die Definition von S nicht korrekt den Sinn von ‘ist wahr’ wiedergeben. In diesem Sinne sagt (RTW) etwas zum Sinn von ‘ist wahr’. Aber Belnap und Gupta beanspruchen mit (RTW) nicht, den Sinn von ‘ist wahr’ anzugeben. Auch sagt die (RTW) nichts darüber aus, ob es möglich ist, mit einer Definition den Sinn von ‘ist wahr’ derart festzulegen, daß jemand, der nicht über den Begriff der Wahrheit verfügt, mit Hilfe dieser Definition den Begriff der Wahrheit erwerben kann.[173] Ob der Wahrheitsbegriff in diesem Sinne psychologisch primitiv ist, d.h. nicht durch eine Definition erworben werden kann, wird von (RTW) nicht entschieden. Die Zirkularität des Wahrheitsbegriffs in dem Sinne, wie es Belanp und Gupta erläutern, schließt aber nicht von vornherein die Möglichkeit aus, daß natürlichsprachliche Wahrheitsbegriff nicht psychologisch primitiv ist. Daß Gupta und Belnap die Zirkularität des Wahrheitsbegriffs in dem aus der (RTD) folgenden Sinne nachweisen, darf nicht verglichen werden mit dem Nachweisversuch Freges, daß jede Definition des Wahrheitsprädikats zu einer Zirkularität führe und daher der Wahrheitsbegriff überhaupt nicht definierbar sei. Belnap und Gupta schlagen mit ihrer (RTW) also nicht in dieselbe Kerbe wie Frege. Mit der Entscheidung, an die Existenz zirkulärer Begriffe zu glauben, verpflichtet man sich nicht auf die These, es gebe keine psychologisch primitiven Begriffe.

### 10.2.3 Erhaltung der klassischen Logik

Eine der bemerkenswerten Eigenschaften der dargestellten revisionssemantischen Systeme ist die folgende: Alle Satzschemata aus der um die Defininienda erweiterten Sprache, die von der Form eines im klassischen Sinne allgemeingültigen Schemas sind, sind auch gültig in dem von (RTD) explizierten Sinne. Da sich Belnap und Gupta bei der Entwicklung ihrer Revisionstheorie auf klassische bivalente Sprachen beschränkt haben und in der klassischen Logik das Schema $\ulcorner A \vee \neg A \urcorner$ für sämtliche wohlgeformte Formeln A allgemeingültig ist, gilt das auch in den Systemen $S^*$ und $S^\#$. Damit wäre in Guptas und Belnaps Theorie auch der folgende Satz eine logische Wahrheit:

(1) Der Lügnersatz ist wahr oder es ist nicht der Fall, daß der Lügnersatz wahr ist.[174]

Dieser Satz ist also kategorisch wahr in jedem Modell; wir müssen ihn uneingeschränkt bejahen. Das haben einige Kritiker der Revisionstheorie wie z.B. Stephen Yablo in [49] für intuitiv falsch gehalten: Wir würden doch gerade wegen der Pathologizität des Lügnersatzes nicht den oben angeführten Satz behaupten

[173] Siehe hierzu auch [19], S. 633

[174] Beim Lügnersatz denke ich wieder an einen Satz wie den Satz (*):

(*) Der Satz (*) ist nicht wahr.

wollen. Und ist es nicht so, daß wenn wir nach Sätzen gefragt würden, die das tertium-non-datur verletzten, wir dann auch den Lügnersatz anführen würden? Auch wenn es Belnaps und Guptas bewußt aufgestelltes Ziel war, die klassischen Gesetze zu erhalten, müssen sie sich dieser Kritik stellen. Belnap und Gupta antworten hierauf mit einer Gegenintuition.[175] Wir würden doch sicher den folgenden Satz uneingeschränkt bejahen wollen:

(2) Wenn der Lügnersatz wahr ist, dann ist der Lügnersatz wahr.

Aber aus (2) folgt (1). Wenn wir also (2) uneingeschränkt bejahen wollen, dann auch (1). Ich glaube, daß der erste Punkt völlig korrekt ist. Ob wir nun in (2) irgendein logisches Wenn-dann oder das der Umgangssprache zugrunde legen – wir werden (2) für in jedem Fall richtig oder logisch wahr erklären. Was den zweiten Schritt betrifft wird der Kritiker eher vorsichtig sein; und ob der Nichtphilosoph der Folgerung zustimmen wird, wage ich zu bezweifeln. Er hat vermutlich nie mit solchen kruden Argumenten zu tun gehabt. Nun wird sich der Logiker oder Sprachphilosoph bemühen, den vielleicht vorhanden Intuitionen gerecht zu werden und die Semantik der Junktoren richtig zu fassen. Eine Möglichkeit, so erwägen Belnap und Gupta, könnte sein, zu sagen, daß es in mehrwertigen Logiken, in denen wir die „richtige“ Semantik der deutschen Junktoren ‘und’, ‘oder’, ‘wenn-dann’ erfassen können, durchaus nicht der Fall sein muß, daß das Ableitungsschema

$$\text{Wenn P, dann Q} \quad \vdash \quad \text{Q oder } \neg \text{ P}$$

korrekt ist. Belnaps und Guptas absolut korrekte Antwort hierauf ist, daß in diesem Falle auch gemäß der Revisionstheorie nicht aus (2) (1) folgen wird. Wenn die zugrunde liegende Logik den Satz (1) für nicht logisch wahr erklärt, aber (2) schon, dann wird das auch in der Revisionstheorie der Fall sein.

Andererseits könnte der Kritiker behaupten, daß die Änderung der Semantik der logischen Junktoren von der Zweiwertigkeit zur Mehrwertigkeit auf die Einführung des Wahrheitsbegriffs zurückzuführen ist. Das Wahrheitsprädikat verändert die Semantik der Junktoren. Damit könnt man aber einsehen, daß Satz (1) intuitiv nicht gültig, sondern pathologisch ist. Belnaps und Guptas dritter Einwand hieregegen scheint mir der einschlägigste zu sein.[176] Er besagt, daß dieses Verfahren mit zwei natürlichen Ideen konfligiert: Zum einen die Idee, daß die T-Bikonditionale den Wahrheitsbegriff definieren; zum anderen die Idee, daß Definitionen nicht die Logik der Sprache ändern, für die sie aufgestellt werden. Ich kann nicht sehen, daß an diesem Einwand viel davon abhängen würde, daß es auf jeden Fall die T-Bikonditionale sind, welche den Wahrheitsbegriff definieren. Nehmen wir einfach an, es gebe eine Definition des Wahrheitsprädikats, das wir in eine Sprache L einführen. Eine Definition soll im Prinzip nicht mehr tun

[175] [18], S.262
[176] [18], S.263

als die Bedeutung des Wahrheitsprädikats festlegen. Die Bedeutung der anderen Komponenten der Sprache, ob es nun logische oder nichtlogische Ausdrücke sind, darf durch die Definition nicht tangiert werden. Ansonsten verletzt die Definition das Kriterium (NK).

Hierauf mag der Kritiker antworten, daß sein Ziel nicht eigentlich die Definition allein des Wahrheitsprädikats sei, sondern ein weit ambitionierteres Projekt: nämlich die gleichzeitige Fixierung der Bedeutung des Wahrheitsprädikats und der Bedeutung der einzelnen logischen Ausdrücke. Der Begriff der Wahrheit ist mit der Bedeutung der logischen Ausdrücke derart eng verknüpft, da sei schon die Vorstellung, man könne die Bedeutung der logischen Ausdrücke als gegeben voraussetzen und darauf basierend eine Definition des Wahrheitsprädikats geben, abwegig. Ob es je solch eine Kritik gegen das Projekt der Revisionstheorie gegeben hat, weiß ich nicht.[177] Ich kann mir jedenfalls nicht vorstellen, wie das ambitioniertere Projekt aussehen könnte und was für ein Definitionsbegriff hinter diesem Projekt steckt. Was die innige Verknüpfung des Wahrheitsbegriffs mit den logischen Ausdrücken betrifft, so kann man vermutlich einsehen, daß für ein Verständnis der logischen Ausdrücke in irgendeiner Weise das Verständnis des Wahrheitsprädikats nötig ist. Aber daraus folgt nicht, daß wir nichts Erhellendes über die Bedeutung – im schwachen von Belnap und Gupta zugrunde gelegten Sinne – der Wahrheit sagen können, wenn wir annehmen, die Bedeutung der logischen Ausdrücke sei fest vorgegeben.

Belnap und Gupta geben auch eine psychologische Erklärung dafür, daß man meinen könnte, (1) dürfe nicht als kategorisch wahr in allen möglichen Welten ausgewiesen werden. Man könnte nämlich meinen, Satz (1) besage, daß entweder der Lügnersatz oder seine Negation kategorisch wahr ist. De facto besagt aber (1) nichts dergleichen, und tatsächlich sind der Revisionstheorie gemäß weder der Lügnersatz noch dessen Negation kategorisch. Die Übertragung der Überlegung, daß aus der Wahrheit eines Satzes der Form $\ulcorner A \vee B \urcorner$ folgen muß, daß A wahr ist oder daß B wahr ist, auf den Begriff des kategorischen ist schlichtweg nicht korrekt. Aus der kategorischen Wahrheit eines Satzes der Form $\ulcorner A \vee B \urcorner$ folgt eben nicht, daß A kategorisch wahr ist oder B kategorisch wahr ist. Was hiermit bestätigt wird ist, daß der Begriff des Kategorischen genuin verschieden ist von dem der Wahrheit. Man kann nicht behaupten, daß das, was Belnap und Gupta kategorisch wahr nennen, eigentlich das sei, was wir in der Umgangssprache vorher einfach nur wahr genannt haben.

[177] Vielleicht könnte ein Bedeutungsholist dergleichen erwägen?

## 10.3 Weitere Einwände gegen (RTW) und (RTD)

### 10.3.1 Materiale Adäquatheit, $\omega$-Inkonsistenz, Tarskische Wahrheitsregeln

Einige einschlägige Einwände, die nicht den ursprünglichen Gedanken der Revisionstheorie betreffen, nämlich, daß man um kategorische Information aus zirkulären Definitionen zu gewinnen, sich Revisionsregeln anzuschauen hat, sondern vielmehr auf die speziellen Semantiken – und hier insbesondere die Limesregel – abzielen, findet man z.B. bei Yaqub in seinem Buch [50], S. 69-98. Da ich diese Einwände nicht besprechen werde, nur eine kurze Anmerkung: Yaqubs Einwände laufen darauf hinaus, daß die Semantiken angewandt auf das Wahrheitsprädikat einige unintuitive Einordnung von Sätzen in bestimmte Kategorien liefert. Die gröbste vernünftige Einteilung, die sich aus der Revisionssemantik ergibt, ist die Einteilung der Sätze in solche, die kategorisch wahr sind bzw. kategorsisch falsch sind bzw. paradox sind bzw. pathologisch aber nicht paradox sind. Neben dieser groben Einteilung gibt es andere feinere Einteilungen, bei denen sich noch gewisse Intuitionen mobilisieren lassen, um entscheiden zu können, ob ein Satz in eine Kategorie dieser feinen Einteilung gehört oder nicht. Yaqub zeigt nun, daß es Sätze gibt, die gemäß der von Belnap und Gupta vorgeschlagenen Semantiken in eine bestimmte Kategorie der von ihm vorgeschlagenen feineren Kategorierung fallen, obwohl man intuititiv sagen würde, diese Sätze gehörten nicht in diese Kategorie. Will man weiterhin an der Idee festhalten, daß die T-Bikonditionale das Wahrheitsprädikat material adäquat definieren, dann muß man mit Yaqub diese Semantiken verwerfen. Yaqub gibt schließlich selbst eine revisionstheoretische Semantik an, mit der sich die von ihm gezeigten Probleme beheben lassen.

Die beiden Systeme $S^*$ und $S^\#$ bilden eines der drei semantischen Systeme, auf denen Belnap und Gupta letztendlich drei Wahrheitstheorien $T^\#$, $T^*$ und $T^C$ entwickeln. Das dritte System ist durch McGees Arbeit über maximal konsistente Mengen inspiriert, welches ich hier übergehen werde. Für die folgende Diskussion sei definiert

$$V^*_M = \{A : A \text{ ist ein Satz, der gültig in } S^* \text{ in M ist}\}$$

$$V^\#_M = \{A : A \text{ ist ein Satz, der gültig in } S^\# \text{ in M ist }\}$$

Eines der Forderungen an eine Wahrheitstheorie ist, daß sie die von Tarski aufgestellten Regeln dafür, wie die Wahrheit eines komplexen Satzes mit der Wahrheit seiner „Teile“ zusammenzuhängen hat, wahr macht. Z.B. sollten die Wahrheitstheorien von Belnap und Gupta als korrekt herausstellen, daß ein Satz, dessen Hauptoperator ein ‘und’ ist, genau dann wahr ist, wenn die einzelnen Konjunkte wahr sind.[178] In einer hinreichend ausdrucksstarken Sprache kann man sicher-

[178]Diese Regeln wurden bei der rekursiven Definition von ‘wahr in L3’ in meinem Abschnitt zu Tarskis Wahrheitstheorie angewandt.

stellen, daß es Formeln $Neg(x,y)$, $AQ(x,y)$, $In(x,y,z)$ und $Konj(x,y,z)$ gibt derart, daß $Neg(x,y)$ „x ist ein Satz , der die Negation von y ist“ ausdrückt und $AQ(x,y)$ „x ist ein Satz, der die Allquantifikation von y ist“ ausdrückt und $In(x,y,z)$ „y ist eine Formel, die nicht mehr als eine freie Variable hat, und x erhält man aus y, indem man in ihm die freien Variablen durch den Standardnamen für z ersetzt“ ausdrückt und $Konj(x,y,z)$ „x ist die Konjunktion der Sätze y und z“ ausdrückt. Die Tarskischen Wahrheitsregeln lassen sich in einer solchen Sprache dann folgendermaßen formulieren:

$$(T\neg)\ \ Neg(x,y) \ \rightarrow\ [T(x) \leftrightarrow \neg T(y)]$$

$$(T\&)\ \ Konj(x,y,z) \ \rightarrow\ [T(x) \leftrightarrow (T(y)\ \&\ T(z))]$$

$$(T\forall)\ \ AQ(x,y) \ \rightarrow\ [T(x) \leftrightarrow (\forall u,v)(In(u,y,v) \leftrightarrow T(u))]$$

Wollen also die Theorien $T^{\#}$ und $T^{*}$ ernst genommen werden, dann muß sich herausstellen, daß die universellen Abschlüsse aller dieser Regeln sich in $V^{\#}$ bzw. $V^{*}$ befinden. Die Schwäche von $S^{*}$ bringt es mit sich, daß keines der Regeln in $V^{*}$ zu finden sind. In $V^{\#}$ allerdings sind sämtliche Regeln zu finden. Allein wie wir schon bei der Besprechung des Systems $S^{\#}$ festgestellt hatten, ist $V^{\#}$ $\omega$-inkonsistent. Das folgt aus McGees Theorem ([18], S.225, 6C.8), welches grob gesprochen besagt, daß jede Menge von Sätzen, die neben anderen Bedingungen noch die erfüllt, daß es sämtliche Wahrheitsregeln enthält, $\omega$-inkonsistent ist. Nun gilt es, sich für die eine und wider die andere Intuition zu entscheiden. Belnap und Gupta würden behaupten, daß diese Entscheidung natürlich nicht nur ihre Wahrheitstheorie betrifft, sondern sämtliche Wahrheitstheorien. Die Allgemeinheit, in der McGees Theorem formuliert ist, könnte diese Überlegung bestätigen. Daß es hier gegen Einwände geben wird, ist verständlich. Ich begnüge mich mit einem Hinweis auf die Literatur: Z.B. behauptet Koons in [23] S. 624-625, daß in der Theorie von Burge[179] die aus McGees Theorem folgende $\omega$-Inkonsistenz nur eine scheinbare ist. Eine Wahrheitstheorie, die dem durch McGees Theorem gestellten Dilemma entkommen kann, wäre bei gleicher Erklärungskraft allen anderen Wahrheitstheorien vorzuziehen.

### 10.3.2 Induktive und implizite Definitionen

Der folgende Einwand betrifft nur die (RTD). McGee führt in [34] Beipiele von zirkulären Definitionen eines einstelligen Prädikats F an, die klassisch interpretiert eine andere Extension für F ergeben als revisionstheoretisch interpretiert. Betrachten wir hierzu die folgende Definition von F:

$$Fx \quad =_{Df} \quad (x = 0\ \&\ F1) \vee (x = 1\ \&\ F0)$$

[179] Diese geht davon aus, daß die natürlichen Sprachen hierarchisch struktiert sind – mit einem Wahrheitsprädikate in jeder Sprachebene.

Nach der induktiven Lesart ist die Extension von F (kurz Ex(F)) gleich der leeren Menge $\emptyset$. Denn nach dieser Lesart ist Ex(F) der kleinste Fixpunkt des Revisionsoperators $\delta$. Da $\delta$ in diesem Fall monoton ist, d.h. die Bedingung '$(\forall X)(\forall Y)(X \subset Y \rightarrow \delta(X) \subset \delta(Y)$' erfüllt, existiert nach einem Theorem der Fixpunkttheorie solch ein Fixpunkt.
Nach der Revisionstheorie aber ist Ex(F) verschiedenen von der leeren Menge. Denn für die Anfangshypothese $X = \{0\}$ oszilliert $\delta^n(X)$ und gibt nacheinander die Werte $\{0\}, \{1\}, \{0\}, \{1\}, ...$ aus.

Auch für implizite Definitionen ergeben sich Unterschiede. Dazu führt McGee das folgende Beispiel an:

$$Hx \quad =_{Df} \quad \neg H0 \,\&\, (x = 1 \vee (x = 0 \,\&\, \neg H1))$$

Nach der impliziten Lesart ist Ex(H) = $\{1\}$. Denn unter der impliziten Lesart wird dem Definiendum die einzige mögliche Extension zugeordnet, so daß das entsprechende (Bi)Konditional wahr wird. Anders gesagt: Hat der Revisionsoperator $\delta$ lediglich einen einzigen Fixpunkt, so wird dieser unter der impliziten Lesart dem Definiendum zugeordnet.
Nach der revisionstheoretischen Lesart aber ist *nicht* Ex(H) =$\{1\}$. Denn mit der Anfangshypothese $X = \emptyset$ ergibt sich eine Revisionsfolge $\emptyset, \{0, 1\}, \emptyset, \{0, 1\}, ...$ - und diese stabilisiert sich nicht auf $\{1\}$.

Guptas Antwort[180] auf diese Kritik McGees besteht in zweierlei Dingen: Zum einen führt Gupta allgemeine Gründe dagegen an, daß man die in die Kritik eingegangenen Forderungen an eine Definitionstheorie zu akzeptieren hat. (Siehe unten die Forderungen (i) und (ii)). Zum anderen aber skizziert er eine modifizierte Version der Revisionstheorie, die konform mit der üblichen Praxis induktiver und impliziter Definitionen ist.

Die beiden grundsätzlichen Forderungen an eine Definitionstheorie, die nach Gupta Eingang in die Kritik gefunden haben, lauten:

(i) Falls die Revisionsregel $\delta$ einen einzigen Fixpunkt hat, dann sollte das Definiendum (der zu $\delta$ gehörigen Definition) über diesen Fixpunkt von $\delta$ interpretiert werden (*Implizite-Definition-Forderung*).

(ii) Falls $\delta$ monoton ist, dann sollte das Definiendum über den kleinsten Fixpunkt von $\delta$ interpretiert werden (*Induktive-Definition-Forderung*).

Diese beiden Forderungen könnten nicht aufrecht erhalten werden, da sie das Desiderat der Nicht-Kreativität nicht erfüllten, d.h. würde man die Forderung (i) und (ii) befolgen, dann würden sich eventuell bei Einführung einer Definition in eine Theorie der semantische Status einiger Sätze aus der Theorie, die das definierte Zeichen nicht enthalten, durch Einführung der Definition ändern. Genauer lautet das Nicht-Kreativitäts-Desiderat wie folgt:[181]

[180] [17], S.423-430
[181] Das folgende ist ein Spezialfall des allgemein formulierten Kriteriums (NK).

> **Nicht-Kreativitäts-Desiderat** Seien die Definitionen $D_1$ zu einer Sprache $L$ hinzugefügt worden und sei das Resultat dieser Erweiterung der Sprache mit $L + D_1$ benannt. Seien weiter die Definitionen $D_2$ zu $L + D_1$ hinzugefügt worden, so daß man die Sprache $L + D_1 + D_2$ erhält. Dann sollte der semantische Status von Ausdrücken aus $L + D_1$ derselbe bleiben in $L + D_1 + D_2$.

Gupta führt an einem Beispiel vor, daß mit (i) und (ii) die Nicht-Kreativität verletzt ist. Das Beispiel lautet folgendermaßen: $D_1$ sei die Definition:

$Jx \quad =_{Df} \quad Jx$

Und $D_2$ sei die Definition:

$Kx \quad =_{Df} \quad (Jx \vee \neg Kx)$

Nach der Induktive-Definitionen-Forderung muß in der Sprache $L + D_1$ die Interpretation die leere Menge $\emptyset$ sein, denn einerseits ist die leere Menge Fixpunkt des Revisionsoperators, also $\delta(\emptyset) = \emptyset$, und andererseits ist sie der kleinste Fixpunkt, da sogar für jede Menge $X$ (und nicht nur Fixpunkte von $\delta$) gilt: $\emptyset \subset X$. Die Implizite-Definitionen-Forderungen allerdings fordert in der Sprache $L + D_1 + D_2$, daß die Interpretation von $J$ der gesamte Wertebereich, nennen wir ihn $W$, zu sein hat. Denn nur für diese Interpretation von $J$ haben die beiden Definitionen einen Fixpunkt. Man beachte, daß $\delta$ nun nicht bloß Teilmengen des Wertebereichs W als Input aufnimmt, sondern Paare von Teilmengen W: Die eine Teilmenge des Paares ist die für das Prädikat $J$ angenommenene Extension, die andere die für das Prädikat $K$ angenommenen Extension. Der Output ist dann wieder eine Paar von Teilmengen. Der einzige Fixpunkt für den so ins rechte Licht gerückten Revisionsoperator $\delta$ ist das Paar $(Ex(J), Ex(K)) = (W, \emptyset)$. Durch die Einführung von $D_2$ ändert sich somit der semantische Status einiger Sätze, wie z.B. der des Satzes '$Ja$' (mit einem denotierenden Namenbuchstaben 'a'; keine Free Logic). Vor der Einführung ist dieser Satz falsch, danach wahr. An einem weiteren ergänzenden Beispiel versucht Gupta nahezulegen, daß hierfür nur die erste Forderung, also die Implizite-Definition-Forderung, verantwortlich ist.[182] Die Ablehnung dieser Forderung, für die Gupta plädiert, bedeute aber nicht die Verwerfung der generellen Praxis impliziten Definierens, sondern lediglich die Ablehnung einer falschen Methode, die generelle Theorie auf implizite Definitionen anzuwenden. Gupta gibt dann an, wie die wahren Definientia für die Definitionen aus seinem ersten Beispiel auszusähen hätten: Die richtige Formulierung greift auf die Logik

[182]Es ist das folgende Beispiel: $D_1$ ist wieder wie im ersten Beispiel. $D_3$ besteht aus der Definition: $Kx \quad =_{Df} \quad (\neg Jx \vee \neg Kx)$ Die Implizite-Definitionen-Forderungen verlangt, daß die Interpretation von $J$ die leere Menge ist. Das bedeutet, daß mit der Hinzufügung einer Definition zu $L + D_1$ der semantische Status von Sätzen aus $L + D_1$ nach Belieben bestimmt werden kann. Mit der Hinzufügung einer Definition erhält $J$ als Extension den gesamten Wertebereich, mit einer anderen die leere Menge.

zweiter Stufe zurück, d.h. für die Formulierung der wahren Definientia ist Quantifikation über Eigenschaften nötig. Besagte Formulierungen erhält man nach dem bereits bekannten Verfahren. Man ersetzt die Prädikatbuchstaben $J$ und $K$ jeweils durch Prädikatvariablen Y und Z, das Definitionszeichen $=_{Df}$ durch das materiale Bikonditional $\leftrightarrow$. Sei $B(Y, Z)$ die Formel

$(\forall x)((Yx \leftrightarrow Yx)\ \&\ (Zx \leftrightarrow Yx \vee \neg Zx))$

Damit lauten die wahren Definientia für $Jx$ und $Kx$ jeweils:

$((\exists!Y, Z)(B(Y, Z)\ \&\ (\forall Y, Z)(B(Y, Z)\ \rightarrow\ Yx)$

$((\exists!Y, Z)(B(Y, Z)\ \&\ (\forall Y, Z)(B(Y, Z)\ \rightarrow\ Zx)$

Diese Definitionen geben nun die zusätzliche Information bei impliziten Definitionen exakt wieder. Hier liegt keine Zirkularität mehr vor, und somit kann eine allgemeine Definitionstheorie wie z.B. eben die Theorie zirkulärer Definitionen wie erwartet arbeiten. Implizite und ähnlich auch induktive Definitionen werden also in der Theorie zirkulärer Definitionen nicht verworfen, sondern – richtig formuliert – subsumiert.

Im zweiten Teil seiner Anwort stellt Gupta eine von $S^*$ und $S^\#$ verschiedene Revisionssemantik $S^*_{FV}$ vor, das sich von $S^*$ ein wenig in der Behandlung der Limespunkte – worin auch sonst? – unterscheidet. Diese Semantik erfüllt eine andere Forderung, die nur ein wenig von der Induktive-Definition-Forderung abweicht, nämlich die Positive-Definition-Forderung, welche folgendes besagt: Für jede Menge D von Definitionen, die alle positive Vorkommnisse ihrer Definienda haben, und die zugehörige Revisionsregel $\delta$ gilt: Jede Revisionsfolge muß sich auf einen Fixpunkt stabilisieren. Dieses neue semantische System $S^*_{FV}$ ergibt zumindest für McGees zweites Beispiel dasselbe Ergebnis.

Die (RTD) widerspricht also auf keinen Fall der gängigen Praxis des Definierens.

# Literatur

[1] Antonelli, A. „Non-well-founded sets via revision rules“, *Notre Dame Journal of Formal Logic* 23 (1994), 633-79.

[2] Antonelli, A. „The complexity of revision“, *Notre Dame Journal of Formal Logic* 35 (1994), 67-72.

[3] Aristoteles: *Topik (Organon V)*. (Deutsche Übersetzung von Eugen Rolfes.) 3. Aufl., Hamburg: Felix Meiner Verlag, 1922.

[4] Belnap, N. „Gupta's rule of revision theory of truth“, *Journal of Philosophical Logic* 11 (1982), 103-116.

[5] Belnap, N. „On rigorous definitions“, *Philosophical Studies* 72 (1993), 115-146.

[6] Berka, K. und Kreiser L. (Hrsg.) *Logik-Texte*. 2.Aufl., Berlin: Akademie-Verlag, 1983.

[7] Boolos, G. *The Logic of Provability*. Cambridge: Cambridge University Press, 1993.

[8] Burgess, J. „The truth is never simple“, *Journal of Symbolic Logic* 51 (1986), 663-81.

[9] Carnap, R. *Meaning and Necessity*, 2. Aufl., Chicago, Illinois: University of Chicago Press, 1956.

[10] Dubislav, W. *Die Definition*, 3. Aufl., Leipzig: Felix Meiner Verlag, 1931.

[11] Essler, W.K. *Wissenschaftstheorie I - Definition und Reduktion*. Freiburg/München: Verlag Karl Alber, 1982.

[12] Fetzer, J.H., Shatz, D. und Schlesinger, G. (Hrsg.) *Definitions and Definability: Philosophical Perspectives*. Netherlands: Kluwer Academic Publishers, 1991.

[13] Fleischer, M. (Hrsg.) *Philosophen des 20. Jahrhunderts*, 4. Aufl., Darmstadt: Wissenschaftliche Buchgesellschaft, 1995.

[14] Gaifmann, H. „Operational pointer semantics: Solution to self-referential puzzles I“, in [45].

[15] Gupta, A. „Truth an paradox“, *Journal of Philosophical Logic* 11 (1982), 1-60. (Wiederabgedruckt in [32], 175-235.)

[16] Gupta, A. „Remarks on definitions and the concept of truth“, *Proceedings of the Aristotelian Society* 89 (1988-89), 227-246.

[17] Gupta, A. „Definition and revision: A response to McGee and Martin“, in [46], 419-443.

[18] Gupta, A. und Belnap, N. *The Revision Theory of Truth*. Cambridge: MIT Press, 1993.

[19] Gupta, A. und Belnap, N. „Reply to book review“, *Notre Dame Journal of Formal Logic* 35 (1994), 632-636.

[20] Herzberger, H. „Notes on naive semantics“, *Journal of Philosophical Logic* 11 (1992), 61-102. (Wiederabgedruckt in [32], 133-174).

[21] Herzberger, H. „Naive semantics and the liar paradox“, *Journal of Philosophy* 19 (1992), 479-97.

[22] Humberstone, I.L. „Two types of circularity“, *Philosophy and Phenomenological Research* 57 (1997), 249-280.

[23] Koons, R.C. „Book review“, *Notre Dame Journal of Formal Logic* 35 (1994), 606-631.

[24] Kremer, P. „The Gupta-Belnap systems $S^{\#}$ and $S^{*}$ are not axiomatisable“, *Notre Dame Journal of formal Logic* 34 (1993), 583-96.

[25] Kripke, S. „Outline of a theory of truth“, *Journal of Philosophy* 72 (1975), 690-716. (Wiederabgedruckt in [32])

[26] Künne, W. „Wahrheit“, in [30], 116-171.

[27] Künne, W. „George Edward Moore - Was ist Begriffsanalyse?“, in [13], 27-40.

[28] Langford, C.H. „The notion of analysis in Moores philosophy“, in [38], 323.

[29] Leonard, H. *Principles of Reasoning: Introduction to Logic, Methodology, and the Theory of Signs*. London: Dover, 1967.

[30] Martens, E. und Schnädelbach, H.(Hrsg.) *Philosophie – Ein Grundkurs*, Bd.1, 2. Aufl., Hamburg: Rowohlt, 1994.

[31] Martin, D.A. „Revision and it rivals“, in [46], 407-448.

[32] Martin, R. L. (Hrsg.) *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford University Press, 1984.

[33] McGee, V. *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth.* Indianapolis, Ind.: Hacket, 1991.

[34] McGee, V. „Revisions“, in [46], 387-406.

[35] Moschovakis, Y.N. *Notes on Set Theory.* Heidelberg: Springer-Verlag, 1994.

[36] Orilia, F. „Meaning and circular definitions“. *Journal of Philosophical Logic* 29 (2000), 155-169.

[37] Rantala, V. „Definitions and definability“, in [12], 135-159.

[38] Schilpp, P.A. (Hrsg.) *The Philosophy of G.E.Moore.* New York, 1942.

[39] Skirbekk, G. (Hrsg.) *Wahrheitstheorien*, 7. Aufl., Frankfurt am Main: Suhrkamp, 1996.

[40] Skyrms, B. „Intensional aspects of semantical self-reference“, in [32].

[41] Suppes, P. *Introduction to Logic.* Mineola, N.Y.: Dover Publications, Inc., 1999.

[42] Tarski, A. „Der Wahrheitsbegriff in den formalisierten Sprachen“, in [6], 443-546.

[43] Tarski, A. „The semantic conception of truth“, *Philosophy and Phenomenological Research* 4 (1943/44), 341-75.

[44] Tichy, Pavel. „On the vicious circle in definitions“, *Studia Logica* 28 (1971), 19-38.

[45] Vardi, M.Y. *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge.* Morgan Kaufmann, 1988.

[46] Villanueva, E. (Hrsg.) *Philosophical Issues*, 8: *Truth.* Atascadero, Calif.: Ridgeview, 1997.

[47] Williams, C.J.F. *Being, Identity and Truth.* Oxford: Oxford University Press, 1992.

[48] Yablo, S. „Definitions, consistent and inconsistent“, *Philosophical Studies* 72 (1993), 147-175.

[49] Yablo, S. „Truth and reflection“, *Journal of Philosophical Logic* 14 (1985), 297-349.

[50] Yaqub, A.M. *The Liar Speaks The Truth.* Oxford: Oxford University Press, 1993.