ANSGAR BECKERMANN

SPRACHVERSTEHENDE MASCHINEN

(*Language Understanding Machines*)

Überlegungen zu John Searle's Thesen zur Künstlichen Intelligenz

ABSTRACT. In this paper the author tries to disentangle some of the problems tied up in John Searle's famous Chinese-room-argument. In a first step to answer the question what it would be for a system to have not only syntax, but also semantics the author gives a brief account of the functioning of the language understanding systems (LUS) so far developed in the framework of AI research thereby making clear that systems like Winograd's SHRDLU are indeed doing little more than mere number crunching. But things would be entirely different, the author argues, if the database of a LUS were built up by the system itself via some perceptual component – at least, if this perceptual component had the capacity to distinguish objects having a certain property F from objects which do not. For in this case the system could store an internal representation of the fact that the object has the property F in its database if and only if the object in fact has that property. And this would be a good basis for calling such a system a *genuine* LUS. But Searle has objected to a very similar account of J. Fodor that nothing could be further from true language understanding. The reason for this complaint seems to be that Searle holds the view that a true LUS must e.g., *know* that the word "hamburgers" refers to hamburgers and that he moreover claims that this knowledge must be *explicit* or that the system must be *aware* of the reference of "hamburgers" to hamburgers. The author argues that this is asking too much. For it seems plausible to say that a system is able to understand e.g., the word "hamburger" even if it has only *implicit* knowledge of the fact that "hamburger" refers to hamburgers in the sense that it has the capacity to tell hamburgers from non hamburgers and the capacity to bring the word "hamburger" together just with objects of the former kind.

1.

John Searle ist häufig so verstanden worden, als habe er in seinem Aufsatz "Minds, Brains, and Programs" und in späteren Arbeiten die These vertreten wollen, keine Maschine – welcher Art auch immer – sei im Wortsinn imstande, Sprache zu verstehen oder andere kognitive Zustände anzunehmen. Doch das ist nicht der Fall. Denn in dem genannten Aufsatz schreibt er ausdrücklich auch:

> I see no reason in principle why we couldn't give a machine the capacity to understand English or Chinese, since in an important sense our bodies with our brains are precisely such machines. (MBP, 422)

Es geht Searle also nicht darum zu behaupten, daß überhaupt keine

Erkenntnis **28** (1988) 65–85.

Maschine Sprache verstehen oder kognitive Zustände haben kann. Und es geht ihm daher auch nicht darum zu behaupten, wir selbst, die wir dies alles können, seien keine Maschinen – ganz im Gegenteil. Seine These ist vielmehr, daß *bestimmte* Maschinen – insbesondere Computer – keine Sprache verstehen können und daß wir selbst daher keine Computer sein können. D.h. genauer kann man Searles These so formulieren:

> SEARLE:
> Keine Maschine, deren Verhalten allein durch formale Veränderungen formal definierter Elemente bestimmt ist, d.h. keine Maschine, deren Verhalten als die Instantiierung eines Computerprogramms definiert ist, ist im Wortsinn imstande, Sprache zu verstehen oder andere kognitive Zustände zu haben.

Searles These ist also eingeschränkter, als viele angenommen haben. Aber sie ist deshalb nicht weniger brisant. Denn sie stellt – und das ist ja auch Searles Absicht – das gesamte Programm der Künstlichen Intelligenz-Forschung infrage. Zumindest gilt das, wenn man dieses Programm im Sinne der – wie Searle sagt – starken KI versteht, d.h. wenn man mit diesem Programm die These verbindet:

> STARKE KI:
> (a) Es gibt Maschinen, deren Verhalten als die Instantiierung eines Computerprogramms definiert ist, die im Wortsinn Sprache verstehen oder andere kognitive Zustände annehmen können.
> (b) Die Programme, die das Verhalten dieser Maschinen bestimmen, *erklären* auch die menschliche Fähigkeit, Sprache zu verstehen, bzw. erklären auch, was es für Menschen bedeutet, bestimmte kognitive Zustände anzunehmen.

2.

Das Hauptargument, das Searle für seine These anführt, besteht in einem Gedankenexperiment,[1] das unter dem Namen *Chinese-Room-Experiment* bekannt geworden ist. Mit diesem Gedankenexperiment möchte er zeigen, daß Situationen vorstellbar sind, in denen ein Mensch z.B. in einer Dialogsituation genau das leistet, was ein Com-

puter leisten kann, ohne dabei jedoch auch nur das Geringste von der Sprache zu verstehen, in der "der Dialog geführt wird". Im einzelnen orientiert sich Searle bei seinem Gedankenexperiment an einem Programm von Roger Schank,[2] bei dem es im Kern darum geht, daß ein Computer, dem zuvor der Normalablauf bestimmter Situationen (z.B. eines Restaurantbesuchs) in Form von "scripts" eingegeben wurde, Geschichten über konkrete einzelne Situationen dieser Art verstehen und Fragen im Hinblick auf diese Geschichten korrekt beantworten soll. Searle stellt uns nun folgende Parallelsituation vor:

> . . . stellen Sie sich vor, Sie wären in ein Zimmer eingesperrt, in dem mehrere Körbe mit chinesischen Symbolen stehen. Und stellen Sie sich vor, daß Sie . . . kein Wort Chinesisch verstehen, daß Ihnen allerdings ein auf Deutsch abgefaßtes Regelwerk für die Handhabung dieser chinesischen Symbole gegeben worden wäre. Die Regeln geben rein formal – nur mit Rückgriff auf die Syntax und nicht auf die Semantik der Symbole – an, was mit den Symbolen gemacht werden soll. Eine solche Regel mag lauten: 'Nimm ein Kritzel-Kratzel-Zeichen aus Korb 1 und lege es neben ein Schnörkel-Schnarkel-Zeichen aus Korb 2'. Nehmen wir nun an, daß irgendwelche anderen chinesischen Symbole in das Zimmer gereicht werden, und daß Ihnen noch zusätzliche Regeln dafür gegeben werden, welche chinesischen Symbole jeweils aus dem Zimmer herauszureichen sind. Die hereingereichten Symbole werden von den Leuten draußen 'Fragen' genannt, und die Symbole, die Sie dann aus dem Zimmer herausreichen, 'Antworten' – aber dies geschieht ohne Ihr Wissen. Nehmen wir außerdem an, daß die Programme so trefflich und Ihre Ausführung so brav sind, daß Ihre Antworten sich schon bald nicht mehr von denen eines chinesischen Muttersprachlers unterscheiden lassen. (GHW, 31)

Der Punkt dieser Geschichte liegt auf der Hand. Der in das Zimmer Eingesperrte wird als Reaktion auf die ihm in schriftlicher Form gestellten chinesischen Fragen, die ihm selbst aber nur als Folgen von durch ihre graphische Form charakterisierten Symbolen erscheinen, Folgen von solchen Symbolen zurückgeben, die von den Fragestellern als vernünftige Antworten auf die von ihnen gestellten Fragen aufgefaßt werden können. Doch das bedeutet nicht, daß der Eingesperrte selbst deshalb verstehen müßte, was die Fragen und was seine "Antworten" bedeuten. Der ganze Witz der geschilderten Situation liegt ja gerade darin, daß er voraussetzungsgemäß die genannten Leistungen erbringt, ohne auch nur ein Wort Chinesisch zu verstehen. Und daher, schließt Searle, verstehen auch Computer nichts von den Sprachen, in denen sie "Dialoge" führen können. Denn, so lautet sein Argument:

> Wenn . . . die Ausführung eines passenden Computerprogramms ***in Ihrem Fall*** nicht ausreicht, um Chinesisch zu verstehen, dann reicht das auch bei ***keinem anderen digitalen Computer*** aus . . . Wenn Sie kein Chinesisch verstehen, dann könnte auch kein

> anderer Computer Chinesisch verstehen; denn kraft seiner Ausführung eines Programms hat kein digitaler Computer irgendetwas, das Sie nicht haben. Der Computer hat – genau wie Sie – nichts außer einem formalen Programm für die Handhabung uninterpretierter chinesischer Symbole. (GHW, 31f.)

Obwohl diese Argumentation Searles auf sehr heftige Kritik gestoßen ist, muß man doch sagen, daß Searle mit ihr einen wichtigen Punkt getroffen hat. Denn im Kern läuft diese Argumentation darauf hinaus, daß er noch einmal mit allem Nachdruck und sehr anschaulich auf die Tatsache hinweist, daß Computer *ihrer Natur nach* weder Rechenmaschinen noch informationsverarbeitende Maschinen, sondern einfach symbolmanipulierende Maschinen sind. Oder, um auch noch den Gebrauch des Wortes "Symbol", das ja ebenfalls auf bedeutungstragende Entitäten verweist, zu vermeiden: daß Computer ihrer Natur nach *musterverarbeitende* Maschinen sind. Denn in Computern geschieht letzten Endes nichts anderes, als daß Muster von 8, 16 oder mehr Bits, die sich in verschiedenen Registern oder an verschiedenen Speicherplätzen befinden, nach bestimmten Regeln hin- und hergeschoben oder verändert werden. Dabei orientieren sich diese Regeln selbst nur an der äußeren Gestalt der einzelnen Bitmuster und nicht etwa an einem möglichen Bedeutungsinhalt, den diese Muster haben könnten.

Wenn man diese Tatsache mit der in der Sprachwissenschaft und Sprachphilosophie geläufigen Unterscheidung von Syntax und Semantik in Zusammenhang bringt, scheint daher klar, daß Computer eingegebene sprachliche Äußerungen immer nur syntaktisch, nicht aber semantisch behandeln können. Denn während sich die Syntax herkömmlichen Definitionen zufolge nur auf die äußere Form, d.h. insbesondere auf die Wohlgeformtheit sprachlicher Ausdrücke bezieht, geht es bei der Semantik um die Bedeutung dieser Ausdrücke. Und diese Bedeutung kann, wie viele meinen, nicht auf die äußere Form sprachlicher Ausdrücke reduziert werden. Die Erfassung von Bedeutungen scheint damit etwas zu sein, was grundsätzlich außerhalb der Reichweite von Maschinen liegt, die im Prinzip nur durch ihre äußere Form charakterisierte Muster manipulieren können.

Diese Argumentation ist jedoch nicht ganz so unproblematisch, wie sie auf den ersten Blick aussehen mag. Ihr Hauptproblem liegt meiner Meinung nach darin, daß sie auf einem ziemlich unklaren Semantikkonzept beruht, in dem Semantik letztlich nur negativ in Abgrenzung gegen die Syntax definiert wird. Allerdings ist es nicht leicht, dieses

Semantikkonzept durch ein genaueres und besser durchformuliertes zu ersetzen. Deshalb möchte ich hier einen anderen Weg gehen und zunächst untersuchen, wie die von Searle kritisierten natürlich-sprachlichen Computersysteme im einzelnen arbeiten, um dann möglicherweise auf diesem Weg einer Antwort auf die Frage, ob solche Systeme tatsächlich Sprache verstehen können oder nicht, etwas näher zu kommen.

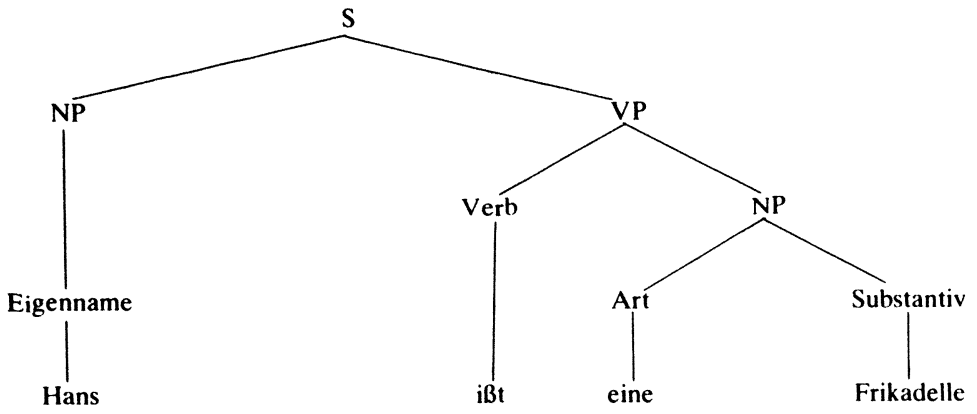
3.

In sehr groben Zügen kann man die Arbeitsweise neuerer natürlich-sprachlicher Systeme (NSS) so zusammenfassen.

In einem ersten Schritt wird jeder eingegebene Satz einer morphologisch-lexikalischen Analyse unterzogen. Als Ergebnis dieser Analyse werden jedem Wort dieses Satzes seine lexikalische Kategorie (Substantiv, Verb, Präposition, Determinator usw.) sowie, falls sinnvoll, einige syntaktische Merkmale (Singular, Plural, definit, indefinit, usw.) zugeordnet. Als Ergebnis ergibt sich etwa für den Eingabesatz "Hans ißt eine Frikadelle" die folgende Aufstellung.[3]

Wort	Lexikalische Kategorie	Merkmale
Hans	Eigenname	
ißt	Verb	3. Pers., Sing., Präsens
eine	Artikel	indefinit, Sing., weibl.
Frikadelle	Substantiv	Sing., weibl.

Im zweiten Schritt wird jeder Eingabesatz unter Berücksichtigung der Ergebnisse der morphologisch-lexikalischen Analyse mit Hilfe einer Parsing-Komponente einer syntaktischen Analyse unterzogen. Dabei wird zunächst geprüft, ob der Satz syntaktisch korrekt ist. Falls nicht, wird er zurückgewiesen. Falls er jedoch korrekt ist, wird ihm als Ergebnis der syntaktischen Analyse seine syntaktische Struktur zugeordnet, z.B. in der Form eines Strukturbaumes. Für den angegebenen Beispielsatz etwa der im oberen Teil der Abb. 1 gezeigte Strukturbaum. Dieser Strukturbaum kann auch – was für die Verarbeitung durch einen Computer natürlicher ist – in der Form der im unteren Teil der Abb. 1 angeführten Liste dargestellt werden.



(S (NP (Eigenname Hans))
(VP (Verb ißt)
(NP (Art eine) (Substantiv Frikadelle))))

Abb. 1.

In der Praxis kann es darüberhinaus sinnvoll sein, in die durch eine solche Liste gegebene Beschreibung der syntaktischen Struktur eines Satzes noch weitere Informationen mitaufzunehmen und z.B. am Hauptknoten S zu notieren, ob es sich bei dem analysierten Satz um eine Behauptung, einen Fragesatz oder einen Befehl handelt.

Diese wenigen Sätze sollen für die Darstellung der ersten beiden Schritte der Sprachverarbeitung in NSS ausreichen. Für die Ausgangsfrage ist nämlich der dritte Schritt wichtiger. Denn dieser Schritt wird herkömmlich als "semantische Analyse" bezeichnet. In ihm scheint es also tatsächlich um die Bedeutung der eingegebenen Sätze zu gehen. Es lohnt sich daher, genauer zu untersuchen, was bei der semantischen Analyse von Sätzen konkret geschieht. Der unmittelbare Zweck dieser Analyse ist die Überführung des eingegebenen Satzes in eine *interne Repräsentation*, wenn man so will, also die Übersetzung des eingegebenen Satzes in einen entsprechenden Satz einer internen Sprache des Computers. Es gibt inzwischen sehr verschiedene Methoden der internen Repräsentation. Als Beispiel soll hier die von Winograd in seinem System SHRDLU eingesetzte Methode dienen, als semantische Repräsentationssprache Ausdrücke der KI-Programmiersprache PLANNER zu verwenden. Bei dieser Methode ergibt sich etwa für den Eingabesatz

(*) Welche Pyramiden werden von einem Block gestützt?

als semantische Repräsentation der PLANNER-Ausdruck

```
(**)  (FIND  ALL  ?X1  (X1)
             (GOAL  (ISA  ?X1  PYRAMIDE))
             (FIND  1  ?X2  (X2)
                    (GOAL  (ISA  ?X2  BLOCK))
                    (GOAL  (STUETZT  ?X2  ?X1)))).
```

Ich kann hier nicht darauf eingehen, wie der Eingabesatz im einzelnen in diesen PLANNER-Ausdruck überführt wird. Das ist aber auch nicht unbedingt erforderlich. Denn für diesen Zusammenhang reicht es aus zu wissen, daß alle Informationen, die dafür nötig sind, aus der Beschreibung der syntaktischen Struktur dieses Satzes, aus den Lexikoneinträgen für die in diesem Satz vorkommenden Wörter und aus einigen anderen im System vorhandenen "Wissenskomponenten" gewonnen werden können. Die entscheidende Frage ist vielmehr, wozu diese Übersetzung in eine interne Repräsentation dient und was sie mit einem wirklichen Verstehen der Bedeutung des eingegebenen Satzes zu tun hat.

Der Grund für die Übersetzung eingegebener Sätze in interne Repräsentationen liegt zunächst darin, daß das System diese internen Repräsentationen – anders als den eingegebenen Satz – in einer zusätzlichen Auswertungs-Komponente weiter verarbeiten kann. So ist z.B. der PLANNER-Ausdruck (**) selbst eine PLANNER-Funktion, die als Teil eines PLANNER-Programms die internen Namen aller Pyramiden liefert, die tatsächlich von einem Block gestützt werden. Dabei ist allerdings die Existenz einer Datenbasis vorausgesetzt, in der das System sein "Wissen" über die "äußere Welt" speichert. Dies muß vielleicht noch etwas erläutert werden.

Das System SHRDLU von Terry Winograd ist so eingerichtet, daß es – auf Anweisung eines menschlichen Partners – Manipulationen in einer fiktiven Blockwelt ausführt und darüberhinaus von diesem Partner gestellte Fragen über diese Blockwelt und über die eigenen "Handlungen" des Systems beantwortet. Die fiktive Blockwelt selbst besteht aus einer Platte, auf der (zum Teil auch übereinander) einige Blöcke und Pyramiden sowie eine größere Box stehen. Die Abb. 2 zeigt einen möglichen Zustand dieser Blockwelt.

Eine solche Abbildung ist jedoch nur eine visuelle Hilfe für das

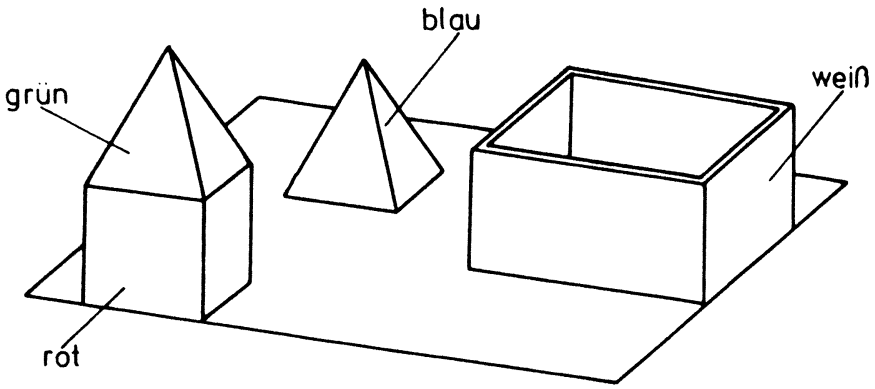


Abb. 2.

menschliche Verständnis. Intern wird der in der Abb. 2 gezeigte Zustand eher durch eine Reihe von Ausdrücken (Listen) repräsentiert sein, z.B. durch die Ausdrücke:

(IST-EIN BLOCK1 BLOCK)
(ORT BLOCK1 (2,1,0))
(FARBE BLOCK1 ROT)
(STUETZT BLOCK1 BLOCK3)
(IST-EIN BLOCK2 PYRAMIDE)
(ORT BLOCK2 (3,7,0))
(FARBE BLOCK2 BLAU)
(IST-EIN BLOCK3 PYRAMIDE)
(ORT BLOCK3 (2,1,2))
(FARBE BLOCK3 GRÜN)
(IST-EIN BLOCK4 BOX)
(ORT BLOCK4 (8,6,0))
(FARBE BLOCK4 WEISS)

Diese Menge von Ausdrücken bildet die Datenbasis, in der das "Wissen" des Systems um den gegenwärtigen Zustand der Blockwelt gespeichert ist. Und auf diese Datenbasis angewandt liefert der PLANNER-Ausdruck (**) z.B. den Wert "BLOCK3" als internen Namen für die grüne Pyramide. Dieser Name kann dann weiter verwendet werden – etwa zur Generierung einer Antwort auf die eingegebene Frage.

In gewisser Weise sieht es also so aus, als käme die Übersetzung des Eingabesatzes (*) in den PLANNER-Ausdruck (**) tatsächlich dem

Verstehen des Satzes (*) gleich. Denn dieser PLANNER-Ausdruck liefert als Teil eines PLANNER-Programms genau die richtige Antwort auf die mit dem Satz (*) gestellte Frage. Beim zweiten Hinsehen erhärtet sich jedoch die Vermutung, daß Searle mit seiner Kritik auch im Hinblick auf die in diesem Abschnitt geschilderten NSS recht hat. Denn in Wirklichkeit gibt es weder eine Blockwelt noch eine grüne Pyramide noch einen roten Block. Und in Wirklichkeit wird auch die grüne Pyramide nicht durch einen Block gestützt. Die anschauliche Darstellung von Blockweltzuständen in Abbildungen wie der Abb. 2 verführt immer wieder dazu, zu vergessen, daß diese Welt nur fiktiv ist. Was es wirklich gibt, sind nur die Datenstrukturen im Computer. Und diese Datenstrukturen ihrerseits sind nichts anderes als an bestimmten Speicherplätzen befindliche Bit-Muster. Wenn ein NSS, das nach den gerade geschilderten Strategien arbeitet, auf die Frage "Welche Pyramiden werden von einem Block gestützt?" antwortet "Die grüne Pyramide wird von einem Block gestützt", dann geschieht also genau das, was Searle mit seinem Chinese-Room-Experiment veranschaulichen wollte. Die Maschine analysiert und verändert eingegebene Zeichenfolgen und Bit-Muster nach Regeln, die sich allein an der physischen Gestalt dieser Muster orientieren. Und am Ende all ihrer Arbeit gibt sie eine Zeichenfolge aus, die sie allein aufgrund der äußeren Gestalt der Zeichen zusammengestellt hat. Es gibt also keinen Grund anzunehmen, die Maschine verstünde tatsächlich, was man ihr eingibt und was sie selbst ausgibt. Denn die Maschine weiß offenbar weder, was eine Pyramide ist, noch, was grün ist, noch, wann etwas von etwas gestützt wird.

4.

Ich denke, die Dinge lägen jedoch ganz anders, wenn sich die Datenbasis des im letzten Abschnitt besprochenen Systems nicht auf eine fiktive Blockwelt bezöge, sondern auf die wirkliche Umgebung des Systems, d.h. wenn das System eine Wahrnehmungskomponente enthielte, die ihrerseits diese Datenbasis als Modell der das System umgebenden Umwelt erst aufbauen würde. Dies ist durchaus keine utopische Idee. Denn nach den bisherigen Ergebnissen der KI-Forschung spricht nichts gegen die Möglichkeit visueller Systeme, die zumindest bei relativ einfachen Szenen in der Lage sind, die zu diesen Szenen gehörenden Objekte zu *identifizieren*, diese Objekte ihrer

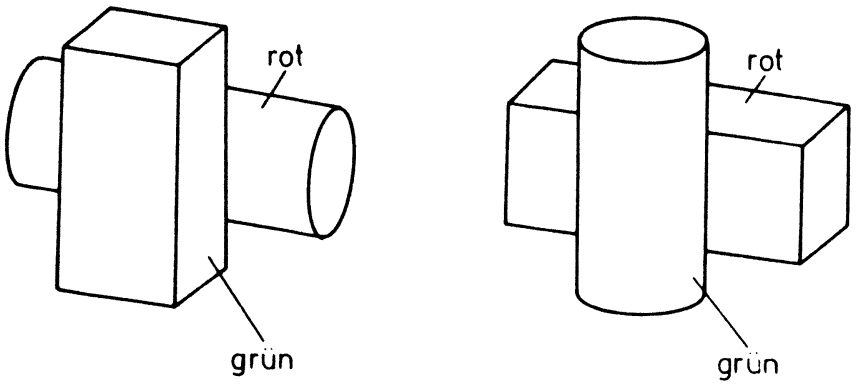


Abb. 3.

geometrischen Gestalt nach zu *klassifizieren* und die zwischen den Objekten bestehenden *räumlichen Beziehungen zu erkennen*.

Ein solches System könnte also, wenn seine Kamera auf die in der Abb. 3 gezeigte Szene gerichtet wäre, erkennen, daß an dieser Szene vier Objekte beteiligt sind, es könnte diesen Objekten interne Namen geben (etwa OBJEKT1, . . . , OBJEKT4), es könnte erkennen, daß es sich bei den Objekten 1 und 4 um Quader handelt und bei den Objekten 2 und 3 um Zylinder, es könnte erkennen, daß der Zylinder 2 liegt, während der Zylinder 3 steht, es könnte erkennen, daß sich das Objekt 1 vor dem Objekt 2 befindet und das Objekt 3 vor dem Objekt 4, es könnte möglicherweise auch die Farbe der Objekte erkennen, d.h. es würde alle diese Sachverhalte in internen "Aussagen" festhalten, z.B. in der folgende Reihe von Ausdrücken:

(IST-EIN OBJEKT1 QUADER)
(FARBE OBJEKT1 GRÜN)
(VOR OBJEKT1 OBJEKT2)
(IST-EIN OBJEKT2 ZYLINDER)
(FARBE OBJEKT2 ROT)
(LIEGT OBJEKT2)
(IST-EIN OBJEKT3 ZYLINDER)
(FARBE OBJEKT3 GRÜN)
(VOR OBJEKT3 OBJEKT4)
(IST-EIN OBJEKT4 QUADER)
(FARBE OBJEKT4 ROT)

Nehmen wir nun an, daß sich ein System der gerade beschriebenen

Art in einer Situation befindet, in der seine Kamera auf die in der Abb. 3 gezeigte Szene gerichtet ist, daß das System daraufhin die gerade beschriebene Datenbasis aufbaut und daß es daher, wenn man ihm die Frage stellt "Welcher Zylinder steht vor einem Quader?" antwortet "Der grüne Zylinder steht vor einem Quader". Kann man dann auch von diesem System noch sagen, es verstünde nicht wirklich, was man es fragt und was es selbst sagt?

Offenbar gehören zum Sprachverstehen sehr verschiedene Aspekte – so z.B. das Verstehen verschiedener Sprechakttypen. Aber eine zentrale These der Theorie des Sprachverstehens ist wohl die Auffassung: Die Bedeutung eines Satzes verstehen, heißt wissen, unter welchen Bedingungen dieser Satz wahr ist. D.h. die Bedeutung eines Satzes verstehen, heißt wissen, welche Wahrheitsbedingungen dieser Satz hat. Wenn man von diesem Grundsatz ausgeht, stellt sich die Frage jedoch so: Kann man von einem System wie dem geschilderten mit Recht sagen, daß es die Wahrheitsbedingungen von Sätzen kennt? Und wenn man auf diese Frage eine Antwort geben will, dann muß man sich darüber klar werden, was es heißt, die Wahrheitsbedingungen eines Satzes zu kennen.

Meiner Meinung nach kann man das Wissen um die Wahrheitsbedingungen von Sätzen jedoch einfach mit einer bestimmten Diskriminierungsfähigkeit identifizieren, nämlich mit der Fähigkeit, Situationen, in denen ein Satz wahr ist, von Situationen zu unterscheiden, in denen das nicht der Fall ist. Somit ist die Frage: Was ist erforderlich dafür, daß ein System S z.B. über die Fähigkeit verfügt, Situationen, in denen der Satz "Der grüne Zylinder steht vor einem Quader" wahr ist, von solchen Situationen zu unterscheiden, in denen das nicht der Fall ist? Soweit ich sehen kann, gehört zum Besitz dieser Fähigkeit unter anderem, daß das System S in dem Sinne über die in diesem Satz vorkommenden Begriffe "Zylinder", "Quader" und "vor" verfügt, daß es Situationen, in denen diese Begriffe zutreffen, von Situationen unterscheiden kann, in denen das nicht der Fall ist. Um Situationen, in denen der Satz wahr ist, identifizieren zu können, muß S also unter anderem in diesem Sinne über den Begriff "vor" verfügen. Ist das für das gerade geschilderte System der Fall?

Die Antwort auf diese Frage hängt davon ab, wie die Wahrnehmungs-Komponente dieses Systems im einzelnen arbeitet. Denn es könnte z.B. sein, daß diese Komponente einen Gegenstand A dann und nur dann als vor einem anderen Gegenstand B stehend

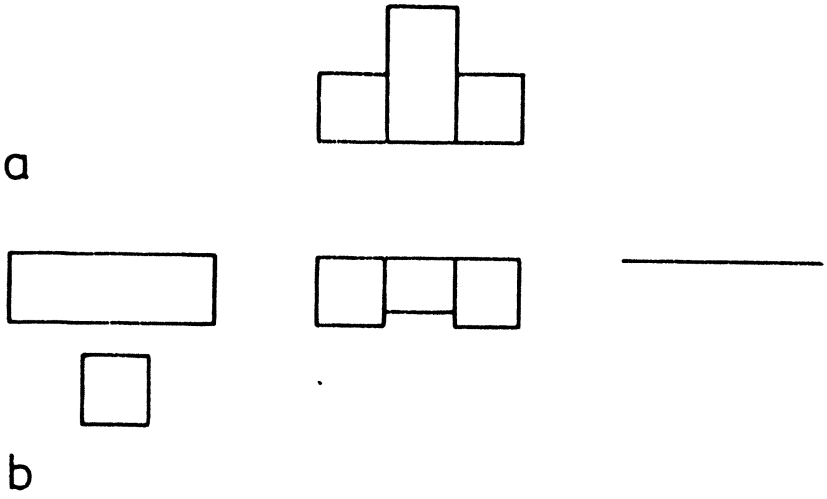


Abb. 4.

klassifiziert, wenn A einen Teil von B verdeckt, bzw. genauer: wenn eine Ansicht der in der Abb. 4a gezeigten Art vorliegt.

In diesem Fall könnte man aber wohl nicht sagen, daß das System tatsächlich über den Begriff "vor" verfügt. Denn Bilder dieser Art können auf sehr verschiedene Weise zustande kommen. Es kann sein, daß A sich tatsächlich vor B befindet, es kann aber auch sein, daß dieses Bild von einem einzigen Objekt stammt mit drei nebeneinander angeordneten Teilen, von denen der mittlere etwas zurückgesetzt ist, und es kann auch sein, daß wir es tatsächlich nur mit einem zweidimensionalen Objekt, d.h. wirklich nur mit einem Bild zu tun haben. Von oben würden die drei geschilderten Situationen in etwa so aussehen, wie es die Abb. 4b zeigt.

Es gibt natürlich Möglichkeiten, diese drei Situationen voneinander zu unterscheiden. Beim binocularen Sehen etwa hilft die Parallaxe bei der Bestimmung von Entfernungen, auch Texturunterschiede können entsprechende Hinweise geben. Besonders verläßlich sind jedoch die Hinweise, die sich daraus ergeben, wie sich die Ansicht einer Szene verändert, wenn sich die Objekte in dieser Szene bewegen oder wenn sich der Beobachter selbst bewegt. In den drei geschilderten Situationen z.B. würde sich die Ansicht für einen Beobachter, der sich im Uhrzeigersinn um die Szene herumbewegt, sehr verschieden entwickeln. In etwa so, wie es in den verschiedenen Bildern der Abb. 5 gezeigt ist.

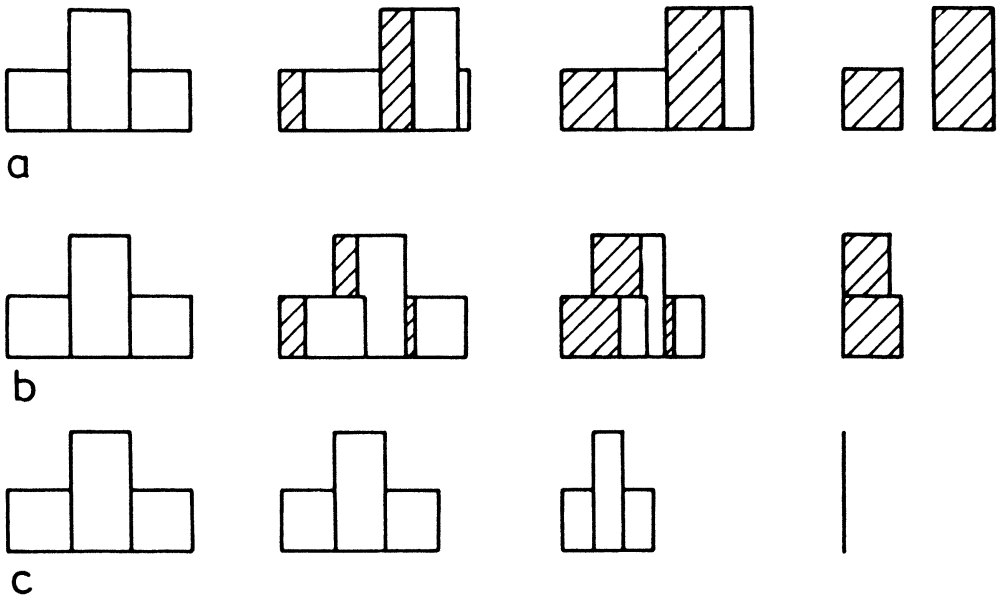


Abb. 5.

Meines Wissens ist es nicht unmöglich, ein visuelles System so einzurichten, daß es in der Lage ist, bei einer kontinuierlichen Veränderung der Ansicht einer Szene, wie sie durch normale Bewegungen hervorgerufen wird, die zu dieser Szene gehörenden Objekte zu fixieren und daher nicht nur statische Bilder, sondern auch Sequenzen aufeinanderfolgender Bilder zur Analyse der betreffenden Szene zu verwenden. Ein solches System könnte dann aber auch so eingerichtet werden, daß es vorläufige Meinungen immer wieder überprüft und gegebenfalls verändert. D.h. es scheint mir nicht unmöglich, ein System so einzurichten, daß es, auch wenn es beim ersten Anblick einer Szene zu der Überzeugung gekommen ist, daß zu dieser Szene zwei Objekte gehören, von denen sich das eine vor dem anderen befindet, diese Überzeugung revidiert, wenn es sich selbst um die Szene herumbewegt und sich dabei seine Ansicht der Szene nicht im Sinne der Folge 5a, sondern im Sinne der Folge 5b verändert. Wenn jedoch ein System in der Lage ist, auf diese Weise eine zunächst getroffene Fehleinschätzung zu korrigieren, d.h. die von ihm erzeugte Datenbasis entsprechend zu verändern, dann ist dies zumindest ein wichtiger Schritt hin zu der Fähigkeit, Situationen, in denen ein Gegenstand vor einem anderen steht, von Situationen zu unterscheiden, in denen das nicht so ist, und dann scheint mir nichts mehr

dagegen zu sprechen, daß ein solches System tatsächlich über den Begriff "vor" verfügt, so wie wir ihn verwenden.

Wenn das aber der Fall ist und wenn ein System in ähnlicher Weise auch über die Begriffe "Zylinder", "Quader" und über die Farbbegriffe verfügt, dann kann das System auch so eingerichtet werden, daß es Situationen, in denen der Satz "Der grüne Zylinder steht vor einem Quader" wahr ist, von Situationen unterscheiden kann, in denen das nicht der Fall ist, d.h. so, daß es die Wahrheitsbedingungen dieses Satzes kennt. Und in diesem Fall scheint mir dann nichts mehr gegen die Annahme zu sprechen, daß ein solches System den gerade noch einmal angeführten Satz tatsächlich im Wortsinn *verstehen* kann.

Gegen diese Auffassung könnte man versucht sein einzuwenden, daß Systeme der gerade geschilderten Art zwar die *Extension* von Begriffen wie "vor", "Zylinder", "Quader" usw. verstehen können, daß es ihnen aber unmöglich ist, auch die *Intension* dieser Begriffe zu verstehen, daß also allen diesen Systemen *Bedeutungsverstehen* nur im Hinblick auf Extensionen und nicht im Hinblick auf Intensionen zukommt. Ein solcher Einwand würde jedoch auf einem grundlegenden Mißverständnis beruhen. Denn wenn man der herkömmlichen intensionalen Semantik folgt, dann kann die Intension eines Begriffs als eine Funktion aufgefaßt werden, die jeder möglichen Welt die entsprechende Extension dieses Begriffs zuordnet, d.h. die Menge all der Gegenstände, die in dieser möglichen Welt unter diesen Begriff fallen. Dies mag abstrakt klingen; aber es ist in diesem Zusammenhang durchaus von Bedeutung. Denn eine Funktion ist eine Zuordnung, die jedem Gegenstand aus dem Definitionsbereich der Funktion einen bestimmten Wert zuordnet. Eine Funktion kann somit durch jeden Mechanismus realisiert werden, der, angewandt auf einen Gegenstand aus dem Definitionsbereich der Funktion, den entsprechenden Wert dieses Gegenstandes erzeugt. Nach diesem Prinzip funktionieren z.B. Taschenrechner, die, wenn man zwei Zahlen eingibt und die "+"-Taste drückt, als Ergebnis die Summe dieser beiden Zahlen ausgeben. Für die hier diskutierte Frage bedeutet das, daß man jeden Mechanismus, der, wenn man ihm einen beliebigen Gegenstand vorlegt, genau dann z.B den Wert "1" ausgibt, wenn dieser Gegenstand unter den Begriff "Quader" fällt, und genau dann den Wert "0", wenn der Gegenstand nicht unter den Begriff "Quader" fällt, als eine ***Realisierung der Intension*** des Begriffs "Quader" auffassen kann. Jedes System, das über einen solchen Mechanismus verfügt, verfügt damit also über eine Realisierung der Funktion, die die Intension des

Begriffs "Quader" ausmacht. Und von jedem System, das über einen solchen Mechanismus verfügt, kann man daher eher sagen, daß es die Intension, als daß es die Extension des Begriffs "Quader" versteht.

Es ist klar, daß mancher Vertreter des zuvor angeführten Einwandes bei der Intension eines Begriffs weniger an eine bestimmte Funktion denkt, als vielmehr an die Beziehungen, in denen dieser Begriff zu anderen Begriffen steht, oder an das Wortfeld dieses Begriffes. Dazu ist zweierlei zu sagen. Erstens scheint mir der Begriff der Intension, wie er in der intensionalen Semantik verstanden wird, der grundlegendere und systematisch wichtigere Begriff zu sein. Und zweitens ist es kein Problem, Systeme der in diesem Abschnitt geschilderten Art – z.B. durch Implementation semantischer Netze – so zu erweitern, daß sie auch wissen, in welchen Beziehungen ein bestimmter Begriff zu anderen Begriffen steht.

5.

Wenn die bisherigen Überlegungen zutreffen, dann ist es offenbar nicht unmöglich, daß bestimmte Computer bzw. bestimmte Maschinen, deren Kern ein Computer bildet, doch im Wortsinn Sprache verstehen. Wie verhält sich dieses Ergebnis zu den Überlegungen Searles? Bedeutet es, daß Searle mit seiner These unrecht hat, daß Maschinen, deren Verhalten allein durch ein formales Programm bestimmt ist, keine Sprache verstehen können? So einfach ist es sicher nicht. Denn das Verhalten von Systemen, wie sie im letzten Abschnitt angedeutet wurden, ist offenbar *nicht nur* durch formale Programme, sondern auch dadurch bestimmt, daß diese Systeme über ihre visuellen Komponenten in bestimmten *kausalen* Beziehungen zu ihrer Umwelt stehen. Die These, daß solche Systeme möglicherweise in der Lage sind, Sprache zu verstehen, entspricht also ziemlich genau der Auffassung, die J. Fodor in seinen kommentierenden Bemerkungen zu Searles Aufsatz so formuliert hat:

> Searle is certainly right that instantiating the same program that the brain does is not, in and of itself, a sufficient condition for having those propositional attitudes characteristic of the organism that has the brain. . . . However, Searles treatment of the 'robot reply' is quite unconvincing. Given that there are *the right kinds of causal linkages* between the symbols that the device manipulates and things in the world – including the afferent and efferent transducers of the device – it is quite unclear that intuition rejects ascribing propositional attitudes to it. (SB, 431 – Hervorh. vom Verf.)

Aber obwohl Searle Fodors Zugeständnis, "daß die Instantiierung eines Programms keine hinreichende Bedingung für Intentionalität darstellt", offenbar mit großer Freude vermerkt, hält er Fodors Gegenvorschlag immer noch für völlig unzureichend. Auch "geeignete" kausale Verbindungen eines Computers mit der ihn umgebenden Welt befähigen diesen nicht, wirklich Sprache zu verstehen.

> . . . no matter what outside causal impacts there are on the formal tokens, these are not by themselves sufficient to give the tokens any intentional content. No matter what caused the tokens, the agent still doesn't understand Chinese. Let the egg foo yung symbol be causally connected to egg foo yung in any way you like, that connection by itself will never enable the agent to interpret the symbol as meaning egg foo yung. To do that he would have to have, for example, some *awareness* of the causal relation between the symbol and the referent; but now we are no longer explaining intentionality in terms of symbols and causes but in terms of symbols, causes, and intentionality, and we have abandoned both strong AI and the robot reply. (MBP, 454)

Was ist der Grund für diese immer noch ablehnende Haltung Searles? Welche Argumente hat er für seine Auffassung, daß auch Systeme der zuvor geschilderten Art nicht wirklich in der Lage sind, Sprache zu verstehen? Einen ersten Hinweis kann man in den Erwiderungen finden, mit denen Searle auf zwei sehr interessante Einwände gegen seine Thesen reagiert hat: den System-Einwand und den Roboter-Einwand. Den System-Einwand formuliert Searle selbst so:

> While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story. The person has a large ledger in front of him in which are written the rules, he has a lot of scratch paper and pencils for doing calculations, he has 'data banks' of sets of Chinese symbols. Now, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part. (MBP, 419)

Auf diesen Einwand erwidert Searle damit, daß er die Ausgangssituation etwas modifiziert. Man könne durchaus annehmen, so schreibt er, daß die im Zimmer eingesperrte Person alle Elemente des gerade noch einmal geschilderten Systems *internalisiert*. Sie lernt die Regeln des Regelbuchs auswendig; ebenso alles, was in den beiden Körben auf den verschiedenen Blättern mit chinesischen Zeichen steht. Sie führt alle Berechnungen im Kopf aus. Kurz, die Person inkorporiert alles, was für das System wichtig ist. Man kann sogar annehmen, daß die Person nicht in einem Zimmer eingesperrt ist, sondern irgendwo im Freien arbeitet. Auch in diesem Fall versteht die

Person jedoch kein Chinesisch; denn an den Grundzügen der Situation hat sich nichts verändert. Und das bedeutet, daß auch das Gesamtsystem kein Chinesisch versteht; denn, so wie die Situation jetzt konstruiert ist, *ist* die Person das System.

Der Roboter-Einwand stellt einen anderen Aspekt in den Vordergrund. Diesen Einwand formuliert Searle so:

> Suppose we wrote a different kind of program from Schank's program. Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking – anything you like. The robot would, for example, have a television camera attached to it that enabled it to 'see', it would have arms and legs that enabled it to 'act', and all of this would be controlled by its computer 'brain'. Such a robot would, unlike Schank's computer, have genuine understanding and other mental states. (MBP 420)

Auch dieser Einwand führt jedoch Searle zufolge nicht zum Ziel. Denn seiner Meinung nach ändert die Hinzufügung von "Wahrnehmungs- und Bewegungskomponenten" nichts an der Fähigkeit (oder besser: Unfähigkeit) des Systems, wirklich Sprache zu verstehen. Dies zeigt sich, so Searle, daran, daß man auf das von den Vertretern des Roboter-Einwandes ins Spiel gebrachte System dasselbe Gedankenexperiment anwenden kann. Angenommen, statt eines Computers in einem Roboter sitzt – wie in der Ausgangssituation – eine Person in einem Zimmer. Diese Person bekommt noch mehr chinesische Schriftzeichen und noch mehr in einer für sie verständlichen Sprache abgefaßte Regeln, nach denen sie auf eingehende Symbole mit der Ausgabe von Symbolen reagiert. Weiter angenommen, einige der eingehenden Symbole kommen, ohne daß die Person dies weiß, von einer Fernsehkamera und einige von ihr ausgegebene Symbole steuern die Motoren des Roboters so, daß sich seine Arme oder Beine auf eine bestimmte Weise bewegen. Offenbar ändert dies, so schreibt Searle, überhaupt nichts daran, daß die Person im Zimmer nichts anderes tut, als formale Symbole nach formalen Regeln zu manipulieren. Sie empfängt zwar "Informationen" von der Fernsehkamera, und sie gibt "Instruktionen" aus zur Bewegung der Arme und Beine des Roboters. Aber sie weiß nicht, daß sie das tut. Für sie ist die Situation allein dadurch charakterisiert, daß sie formale Symbole erhält und formale Symbole ausgibt. Und das tut sie nach rein formalen Regeln, ohne

auch nur die geringste Kenntnis davon zu haben, was diese Symbole bedeuten könnten.

Mir scheint, daß Searle bei dieser Erwiderung auf den Roboter-Einwand von einem einfachen Trick Gebrauch macht. Und dieser Trick besteht darin, daß er bei der Übertragung der ursprünglichen Chinese Room-Situation auf den Roboterfall die Grenze zwischen dem Computer bzw. der Person auf der einen und den übrigen Teilen des Systems auf der anderen Seite so zieht, daß z.B. der Computer bzw. die Person als Eingabe immer nur das erhält, was die "Wahrnehmungskomponenten" als output liefern. Nur auf diese Weise kann Searle sicherstellen, daß der Computer oder die Person immer nur mit formalen Mustern und nicht direkt mit der Welt konfrontiert sind. Und auch nur auf diese Weise kann er seine Schlußfolgerung erreichen, daß der Computer *in* dem Roboter immer noch kein Chinesisch versteht. Doch das war gar nicht der Streitpunkt. Denn es geht ja auch nicht darum, ob das *Gehirn* einer Person eine Sprache versteht, sondern ob die ganze Person dies tut.

Aus diesem Grund wird der Trick Searles auch sofort sichtbar, wenn man den Roboter-Einwand mit dem System-Einwand verbindet, so wie die Vertreter des Roboter-Einwandes diesen Einwand offenbar auch schon von Anfang an gemeint hatten. Denn selbst Searle formuliert diesen Einwand so, daß er auf die Schlußfolgerung hinausläuft, daß man von "solchen Robotern" (und nicht etwa von den in "solchen Robotern" steckenden Computern) zu recht sagen kann, daß sie tatsächlich Sprache verstehen.

Was könnte Searle auf einen solchen verbundenen Einwand erwidern? Der Strategie seiner Erwiderung auf den System-Einwand folgend müßte er behaupten, daß sich auch in diesem Fall nichts Wesentliches ändern würde, wenn die Person alle für das System (den gesamten Roboter) relevanten Teile inkorporiert. Doch das würde in diesem Fall eben auch die Inkorporierung der Fernsehkamera bedeuten und somit zur Folge haben, daß die Person als Gesamtsystem nicht mehr nur formale Symbole manipuliert, sondern auch in bestimmter Weise auf die Außenwelt reagiert, z.B. bestimmte formale Symbole überhaupt erst als Reaktion auf bestimmte von ihr wahrgenommene Situationen herstellt. Wenn jedoch die Person als Gesamtsystem etwa das Symbol "grauer Hut" immer und nur in Situationen generiert, in denen sie einen grauen Hut wahrnimmt, dann scheint es doch nicht mehr völlig unplausibel anzunehmen, daß diese

Person als Gesamtsystem weiß, was das Symbol "grauer Hut" bedeutet. Fodor hat deshalb offensichtlich recht, wenn er schreibt, Searles Erwiderung auf den Roboter-Einwand sei nicht besonders überzeugend.

Searle jedoch findet seinerseits diese Auffassung Fodors wenig überzeugend. Denn seiner Meinung nach ermöglicht das bloße Bestehen von kausalen Beziehungen – welcher Art auch immer – zwischen einem Symbol und dem durch das Symbol Bezeichneten noch kein wirkliches Sprachverstehen.

> To do that [the system] would have to have, for example, some *awareness* of the causal relation between the symbol and the referent. (MBP, 454)

Möglicherweise findet sich in einer eher beiläufigen Bemerkung Searles der Schlüssel für die nicht leicht zu verstehende Diskrepanz zwischen den Ansichten Searles und Fodors. Denn in seiner Erwiderung auf den System-Einwand schreibt Searle unter anderem:

> . . . the English subsystem *knows* that 'hamburgers' *refers* to hamburgers . . . (MBP, 419 – Hervorh. vom Verf.)

Und:

> But the Chinese system *knows* none of this. (ebd. – Hervorh. vom Verf.)

Die entscheidende Frage im Zusammenhang mit dieser Bemerkung scheint mir zu sein, was hier mit "wissen" gemeint sein soll. Der Hinweis auf die Notwendigkeit von "awareness" in seiner Erwiderung auf Fodor legt die Vermutung nahe, daß Searle davon ausgeht, daß zum Sprachverstehen *explizites* Wissen im Sinne von "wissen, daß" erforderlich ist. Über Wissen dieser Art verfügen die oben im Abschnitt 4 geschilderten Systeme (und an ähnliche Systeme scheint mir auch Fodor zu denken) natürlich nicht. Denn zwar wissen auch diese Systeme in einem gewissen Sinn, was das Wort "Quader" bedeutet, insofern nämlich, als sie Situationen, in denen Gegenstände die durch dieses Wort bezeichnete Eigenschaft haben, von solchen Situationen unterscheiden können, in denen das nicht der Fall ist, und als sie das Wort "Quader" daher in den verschiedensten Situationen richtig verwenden können. Aber ein solches Wissen ist offensichtlich implizit, ein "wissen, wie". Denn es besteht allein aus einer Reihe von Fähigkeiten.

Die entscheidende Frage ist somit, ob Searle recht hat, wenn er

explizites Wissen um die Bedeutung sprachlicher Ausdrücke zur Voraussetzung von Sprachverstehen erklärt. Soweit ich sehen kann, führt er für diese Auffassung keine Gründe an. Und ich kann auch selbst keine systematischen Gründe erkennen. Denn Sprachverstehen ist in der Sprachphilosophie immer wieder mit der Fähigkeit, sprachliche Ausdrücke richtig zu verwenden, in Zusammenhang gebracht worden. Und diese Fähigkeit setzt offenbar nur ein implizites Wissen voraus. Mir scheint daher, daß implizites Wissen um die Bedeutung sprachlicher Ausdrücke zum Sprachverstehen ausreicht und daß man Systeme der im Abschnitt 4. behandelten Art demzufolge völlig zurecht als sprachverstehende Systeme bezeichnen kann.[4] Und dies ist, wie mir scheint, auch genau die Auffassung, die den Überlegungen Fodors zugrundeliegt.

ANMERKUNGEN

[1] Schon an dieser Stelle ist es wichtig, zu betonen, daß das Chinese-Room-Experiment nicht wirklich ein Argument ist, sondern eine Art Gedankenexperiment, in dem an die Intuitionen der Leser oder Hörer appelliert wird. Searle bittet den Leser, sich gewisse Situationen vorzustellen und dann zu sagen, ob in diesen Situationen bestimmte Zuschreibungen berechtigt sind oder nicht. Bei solch einem Gedankenexperiment ist es daher von entscheidender Bedeutung, daß die betreffenden Situationen so klar wie möglich geschildert und nicht durch ungenaue Darstellungen verfälscht werden. Dieser Punkt scheint mir gerade im Hinblick auf Searles Erwiderungen auf den System- und den Roboter-Einwand, auf die ich am Schluß des Aufsatzes zu sprechen kommen werde, von großer Wichtigkeit zu sein.

[2] Searle selbst betont aber, seine Überlegungen träfen auch auf andere Systeme dieser Art zu, z.B. auf das System ELIZA von Weizenbaum und das System SHRDLU von Terry Winograd.

[3] Diese Aufstellung ist sehr verkürzt. Sie berücksichtigt z.B. nicht, daß etwa für die Wörter "eine" und "Frikadelle" die Kasus mit angegeben werden sollten und daß dementsprechend für diese Wörter verschiedene Möglichkeiten berücksichtigt werden müssen. Für diesen Zusammenhang reicht jedoch die angeführte vereinfachte Aufstellung zur Illustration aus.

[4] Vgl. jedoch die Überlegungen R. Cummins', der in *The Nature of Psychological Explanation* an mehreren Stellen (z.B. S. 76 ff.) auf einer ähnlichen Linie wie Searle argumentiert, daß wirkliches Verstehen voraussetzt, daß das System seine eigenen Repräsentationen versteht, daß diese Repräsentationen Repräsentationen auch für das System selbst sind.

LITERATUR

Cummins, R.: 1983, *The Nature of Psychological Explanation*, The MIT Press, Cambridge, Massachusetts.

Fodor, J. A., (SB): 1980, 'Searle on what only Brains can do', in *The Behavioral and Brain Sciences* **3**, 331f.

Searle, J., (MPB): 1980, 'Minds, Brains and Programs', in *The Behavioral and Brain Sciences* **3**, 417–24, 450–56.

Searle, J., (GHW): 1986, *Geist, Hirn und Wissenschaft*, Suhrkamp Verlag, Frankfurt/M. (dt. Übersetzung von *Minds, Brains, and Science. The 1984 Reith Lectures*, British Broadcasting Corporation, London, 1984.)

Winograd, T.: 1972, *Understanding Natural Language*. Academic Press, New York.

Winograd, T.: 1984, 'Software für Sprachverarbeitung', in *Spektrum der Wissenschaft*, S. 88–102.

Manuscript received 27 November 1986

Philosophisches Seminar
Georg-August-Universität
Platz der Göttinger Sieben
D-3400 Göttingen
F.R.G.