Professor Dr. Wolfgang Klein
Universität Heidelberg

Organisation des Wissens durch Sprache

Konsequenzen für die maschinelle Sprachanalyse

Gliederung

Das Wissen, das sich die Menschen zu einer bestimmten Zeit erworben zu haben glauben, wird weithin mit Hilfe der natürlichen Sprache festgehalten (»kodifiziert«) und weitervermittelt. Zu diesem in natürlich-sprachlichen Äußerungen kodifizierten Wissen hat man jedoch mit einem Computer kaum direkt Zugang. Zwar bemüht man sich seit vielen Jahren mit zum Teil erheblichem Aufwand um beispielsweise automatische Informationserschließung, maschinelle Sprach-Übersetzung und Mensch-Maschine-Dialoge in natürlicher Sprache, aber die Ergebnisse sind bescheiden. Verantwortlich für den in diesen Bereichen vergleichsweise geringen Erfolg sind verschiedene Eigenschaften der natürlichen Sprachen, die - im Gegensatz zu formalen Sprachen (wie Programmiersprachen, gängige logische Sprachen) — die maschinelle Informationserschließung erschweren.[*)]

1 Exkurs in die »Logik« der Sprache

Ausgangspunkt der Überlegungen soll eine Betrachtung des Begriffs »Wissen« sein. Im Deutschen gibt es mindestens zwei verschiedene solche Begriffe, die man im Anschluß an die analytische Philosophie oft als »wissen, wie« und als »wissen, daß« auseinanderhält. Bekannter sind die entsprechenden englischen Bezeichnungen des »knowing how« und »knowing that«. Zwischen »wissen« im Sinne von »sich auf etwas verstehen« und »wissen« im Sinne von »die Wahrheit über etwas kennen« besteht natürlich ein enger Zusammenhang - vor allem dann, wenn die Inhalte des »wissen, daß« mögliche Handlungen betreffen, also praktischer Natur sind. Das soll aber hier nicht weiter erörtert werden; im folgenden geht es ausschließlich um das »wissen, daß« - was immer seine Inhalte sein mögen.

Mit diesem »wissen, daß« gerät man nun in der Sprache in eine merkwürdige Paradoxie, die uns vielleicht wenig über das Wissen, aber viel über die Sprache[1)] sagt. Sie besteht darin, daß man entweder jeden Satz als wahr annehmen muß bzw. als wahr nachweisen kann - was offenbar unsinnig ist - oder aber das fundamentale Prinzip des *»tertium non datur«* aufgeben muß, also das Prinzip, nach dem immer entweder ein Aussagesatz wahr ist oder aber seine Negation. Dies hängt damit zusammen, daß man zwar glauben kann, was man will, hingegen *wissen* kann man nur Tatsachen (das heißt Sachverhalte), die wahr sind. Niemand kann wissen, daß die Erde eine Scheibe ist - und wenn er noch so fest davon überzeugt ist. Es ist nämlich nicht so, und wenn jemand sagt: »Fritz weiß, daß die Erde eine Scheibe ist«, dann ist dies falsch. (Fritz kann dies nämlich höchstens glauben oder zu wissen meinen.) Wenn es aber zutrifft, daß er es nicht weiß, dann ist es — nach dem Satz vom ausgeschlossenen Dritten - wahr, daß er es nicht weiß; das heißt der folgende Satz ist wahr: »Fritz weiß nicht, daß die Erde eine Scheibe ist«. Aber auch dieser Satz kann nur wahr sein, wenn die Erde eine Scheibe ist. Man kann sich dies leicht überlegen. Wenn ich — um ein weniger offensichtliches Beispiel zu nehmen - etwa sage: »Fritz weiß nicht, daß seine Frau ihn betrügt«, dann bestreite ich damit gewisse Kenntnisse von Fritz hinsichtlich einer offenkundig feststehenden Tatsache. Es gilt also folgendes: Sowohl aus dem Satz »Fritz weiß, daß die Erde eine Scheibe ist« wie aus dem Satz »Fritz weiß nicht, daß die Erde eine Scheibe ist« ergibt sich zwangsläufig, daß die Erde eine Scheibe ist. Einer der beiden Sätze muß aber wahr sein: entweder er weiß es, oder er weiß es nicht; ein Drittes ist nicht möglich.

Man kann dies etwas allgemeiner ausdrücken: Sowohl aus einem Satz *a* wie aus dessen Negation — *a* ergibt sich die Wahrheit eines anderen Satzes *b;* da aber *a* oder *nicht a* wahr ist, ergibt sich *b* als wahr. Man mag sich drehen und wenden, wie man will: Die Erde ist eine Scheibe. Natürlich kann man auf dieselbe Weise auch zeigen, daß die Erde eine Kugel, eine Banane oder ein Teesieb ist, denn auch hier gilt, daß Fritz dies weiß oder nicht.

*) Textfassung eines Vortrags am 14. September 1976 in Heidelberg auf dem »Forum für Wissenschaft und Verwaltung« der IBM Deutschland. Der Text wurde für den Druck leicht überarbeitet und gekürzt, der Vortragsstil jedoch beibehalten.

[1)] Mit »Sprache« ist hier und im folgenden, soweit nicht ausdrücklich anders gesagt, immer eine *natürliche* Sprache (wie Deutsch, Englisch, Russisch) gemeint.

Offenbar hängt das damit zusammen, daß ein Verbum wie »wissen, daß« nur anwendbar ist, wenn der Inhalt des »daß-Satzes« wahr ist. Man sagt auch, das Verbum hat eine bestimmte *Präsupposition,* eben die Faktizität dessen, was im »daß-Satz« steht. »Wissen, daß« ist übrigens keineswegs der einzige derartige Fall. Ein anderes, bekannteres, aber auch etwas verklausulierteres Beispiel ist die Frage eines Richters an einen Angeklagten: »Haben Sie aufgehört, Ihre Frau zu verprügeln?«. Ganz gleich, ob er antwortet: »Ja, ich habe aufgehört« oder »Nein, ich habe nicht aufgehört«, er akzeptiert als Faktum, daß er seine Frau verprügelt hat. Hier wird noch deutlicher, daß es sich nicht um eine echte Paradoxie unserer Erkenntnis, sondern um eine Paradoxie der natürlichen Sprache handelt.

Es gibt, wenn man wieder an das Beispiel »wissen, daß« denkt, drei Möglichkeiten, sie aufzulösen:

1. Man gibt das Prinzip des *»tertium non datur«* auf, das heißt das Prinzip, daß ein Satz oder aber seine Negation wahr ist. Dies ist wahrscheinlich die einzig korrekte Lösung, aber der Satz vom ausgeschlossenen Dritten ist ein so fundamentales logisches Prinzip, daß man es nur ungern fallen ließe.

2. Man sagt, der Ausdruck »wissen, daß« darf nur in Fällen angewandt werden, in denen die Präsupposition erfüllt, das heißt der »daß-Satz« wahr ist. Dessen kann man aber nie sicher sein, und deshalb liefe dies, konsequent durchgeführt, auf ein Verbot des Verbums »wissen, daß« hinaus. Trotzdem ist dies die Lösung, die man in der Praxis wählt. Wenn Sie noch einmal an das Beispiel des Richters denken: Man erwartet einfach von einem verantwortungsvollen Richter, daß er keine Sätze mit nicht erfüllten Präsuppositionen verwendet.

3. Man gibt den absoluten Begriff von »wissen, daß« auf und ersetzt ihn durch einen relativen. »Wissen, daß« heißt dann »relativ zu bestimmten Standards mit Recht für wahr halten«. Wenn man also dann sagt: »Fritz weiß, daß die Erde eine Scheibe ist«, so bedeutet dies: »Fritz ist zu der festen Überzeugung gekommen, daß die Erde eine Scheibe ist, und relativ zu seinen Überprüfungsmöglichkeiten ist diese Überzeugung berechtigt«. In diesem Fall tritt keine Paradoxie auf: Wissen im absoluten Sinn kann man nur etwas, was wahr ist: wissen im relativen Sinn kann man auch etwas Falsches. Diese Lösung klingt sehr verführerisch, weil wir alle wissen, daß unser Wissen ständig revidiert wird, also relativ ist. Man muß sich aber zwei Dinge klar machen: Zum einen ist der Wissensbegriff unserer Sprache offenbar der absolute, und ihn aufgeben heißt, unsere Sprache künstlich zu normieren, um diese eine Paradoxie - neben vielen anderen - zu vermeiden. Und zum andern: Es dreht sich keineswegs darum, zuzugestehen, daß man oft etwas zu wissen meint und es dann — wie der Fortgang der Wissenschaft zeigt - eben doch nicht gewußt hat. Dies ist selbstverständlich: Was heute als wahr gilt, kann morgen als falsch erwiesen sein und vielleicht übermorgen wieder als wahr. Es handelt sich vielmehr darum, daß der Begriff des absoluten Wissens selbst zu Paradoxien führt und wir ihn deshalb als Begriff aufgeben müßten. Dies hätte aber die Konsequenz, daß wir selbst in Sätzen wie »2 mal 2 ist 4« dies nicht genau wissen, sondern nur ganz, ganz fest glauben können.

Wenn man in der Praxis von »dem Wissen« spricht, so meint man damit gewöhnlich so etwas wie ein »relatives Wissen«, das heißt das, was Menschen zu einer bestimmten Zeit, zu einem bestimmten Grad der Erforschung der Realität und vielleicht relativ zu anderen Faktoren als wahr anzunehmen berechtigt sind. In diesem Sinne spricht man etwa von dem Wissen, das in unseren großen Konversationslexika oder in anderen enzyklopädischen Werken niedergelegt und uns dort zugänglich ist. Dieses Wissen will ich hier als kodifiziertes Wissen bezeichnen. Es ist das, worüber die Fachleute eines Gebietes sich mehr oder minder einig sind und wovon sie andere, die ihnen vertrauen, überzeugt haben. Zu diesem Zweck muß das Wissen weitervermittelt, mitgeteilt werden können, und die bei weitem wichtigste Form sowohl zur Mitteilung wie zur Kodifizierung ist die natürliche Sprache. Dies ist in mancher Hinsicht ein Wunder.

Niemandem ist dies deutlicher geworden als jenen, die nun vor schon fast 30 Jahren - beinahe zugleich mit dem Aufkommen der ersten elektronischen Rechenanlagen überhaupt - mit dem Versuch begonnen haben, Sprachen wie Englisch, Deutsch, Russisch einer maschinellen Behandlung zugänglich zu machen. Die Sprachanalyse[2)] war dabei nicht Selbstzweck, sondern Mittel - zunächst für eine maschinelle Sprachübersetzung, später für das Information Retrieval und Automatic Abstracting, noch später für den Mensch-Maschine-Dialog in natürlicher Sprache. Die Ergebnisse sind bekanntlich bescheiden, ja kläglich, wenn man sie an den Zielen mißt. Dies liegt nicht am mangelnden Bedarf, denn

[2)] Mit »Sprachanalyse« ist hier und im folgenden immer *automalische* Sprachanalyse gemeint

der praktische Nutzen solcher Verfahren ist offenkundig und wurde auch klar gesehen. Es liegt auch nicht an mangelnden Forschungsmitteln oder gar an der Unfähigkeit der Forscher; es liegt vielmehr an der Struktur der Sprache, die wir alle als Kinder so mühelos lernen und dann Tag für Tag benutzen. Letzteres führt uns vielleicht auch dazu, ihre Komplexität zu unterschätzen.

Einige der wichtigsten Eigenschaften der natürlichen Sprache, und zwar jene, die für eine Sprachanalyse besonders problematisch sind, will ich im folgenden skizzieren. Es kann sicherlich aufschlußreich sein, wenn man sich kontrastierend dazu die Struktur einer möglichst reichen Programmiersprache vor Augen führt.

2 Einige wesentliche Eigenschaften natürlicher Sprachen und ihre Bedeutung für die Sprachanalyse

Die erste der hier zu nennenden spezifischen Eigenschaften einer natürlichen Sprache ist ihre *Variabilität.* Wenn man von »deutscher Sprache« spricht, so verbindet sich damit oft die Vorstellung von etwas Feststehendem, klar Umrissenem. Man denkt,es gibt»diedeutscheSprache«, die von den einzelnen Sprechern mehr oder minder gut beherrscht wird. Diese Vorstellung ist sehr verbreitet, aber völlig irreführend. Man kann über die Sprache sinngemäß sagen, was der berühmte Richter *Oliver Wendell Holmes* über das Common Law sagte: »Sie ist nichts Allgegenwärtiges in den Wolken, das über den einzelnen waltet«. - Was es an einer Sprache tatsächlich gibt, ist das sprachliche Verhalten der einzelnen sprechenden und schreibenden Individuen, und erst aus dem, was sie tun, kann man durch komplizierte Abstraktionen auf ein geheimnisvolles Wesen wie »die deutsche Sprache« schließen.

Nun ist, wie jedermann weiß, das sprachliche Verhalten oft sehr unterschiedlich, wenn auch keineswegs regellos. Es kann in Abhängigkeit von bestimmten außersprachlichen Faktoren gewisse Ausprägungen annehmen; solche Faktoren sind beispielsweise die Gegend, in der ein Sprecher zuhause ist: ein Heidelberger spricht gewöhnlich anders als ein Hamburger, dieser anders als ein Berliner, dieser anders als ein Frankfurter oder ein Wiener. Ein zweiter Faktor ist die soziale Schicht: ein Akademiker spricht gewöhnlich anders als ein Hilfsarbeiter. Ein dritter Faktor ist die Zeit: auch vor 500 Jahren sprach man schon deutsch, wenngleich deutlich anders als heute. Ein vierter Faktor schließlich ist die Redesituation: auf der Kanzel spricht der Pfarrer anders als im Wirtshaus. Je nachdem welcher Faktor im Spiel ist, zeigen sich die Unterschiede oft schwerpunktmäßig auf verschiedenen Ebenen: so vor allem im Lautlichen bei räumlichen oder, wie man oft sagt, dialektalen Unterschieden, im Bereich des Wortschatzes bei Variation nach Redesituation, im Satzbau und im Wortschatz bei zeitlicher Variation, in allen drei Bereichen bei sozial bedingter Variation. Dies will ich hier nicht weiter verfolgen. Wichtig ist, daß die Sprache nichts Einheitliches ist, sondern ein sehr komplexes System einzelner, oft sehr verschiedenartiger Ausprägungen. Man kann eine dieser Ausprägungen, z. B. die sogenannte *Dudennorm,* für besonders schön oder nützlich halten, aber sie ist natürlich nicht *die* deutsche Sprache. Man macht sich auch nur selten klar,daß die meisten Deutschen alltäglich unter sich einen Dialekt oder eine dialektähnliche Sprachform sprechen.

Die Variabilität der Sprache, die hier natürlich nur angedeutet werden konnte, stellt für den Linguisten, der die ganze Sprache in ihrem Zusammenhang mit vielen Faktoren beschreiben will, ein gewaltiges Problem dar. Für die automatische Sprachanalyse verursacht sie zum Glück nur gelegentlich Schwierigkeiten. Denn dort ist man gewöhnlich nicht an der Sprache, wie sie tatsächlich gesprochen wird, interessiert, sondern an sehr eingeschränkten Ausprägungen, wie sie sich etwa in wissenschaftlichen Texten, juristischen Texten oder Zeitungstexten zeigen. Schwierigkeiten entstehen hier fast nur bei Unterschieden im Wortschatz, und zwar bei sogenannten Fachsprachen. Ein Wort wie »Familie« bedeutet in der Sprache der Juristen, in der Sprache der Soziologen oder gar in der Sprache der Mathematiker sehr verschiedene Dinge, und für diese semantische Variabilität gibt es Dutzende von Beispielen. Immerhin kann man dieses Problem bei der Sprachanalyse weitgehend in den Griff bekommen, und zwar, indem man so etwas wie einen Fachsprachenindex setzt. Man gibt, grob gesagt, an: bei dem zu analysierenden Text handelt es sich um einen mathematischen, also ist von den verschiedenen Bedeutungen des Wortes »Familie« die besondere herauszusuchen. Dies klappt nicht immer, aber doch in den meisten Fällen.

Das zweite wesentliche Charakteristikum der natürlichen Sprache - im Gegensatz zu einer künstlichen — ist ihre *semantische Universalität* oder, wie man auch sagen könnte, ihre *semantische Offenheit.* Damit ist folgendes gemeint: In der natürlichen Sprache kann man praktisch über alles reden - über juristische Probleme und über Lungenkrebs, über die

Schönheit alter Städte und junger Mädchen, über Fußball und Verwaltungsreformen, über den Mond in lauen Sommernächten und als astronomisches Objekt, über die rote Gefahr und über Einhörner, über die Sprache selbst, kurz: über alles. Wir können auch jederzeit neue Termini in ihr bilden, wann immer dies zweckmäßig oder erforderlich scheint.

Diese semantische Offenheit stellt für die automatische Sprachanalyse allerdings ein erhebliches Problem dar. Um beispielsweise ein Dialogsystem für die Kommunikation in natürlicher Sprache aufzubauen, benötigt man unter anderem ein Lexikon, ein Verzeichnis der Wörter, die in den Sätzen auftreten können. Dieses Wortverzeichnis kann sehr groß sein, aber in der natürlichen Sprache ist es so, daß das Repertoire überhaupt keine feste Obergrenze hat, weil jeder nach Belieben neue Wörter, vor allem Namen, einführen kann und trotzdem verstanden wird. Dieses Problem ist aber nicht generell unlösbar, und in der Tat hat man bereits »lernende« Verfahren entwickelt, die trotzdem eine Analyse durchführen und gegebenenfalls das Lexikon automatisch erweitern. Was also die semantische Universalität betrifft, so gibt es erhebliche praktische Schwierigkeiten, aber man hat durchaus klare Lösungsperspektiven.

Das dritte Merkmal der Sprache, das hier von Belang ist, ist die *Vagheit* ihrer Äußerungen, Die Vagheit tritt bei sehr vielen sprachlichen Äußerungen in Erscheinung. Ich will einige Beispiele nennen, die nicht so augenfällig sind wie die Reden von Politikern. Wenn jemand sagt: »In Heidelberg regnet es leicht«, so ist dieser Satz zunächst einmal in elementarer Weise mehrdeutig. Er kann bedeuten: »In Heidelberg fällt ein leichter Regen«, aber auch »In Heidelberg kommt es leicht dazu, daß es regnet«. Ich weiß nicht, an welche Bedeutung Sie zuerst gedacht haben, aber wie dem auch sei, beide Deutungen sind in sich noch einmal vage. Was soll »ein leichter Regen« heißen ? Der Ausdruck ist nicht festgelegt auf soundsoviel Kubikzentimeter pro Quadratmeter pro Zeiteinheit, und eine solche Festlegung wäre auch unsinnig. Ebenso kann man fragen, was es heißen soll, daß es leicht zu Regen komme. Heißt es »oft«? Wahrscheinlich nicht, aber selbst wenn: Was heißt schon »oft«? Für die Sahara wäre zweimal im Jahr sicher schon oft, für Husum bestimmt nicht. Es bleibt unbestimmt, was »oft« bedeutet. Oder, um ein anderes Beispiel zu geben: Selbst wenn man das Wort »rot« als Farbe und nicht als politische Kennzeichnung meint, ist es höchst unbestimmt. Dieselbe Farbe, die man bei Haaren unbedenklich als »rot« bezeichnen würde, würde man bei einem Mantel eher »braun« nennen. Ein drittes Beispiel: Über das, was »groß« - im Sinne von Körpergröße - ist, herrschen sehr verschiedene Vorstellungen. Was für einen Dackel groß ist, wäre für eine Dogge eher klein. Wenn ein Elefant klein ist, ist er trotzdem groß.

Diese Vagheit, von der kaum ein Satz - für sich genommen - frei ist, ist kein Fehler, sondern einer der größten Vorzüge der natürlichen Sprache. Man kann nämlich jederzeit genauer werden, nur ist dies gewöhnlich nicht notwendig oder nicht erwünscht, weil ohnehin hinlänglich klar ist, was gemeint ist, und jede weitere Angabe wäre dann sinnlos, überflüssig und unökonomisch. Vor allem ist es so möglich, sich zwanglos an die Erfordernisse der Situation einesteils und andernteils an die Genauigkeit unserer alltäglichen Kenntnisse - etwa über den Regen in Heidelberg - anzupassen.

Was nun die Behandlung der Vagheit in der Sprachanalyse anbelangt, wo sie Probleme verursachen kann (etwa bei Frage-Antwort-Systemen), so gibt es hierfür praktisch keine Lösung. Es bleibt, falls die Vagheit erkannt wird, nur die Möglichkeit einer gesteuerten Rückfrage mit der Bitte um Präzisierung - wie man dies auch in der alltäglichen Rede macht. Allerdings hat man bisher auch kein Verfahren, Vagheit automatisch als solche zu erkennen.

Mit der Vagheit in engem Zusammenhang steht die *Mehrdeutigkeit.* Beide sind aber nicht zu verwechseln. Wenn ein Satz mehrdeutig ist, so können - wie wir am Beispiel »In Heidelberg regnet es leicht« gesehen haben - alle Deutungen noch einmal vage sein. Das Phänomen der Mehrdeutigkeit ist - vor allem im Zusammenhang mit maschineller Sprachübersetzung - schon recht gut untersucht worden. Man kann hier drei Stufen unterscheiden, nämlich syntaktische, semantische und pragmatische Mehrdeutigkeit. Die erste - die *syntaktische* - bezieht sich auf den formalen Aufbau von Sätzen aus kleineren und kleinsten Teilen. Man denke etwa an einen Ausdruck wie »Alte Männer und Frauen sind hilflos«. Hier kann sich das Adjektiv »alte« auf Männer allein beziehen (wie in »Alte Männer und Kinder sind hilflos«) oder aber auf Männer und auf Frauen. Ein anderes, sehr bekanntes Beispiel ist der Satz »Ein Junggeselle ist ein Mann, dem zum Glück die Frau fehlt«. Hier kann der Satzteil »zum Glück« einmal Präpositionalobjekt sein - »fehlen zu« - oder aber Umstandsergänzung »glücklicherweise«. Solche syntaktischen Mehrdeutigkeiten sind ausge-

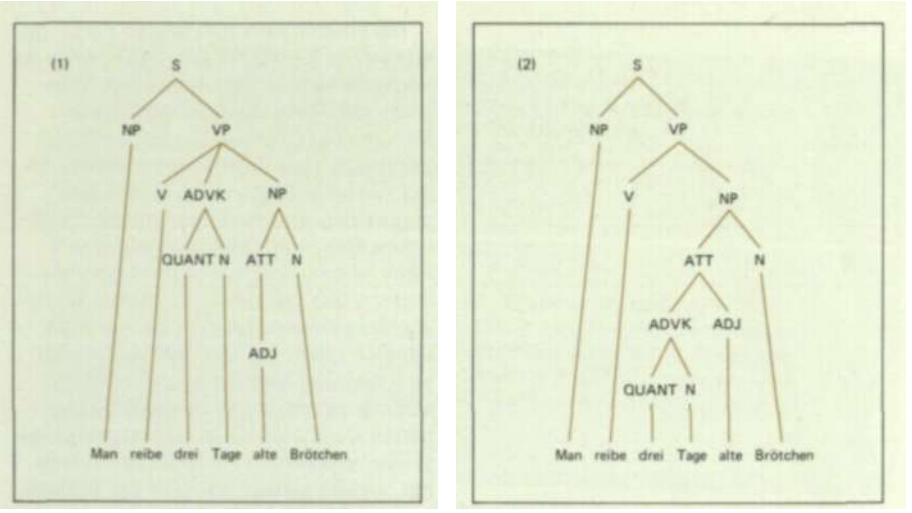

Abbildung 1: Don beiden unterschiedlichen Lesarten des Beispielsatzes entsprechen verschiedene Strukturen, hier angezeigt durch die *Strukturbäume* (1) und (2) - eine in der Linguistik übliche Form der Strukturbeschreibung. Rund 40 Prozent aller Wörter in einem laufenden Text sind syntaktisch mehrdeutig. Die meisten Mehrdeutigkeiten können jedoch innerhalb der Satzgrenze aufgelöst werden, und nicht alle haben semantische Konsequenzen wie in diesem Beispiel

sprochen häufig (vgl. auch *Abbildung 1).* Wir merken sie gewöhnlich gar nicht, aber für eine maschinelle Analyse stellen sie ein ungeheures Problem dar. Immerhin: Dies ist eines der wenigen Probleme, die als weithin gelöst gelten können, und dies ist zugleich einer der wenigen Glanzpunkte in der Geschichte der automatischen Sprachanalyse.

Anders steht es mit den *semantischen* Mehrdeutigkeiten, die, oft aber nicht immer, eine Folge von syntaktischen sind. Damit sind einfach unterschiedliche Bedeutungen eines Wortes, Satzteils oder Satzes gemeint. Solche Mehrdeutigkeiten können, wie wir schon gesehen haben, fachsprachlich bedingt sein; dann fallen sie unter die Variabilität der Sprache. Es gibt aber auch Mehrdeutigkeiten innerhalb einer Fachsprache. Es kann etwa in der Jurisprudenz für die Ermittlung eines Straftatbestandes wichtig sein, ob sich etwas am Tag oder in der Nacht abgespielt hat - beispielsweise wenn es darum geht, ob das Licht am Auto eingeschaltet war. In diesem Fall bedeutet »Tag« ungefähr soviel wie »Zeit, in der es hell ist«. Es kann aber auch jemand zu »dreißig Tagen ersatzweise« verurteilt werden. Dann bedeutet »Tag« leider mehr, nämlich 24 Stunden. Ich selbst verdanke einer juristischen Mehrdeutigkeit ein Schlüsselerlebnis in Sachen Recht, als ich einen bekannten Juristen erklären hörte, daß das Wort »muß« in juristischen Texten nicht immer »muß« bedeuten muß, sondern auch »kann« bedeuten kann. Ich wäre, weder in meiner Eigenschaft als Sprecher des Deutschen noch als Sprachwissenschaftler, je auf diese Mehrdeutigkeit gekommen, aber man muß hier dem Juristen wohl glauben (oder zumindest *kann* man es).

Was nun die Frage der automatischen Sprachanalyse und die Auflösung semantischer Mehrdeutigkeiten angeht, so kann man dies Problemfeld nur als ein Cannae der Sprachanalyse bezeichnen. Rein semantische Mehrdeutigkeiten, die nicht fachsprachlich bedingt sind, kann man fast nie auflösen, und dies, obwohl sie einem Menschen in einer normalen Redesituation fast nie Schwierigkeiten bereiten, ja oft gar nicht bemerkt werden. Am augenfälligsten ist dies bei den sogenannten *Metaphern* - nicht einmal den poetischen, sondern ganz alltäglichen wie etwa, wenn man von einem »Meilenstein in der Compiler-Entwicklung« oder einem »Abgrund an Verlogenheit« spricht. Solche spontan verwendeten Metaphern sind überaus häufig; sie fallen uns kaum auf, aber sie stellen die automatische Sprachanalyse, wenn sie über nichttriviale Anwendungen hinauskommen soll, vor fast unlösbare Probleme.

Die dritte Form der Mehrdeutigkeit, die *pragmatische,* ist eine Mehrdeutigkeit im Hinblick auf die Intention, die ein Sprecher mit seiner Äußerung verfolgt. Wenn z. B. jemand sagt »Ich habe Hunger« oder »Sie sind ein Schafskopf«, so sind dies zunächst einmal Behauptungen über irgendwelche Sachverhalte; pragmatisch gesehen haben sie aber eher die Funktion von Aufforderungen bzw. Beleidigungen. Oft ist diese Funktion sehr unklar, und sie kann nur der Redesituation entnommen werden. Pragmatische Mehrdeutigkeiten sind sehr wichtig, aber zum Glück treten sie in Texten, an denen man in der Sprachanalyse interessiert ist, selten auf. Wenn sie aber auftreten - wie dies zum Beispiel beim Dialog in natürlicher Sprache vorstellbar ist -, sind sie praktisch unlösbar.

Ich komme nun zum letzten und - gemessen an formalen Sprachen - vielleicht wichtigsten Merkmal natürlicher Sprachen, der *Kontextgebundenheit* oder *Kontextabhängigkeit.* Sie zeigt sich an zahlreichen Eigenheiten, von denen ich nur einige andeuten will. Besonders augenfällig sind die sogenannten deiktischen Elemente; das sind Wörter wie »ich«, »hier«, »jetzt«, »heute« und viele andere. Das Wort »ich« kann manchmal soviel bedeuten wie »Bundeskanzler«, nämlich wenn der Bundeskanzler sagt: »Ich begrüße Sie«; es kann bedeuten »Oberbürgermeister von Heidelberg«, nämlich wenn der Oberbürgermeister von Heidelberg sagt: »Ich begrüße Sie«; es hat, anders gesagt, allein in der Bundesrepublik rund 60 Millionen Bedeutungen, und was es bedeutet, hängt einfach davon ab, wer spricht. Das Wort »heute« bedeutet heute zum Beispiel den 14. September und morgen den 15. September, während das Wort »morgen« entsprechend heute den 15. September bedeutet. Man kann also sagen, daß morgen heute gestern ist. Was ein Wort in einem Satz bedeutet, hängt nicht nur von diesem Wort und diesem Satz ab, sondern auch davon, wer es äußert, wann es geäußert wird, wo es geäußert wird - denken Sie an das Wort »hier« -, zu wem es geäußert wird und vielen weiteren Faktoren der jeweiligen Sprechsituation oder, wie man auch sagt, des *Kontextes.*

Diese Kontextgebundenheit zeigt sich aber noch an vielen anderen Stellen. Wenn ein Satz z. B. heißt: »Die eine ist groß, die andere klein«, dann hängt seine genaue Bedeutung davon ab, wovon in diesem Kontext die Rede war: Je nachdem, ob von Frauen, Wahlkampfspenden, Kühen oder Schwierigkeiten der automatischen Sprachanalyse die Rede war, bedeutet der Satz ganz verschiedene Dinge, kann er wahr oder falsch sein. In sehr augenfälliger Weise äußert sich die Kontextgebundenheit auch bei Frage-Antwort-Folgen. Die normale Antwort auf eine Frage wie »Wo liegen denn die Gebäude der Fakultät für Chemie der Universität Heidelberg?« lautet nämlich: »Im Neuenheimer Feld«, nicht aber »Die Gebäude der Fakultät für Chemie der Universität Heidelberg liegen im Neuenheimer Feld«. Man läßt also im gegebenen Kontext normalerweise den ganzen ersten Teil der Antwort weg. Es soll zwar auch heute noch Lehrer geben, die von den Schülern verlangen, im ganzen Satz zu antworten, aber zum Glück halten sich die Schüler daran ebensowenig wie die Lehrer selber, sonst würden sie nämlich rasch zu Sprachgestörten. Die ständige, planvolle Verwendung von Ellipsen ist einer der wichtigsten Unterschiede einer natürlichen von einer formalen Sprache. Der Kontext stellt gewöhnlich die Gesamtbedeutung völlig klar.

Was nun wieder die Sprachanalyse betrifft, so ist die Kontextgebundenheit eine der größten Schwierigkeiten. Man muß hier allerdings etwas unterscheiden. Wenn sie darin besteht, daß auf unmittelbar vorhergehende sprachliche Äußerungen Bezug genommen wird - dies ist vor allem auch bei Personalpronomina wie »er, sie, es« der Fall -, dann gibt es gewisse Lösungsmöglichkeiten. Bezieht sich die Kontextgebundenheit hingegen auf gewisse allgemeine Züge der Sprechsituation wie Ort, Zeit, Sprecher und dergleichen, dann ist man am Ende. Zusammen mit der semantischen Mehrdeutigkeit ist die Kontextgebundenheit nicht nur eine praktische, sondern eine prinzipielle Barriere für den Einsatz der natürlichen Sprache in der Datenverarbeitung.

Alle hier genannten Eigenschaften sind nun keineswegs sporadisch zu beobachtende, störende Merkmale der natürlichen Sprache, sie sind vielmehr für sie konstitutiv. Wenn man sich dies vergegenwärtigt, scheint es vielleicht verständlich, weshalb ich es an früherer Stelle *beinahe ein Wunder* genannt habe, daß mit ihrer Hilfe die Kodifizierung und Vermittlung des Wissens so relativ problemlos möglich ist. Die Sprache ist vage, variabel, offen, mehrdeutig, kontextgebunden ; aber nicht die Wissenschaftler beklagen ihre Unzulänglichkeit, sondern die Dichter.

3 Wieso funktionieren Wissenskodifizierung und -Vermittlung durch die Sprache?

Ich will abschließend kurz zusammenfassen, wie es kommt, daß die natürliche Sprache einerseits ihrer Aufgabe in der Kodifizierung und Vermittlung des Wissens so relativ gut gerecht wird, sich aber andererseits einer automatischen Analyse und damit einem Einsatz in der Datenverarbeitung so weitgehend entzieht.

Der Grund liegt darin, daß unser Verständnis sprachlicher Äußerungen, etwa bestimmter Sätze (vgl. *Abbildung 2),* von mindestens vier Faktoren wesentlich abhängt. Es sind dies:

1. Ein Sprecher muß über etwas verfügen, was man als *Weltwissen* bezeichnen könnte, das heißt wissensmäßige Voraussetzungen, an die ein anderer Sprecher anknüpfen kann. Um einen Satz wie »Die Entscheidung des Bundesverfassungsgerichts zur Reform des § 218 stieß bei den Parteien auf eine un-

(1) »Drei Monate, ich nix abeite. Warum? Nix meine Papie guute. Ich Rathause nix gutt sprächän, un dann nix Papie.«

(Nach: Heidelberger Forschungsprojekt »Pidgin-Deutsch«: Sprache und Kommunikation ausländischer Arbeiter. Kronberg 1975. S.142)

(2) »Ein Versuch, vom Vorstellen des Seienden als solchen in das Denken an die Wahrheit des Seins überzugehen, muß, von jenem Vorstellen ausgehend, in gewisser Weise auch die Wahrheit des Seins noch vorstellen, so daß dieses Vorstellen anderer Art und schließlich als Vorstellen dem Zu-denkenden ungemäß bleibt.«

(M. Heidegger: Was ist Metaphysik? Tübingen 1965. S.18)

Abbildung 2: Für das *Verstehen* einer sprachlichen Äußerung ist es oftmals ganz unwichtig, ob die grammatischen Regeln genau befolgt werden oder nicht. So versteht man Text (1) - die Äußerung eines ausländischen Arbeiters - relativ gut, obwohl die Grammatik der deutschen Hochsprache nicht befolgt wird, Text (2) - eine kurze Textpassage eines Philosophen - nur sehr schwer, obwohl die Grammatik korrekt, einfach und überschaubar ist und in diesem Text eigentlich keine ungewöhnlichen Wörter vorkommen

terschiedliche Aufnahme« zu verstehen, genügt es nicht, die Bedeutung der Wörter »Partei«,»Paragraph«,»Aufnahme« usw. zu kennen. Man benötigt eine Menge an zusätzlichem Faktenwissen. Ein Mensch hat dies oder kann es sich erwerben und vermag dann den Satz zu verstehen. Eine DV-Anlage hat es nicht, und selbst wenn man eine Menge von Fakten einspeicherte, ist derzeit nicht zu sehen, wie sie es zur Interpretation anwenden können sollte.

2. Die zweite Komponente könnte man als *Situationswissen* bezeichnen. Damit sind Kenntnisse über die jeweilige Sprechsituation (z. B. Zeit, Ort, Sprecher, Hörer, deren Verhältnis zueinander) gemeint. Wenn man an das denkt, was vorhin über die Wörter »ich«, »hier«, »heute« usw. gesagt wurde, wird die Bedeutung dieses Faktors klar. Eine DV-Anlage verfügt nicht über die Möglichkeit, Informationen aus der Situation herauszuziehen, von Grenzfällen vielleicht abgesehen.

3. Die dritte Komponente möchte ich hier als *Textwissen* bezeichnen. Damit meine ich jene Information, die dem unmittelbaren sprachlichen Kontext entnommen ist, und die bei der Konstruktion von Sätzen eine besondere Rolle spielt. Ein typischer Fall sind die vorhin erwähnten Auslassungen bei Frage-Antwort-Folgen. Ein Sprecher knüpft in dem, was er sagt, gewöhnlich an das an, was unmittelbar zuvor - sei es von ihm, sei es von einem anderen - gesagt wurde; er kann damit rechnen, daß der Hörer dieses ebenfalls verfügbar hat, und entsprechend konstruiert er seine Sätze. Bei maschineller Analyse kann dies gewöhnlich nicht vorausgesetzt werden. Allerdings gibt es hier gewisse, wenn auch derzeit noch wenig genutzte Möglichkeiten, die Analyse über den einzelnen Satz hinaus auf ganze Texte auszudehnen und die dabei gewonnene Information aufzubewahren.

4. Die vierte und letzte Komponente schließlich ist der *Satz* selbst. Er steht natürlich der Maschine im Prinzip ebenso zur Verfügung wie dem Menschen. Aber er ist eben das einzige, was ihr zur Verfügung steht.

Aus dem Gesagten läßt sich leicht ableiten, weshalb die natürliche Sprache sich trotz aller Vagheit, Mehrdeutigkeit usw. so gut für eine Kodifizierung und Weitergabe des Wissens eignet, nicht aber für eine automatische Sprachanalyse. Sie trägt nur einen Teil der Last und ist so ausgelegt, daß zum Verständnis der Sätze eine Reihe weiterer Voraussetzungen erfüllt sein müssen. Sie rechnet nämlich mit dem Vorwissen der Sprecher in allgemeiner, situativer und textueller Hinsicht und kann darum vieles offen lassen. Dieses Vorwissen trägt so stark zum Verständnis bei, daß wir oft eine Äußerung gegen ihren Wortlaut richtig verstehen — etwa wenn sich jemand verspricht und wir sagen: »Du wolltest wohl dies und jenes sagen«, weil er nach unserem Vorwissen nur dieses gemeint haben konnte. Das ist auch einer der Gründe dafür, weshalb wir *ironische* Äußerungen verstehen, wo ja bekanntlich das Gegenteil von dem gesagt wird, was gemeint ist.

Ein Computer versteht keine Ironie. Er verfügt nicht über das entsprechende Weltwissen, Situationswissen, Textwissen. Man müßte es ihm vermitteln. Ob das gelingt, ist fraglich, aber nicht unwahrscheinlicher, als noch vor 50 Jahren die heutigen Computer jedem erschienen wären.