scientia nova

eine Bibliothek des modernen wissenschaftlichen Denkens

Bisher erschienen:

Karl H. BORCH, Wirtschaftliches Verhalten bei Unsicherheit

C. W. CHURCHMAN/R. L. ACKOFF/E. L. ARNOFF, Operations Research

Morton D. DAVIS, Spieltheorie für Nichtmathematiker

Richard C. JEFFREY, Logik der Entscheidungen

Norman MALCOLM, Ludwig Wittgenstein

Möglichkeiten und Maßstäbe für die Planung der Forschung

Oskar MORGENSTERN, Spieltheorie und Wirtschaftswissenschaft

Ernest NAGEL/James E. NEWMAN, Der Gödelsche Beweis

John von NEUMANN, Die Rechenmaschine und das Gehirn

E. Howard RAIFFA, Einführung in die Entscheidungstheorie

Hubert SCHLEICHERT, Elemente der physikalischen Semantik

Erwin SCHRÖDINGER, Was ist ein Naturgesetz?

Technikfolgen-Abschätzung

Hermann WEYL, Philosophie der Mathematik und Naturwissenschaft

Dean E. WOOLDRIDGE, Mechanik der Gehirnvorgänge

Dean E. WOOLDRIDGE, Mechanik der Lebensvorgänge

CLAUDE E. SHANNON
WARREN WEAVER

Mathematische Grundlagen der Informationstheorie

R. OLDENBOURG VERLAG MÜNCHEN WIEN 1976

Die englischsprachige Originalausgabe,
The Mathematical Theory of Communication,
ist erschienen bei University of Illinois Press.

Deutsche Übersetzung: Dipl. Ing. Helmut Dreßler

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Mathematische Grundlagen der Informationstheorie / Claude E. Shannon ; Warren Weaver.
(Scientia nova)
Einheitssacht.: The mathematical theory of communication (dt.).
ISBN 3-486-39851-2

NE: Shannon , Claude E. [Mitarb.]; Weaver , Warren [Mitarb.]; EST

ISBN 3-486-39851-2

Inhaltsverzeichnis

Vorwort

In den letzten Jahren konnte man in den USA und auch im Ausland eine bemerkenswerte Forschungsaktivität einiger Wissenschaftler auf dem Gebiet der Kommunikations-Theorie feststellen. Im Hinblick auf das weitgespannte Interesse gab Dean L. N. Ridenour die Anregung zu diesem Buch, das aus zwei Abhandlungen zum Thema Kommunikation besteht.

Der erste Aufsatz ist bisher in der vorliegenden Form noch nicht veröffentlicht worden, obwohl schon eine verkürzte Fassung im Juli 1949 in der Zeitschrift *Scientific American* erschienen ist. Zum Teil besteht er aus einer einführenden Darstellung der allgemeinen Theorie und dürfte von denen gern gelesen werden, die einen Überblick über das Thema erhalten wollen, bevor sie sich mit den mehr mathematischen Gesichtspunkten beschäftigen. Zusätzlich werden einige Anregungen zur breiteren Anwendung der fundamentalen Prinzipien der Kommunikations-Theorie gegeben.

Die zweite Arbeit ist ein Nachdruck aus dem *Bell System Technical Journal* vom Juli und Oktober 1948, kaum verändert, bis auf die Berichtigung weniger unbedeutender Druckfehler, ergänzt auch um einige weitere Literaturhinweise. Es ist daran gedacht, nachfolgende Entwicklungen auf diesem Gebiet in einem späteren Werk zu behandeln, das sich mit den mehr allgemeinen Aspekten der Informations-Theorie befaßt.

Wir freuen uns, Dean Ridenour, der dieses Buch ermöglichte, und dem Verlag "University of Illinois-Press" für die ausgezeichnete Zusammenarbeit danken zu können.

September, 1949

C. E. SHANNON
W. WEAVER

Ein aktueller Beitrag zur mathematischen Theorie der Kommunikation

von Warren Weaver

1. Einführung in die analytische Behandlung der Kommunikation [1)]

1.1 Kommunikation

Der Begriff der *Kommunikation* wird hier in einem sehr weitläufigen Sinn gebraucht, um alle Vorgänge einzuschließen, durch die ge-

[1)] Dieser Aufsatz hat drei Kapitel. Im ersten und dritten Kapitel stammen Idee und Ausführung von W. Weaver allein. Das zweite dazwischenliegende Kapitel "Kommunikationsprobleme der Ebene A" interpretiert die mathematischen Abhandlungen von Dr. Claude E. Shannon (Bell Telephone Laboratories). Dr. Shannons Arbeit geht, worauf J. von Neumann hingewiesen hat, zurück auf Bemerkungen von Boltzmann in einigen seiner Arbeiten zur statistischen Physik (1864), wonach die Entropie sich auf "fehlende Information" bezieht, und zwar insoweit, als sie die Anzahl von Alternativen betrifft, die für ein physikalisches System noch offen bleiben, nachdem alle makroskopisch beobachtbare, das System betreffende Information aufgezeichnet ist. L. Szilard (Z. f. Phys. Bd. 53, 1925) erweiterte diese Vorstellung noch zu einer generellen Diskussion über den Begriff der Information in der Physik, und J. von Neumann (*Math. Foundation of Quantum Mechanics*, Berlin 1932, Kap. 5) befaßte sich mit der Information in der Quantenmechanik und der Physik der Elementarteilchen. Dr. Shannons Arbeit ist jedoch enger mit einigen Gedanken verbunden, die vor etwa 20 Jahren von H. Nyquist und R. V. L. Hartley, beide von den Bell Laboratories, entwickelt worden sind; und Dr. Shannon selbst wies mit Nachdruck darauf hin, daß die Kommunikationstheorie viel von ihrem eigentlichen Gehalt Professor Norbert Wiener verdankt. Professor Wiener andererseits betont, daß Shannons frühe Arbeiten über Schaltvorgänge und mathematische Logik erschienen sind, noch bevor sein eigenes Interesse auf diesem Gebiet erwacht ist; und er fügt freundlicherweise hinzu, daß Shannon sicherlich für die selbständige Entwicklung solch grundlegender Aspekte zur Theorie wie die Einführung der Entropie die Ehre gebührt. Shannon hat sich natürlich besonders um die technische Anwendung der Kommunikationstheorie bemüht, während Wiener sich mehr mit biologischen Anwendungen (Erscheinungen im Zentralnervensystem) beschäftigte.

dankliche Vorstellungen einander beeinflussen können. Dies bezieht sich natürlich nicht nur auf die Sprache in Wort und Schrift, sondern auch auf Musik, Malerei, Theater und Ballet, eigentlich auf alles menschliche Verhalten. In manchem Zusammenhang erscheint es wünschenswert, eine noch umfassendere Definition des Begriffs Kommunikation zu verwenden, insbesondere wenn man Vorgänge mit einschließen will, durch die eine Maschine (z.B. ein Automat, der ein Flugzeug aufspürt und dessen wahrscheinliche zukünftige Position berechnet) eine andere Maschine beeinflußt (z.B. eine Lenkwaffe, die dieses Flugzeug verfolgt).

Diese Abhandlung wird sich oft scheinbar nur auf das spezielle, allerdings selbst ausgedehnte und wichtige Gebiet der sprachlichen Kommunikation beziehen; aber praktisch alle Aussagen lassen sich ebenso auf jede Art von Musik, auf ruhende oder bewegte Bilder, wie z.B. das Fernsehen, anwenden.

1.2 Kommunikationsprobleme in drei Ebenen

Bedingt durch das umfangreiche Gebiet der Kommunikation, scheint es Probleme in drei Ebenen zu geben. So erscheint es vernünftig, folgende drei Fragen zu stellen:

EBENE A. Wie genau können die Zeichen der Kommunkation übertragen werden? (Das technische Problem.)

EBENE B. Wie genau entsprechen die übertragenen Zeichen der gewünschten Bedeutung? (Das semantische Problem.)

EBENE C. Wie effektiv beeinflußt die empfangene Nachricht das Verhalten in der gewünschten Weise? (Das Effektivitätsproblem.)

Die *technischen Probleme* betreffen die Genauigkeit der Übertragung vom Sender zum Empfänger von Zeichenfolgen (geschriebene Sprache) oder von kontinuierlich sich ändernden Signalen (telefonische oder drahtlose Übertragung von Stimme oder Musik) oder von kontinuierlich sich ändernden zweidimensionalen Mustern (Fernsehen) usw. Mathematisch gesehen bedeutet das erste die Übertragung einer endlichen Menge von diskreten Zeichen, das zweite die Übertragung einer stetigen Funktion der Zeit und das dritte die Übertragung von mehreren kontinuierlichen Funktionen der Zeit oder von einer kontinuierlichen Funktion der Zeit und von zwei Raumkoordinaten.

Die *semantischen Probleme* betreffen die völlige Übereinstimmung oder genügend gute Näherung der Interpretation der Nachricht beim Empfänger, verglichen mit der vom Sender gewünschten Bedeutung. Dies ist eine sehr schwierige und verwickelte Situation, sogar wenn man sich nur mit den relativ einfachen Problemen der Kommunikation durch die Sprache befaßt.

Eine grundsätzliche Komplikation wird durch folgendes Beispiel anschaulich: Wenn Herr X nicht zu verstehen scheint was Herr Y sagt, so ist es theoretisch nicht möglich, daß diese Situation, solange Herr Y weiterhin nur mit Herrn X redet, in einer endlichen Zeit geklärt werden kann. Wenn Herr Y sagt: "Verstehen Sie mich jetzt?", und Herr X erwidert: "Ja, natürlich", so ist dies nicht unbedingt ein Beweis dafür, daß die Verständigung erreicht wurde. Es kann ja möglich sein, daß Herr X die Frage nicht verstanden hat. Falls dies lächerlich klingt, so stellen Sie sich einmal folgenden Dialog vor: "Czy pañ mnie rozumie?", und der Antwort: "Hai wakkate imasu". Ich glaube, daß diese grundsätzliche Schwierigkeit[2]), zumindest auf dem beschränkten Gebiet der sprachlichen Kommunikation, auf eine annehmbare Größe verringert (aber niemals ganz vermieden) werden kann durch Erklärungen, die (a) vermutlich nie mehr als eine Näherung an die zu erklärenden Gedanken sind, die aber (b) verständlich sind, da sie in einer Sprache formuliert sind, die vorher durch Gesten und Handlungen einigermaßen klar vermittelt worden ist. Zum Beispiel dauert es nicht lange, das Zeichen für "Ja" in irgendeiner Sprache durch eine Handlung verständlich zu machen.

Das semantische Problem ist weit verzweigt, falls man an die Kommunikation im allgemeinen denkt. Man stelle sich zum Beispiel vor, welche Bedeutung eine amerikanische Wochenschau für einen Russen haben könnte.

Die *Effektivitätsprobleme* beziehen sich auf den Erfolg, mit dem die Nachricht, die dem Empfänger übermittelt wurde, zu einem vom Sender beabsichtigten Verhalten führt. Auf den ersten Blick mag die Annahme, es sei der Zweck aller Kommunikation, das Verhal-

[2]) "Als Pfungst 1911 nachwies, daß die Pferde von Elberfeld, die erstaunliche sprachliche und mathematische Fähigkeiten zeigten, lediglich auf die Kopfbewegungen ihres Dompteurs reagierten, begegnete ihr Eigentümer, Herr Krall, dieser Kritik auf eine sehr direkte Art. Er fragte die Pferde, ob sie solch kleine Bewegungen überhaupt erkennen könnten, worauf sie nachdrücklich mit 'Nein' antworteten. Leider können wir nicht alle so sicher sein, daß unsere Fragen verstanden werden oder daß wir so deutliche Antworten erhalten." Siehe K. S. Lashley, "Persistent Problems in the Evolution of Mind" in *Quarterly Review of Biology*, Bd. 24, März 1949, S. 28.

ten des Empfängers zu beeinflussen, als eine unerwünschte Beschränkung erscheinen. Mit einer einigermaßen weiten Auslegung des Begriffs Verhalten ist es jedoch klar, daß Kommunikation entweder das Verhalten beeinflußt oder aber ohne irgendeine ersichtliche und wahrscheinliche Wirkung bleibt.

Das Problem der Effektivität führt zu ästhetischen Betrachtungen im Fall der schönen Künste. Im Fall der geschriebenen oder gesprochenen Sprache werden Überlegungen notwendig, die von der bloßen Technik des Stils über all die psychologischen und emotionalen Gesichtspunkte der Propagandatheorie bis hin zu jenen Wertvorstellungen reichen, die nötig sind, um den Worten "Erfolg" und "beabsichtigt" eine nützliche Bedeutung zu verleihen; diese Worte wurden hier im ersten Satz des Abschnitts über die Effektivität benutzt.

Das Effektivitätsproblem ist eng mit dem semantischen Problem verbunden und überschneidet sich mit ihm auf eine schwer bestimmbare Art; tatsächlich bestehen Überschneidungen zwischen all den hier vorgeschlagenen Kategorien von Problemen.

1.3 Kommentar

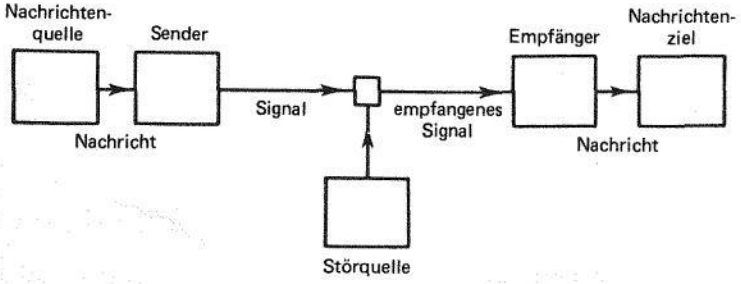
Nach der vorangegangenen Darstellung könnte man zu der Annahme verleitet werden, die Ebene A stelle ein relativ oberflächliches Problem dar, da es nur die technischen Details eines guten Entwurfs für ein Kommunikationssystem betrifft, während die Ebenen B und C den meisten, wenn nicht den gesamten philosophischen Gehalt der Kommunikationstheorie einschließen.

Die mathematische Theorie für die technischen Aspekte der Kommunikation, wie sie hauptsächlich von Claude Shannon bei den Bell Telephone Laboratories entwickelt wurde, bezieht sich in der Tat zunächst nur auf das Problem A, nämlich das technische Problem, wie genau verschiedene Arten von Signalen vom Sender zum Empfänger übertragen werden. Aber die Theorie hat, wie ich glaube, eine tiefe Bedeutung, die eine Unterbewertung der Ebene A nicht gerechtfertigt erscheinen läßt. Ein Teil der Bedeutsamkeit der neuen Theorie kommt daher, daß auf den Ebenen B und C nur von dem Grad der Signalgenauigkeit Gebrauch gemacht werden kann, wie er auf der Ebene A analysiert wurde. So wirkt sich jede Beschränkung, die in der Theorie der Ebene A entdeckt wird, auch auf die Ebenen B und C aus. Die weitaus größere Bedeutsamtkeit der Ebene A ergibt sich jedoch dadurch, daß die Analyse ihrer Probleme eine stärkere Überlappung dieser Ebene mit den anderen beiden offenbart als man sich als Laie vorzustellen vermag. Dadurch ist die Theorie der Ebene A zumindest in einem bedeutsamen Grad auch eine Theorie der Ebenen B und C. Ich hoffe, daß die folgenden Kapitel dieser Abhandlung diesen Kommentar verständlich machen und rechtfertigen werden.

2. Kommunikationsprobleme der Ebene A

2.1 Ein Kommunikationssystem und seine Probleme

Das betrachtete Kommunikationssystem kann symbolisch wie folgt dargestellt werden:



Die *Nachrichtenquelle* wählt aus einer Menge von möglichen Nachrichten eine gewünschte *Nachricht* aus (dies ist eine besonders wichtige Feststellung, die später einer umfangreichen Erklärung bedarf). Die ausgewählte Nachricht kann aus geschriebenen oder gesprochenen Worten oder aus Bildern, aus Musik usw. bestehen.

Der *Sender* übersetzt diese *Nachricht* in das *Signal*, welches dann über den *Übertragungskanal* vom Sender zum *Empfänger* übertragen wird. Im Fall der Telefonie ist der Kanal ein Draht, das Signal ein sich ändernder elektrischer Strom in diesem Draht; der Sender ist die Anlage (Telefonapparat usw.), die den Schalldruck der Stimme in einen sich ändernden elektrischen Strom übersetzt. In der Telegrafie verschlüsselt der Sender die geschriebenen Worte in Folgen von unterbrochenen Strömen verschiedener Länge (Punkte, Striche, Zwischenräume). Bei der gesprochenen Sprache ist die Nachrichtenquelle das Gehirn, der Sender sind die Stimmbänder, die den sich ändernden Schalldruck (das Signal) erzeugen, der durch die Luft (den Kanal) übertragen wird. In der Funktechnik ist der Kanal einfach der Raum (oder der Äther, wenn jemand dieses veraltete

und irreführende Wort noch bevorzugt), und das Signal ist die elektromagnetische Welle, die übertragen wird.

Der *Empfänger* ist eine Art umgekehrter Sender, der das übertragene Signal in eine Nachricht zurückverwandelt und diese Nachricht an das Ziel weitergibt. Wenn ich zu Ihnen spreche, ist mein Gehirn die Nachrichtenquelle und das Ihre das Ziel; meine Stimmbänder sind der Sender und Ihre Ohren und die damit verbundenen Gehörnerven sind der Empfänger.

Während des Übertragungsprozesses werden leider meistens dem Signal bestimmte Dinge hinzugefügt, die von der Nachrichtenquelle nicht beabsichtigt waren. Diese unerwünschten Zusätze können Tonverzerrungen sein (in der Telefonie z. B.) oder atmosphärische Störungen (in der Funktechnik) oder Verzerrungen der Form oder des Schattens eines Bildes (Fernsehen) oder Übertragungsfehler (Telegrafie oder Bildfunk) usw. Alle diese Veränderungen im übertragenen Signal werden *Störungen* genannt.

Wenn man ein solches Kommunikationssystem betrachtet, sollte man sich folgende Art von Fragen stellen:

a. Wie mißt man den *Betrag der Information*?

b. Wie mißt man die *Kapazität* eines Übertragungskanals?

c. Die Übersetzung der Nachricht in das Signal durch den Sender beinhaltet oft einen *Codiervorgang*. Was sind die charakteristischen Merkmale eines effizienten Codiervorganges? Und wenn die Codierung so effizient wie möglich ist, mit welcher Übertragungsrate kann der Kanal Information weiterleiten?

d. Was sind die allgemeinen Merkmale der *Störungen*? Wie beeinflussen Störungen die Genauigkeit der Nachricht, die schließlich das Ziel erreicht? Wie kann man die unerwünschten Effekte der Störungen auf ein Minimum beschränken und bis zu welchem Grad können sie ausgeschaltet werden?

e. Wenn das zu übertragende Signal *kontinuierlich* ist (wie bei der gesprochenen Sprache oder der Musik) und nicht aus diskreten Zeichen besteht, wie beeinflußt diese Tatsache das Problem?

Wir werden nun ohne jeden Beweis und mit einem Minimum an mathematischen Ausdrücken die wichtigsten Ergebnisse darlegen, die Shannon erhalten hat.

2.2 Information

Das Wort *Information* wird in dieser Theorie in einem besonderen Sinn verwendet, der nicht mit dem gewöhnlichen Gebrauch verwechselt werden darf. Insbesondere darf *Information* nicht der Bedeutung gleichgesetzt werden.

Tatsächlich können zwei Nachrichten, von denen eine von besonderer Bedeutung ist, während die andere bloßen Unsinn darstellt, in dem von uns gebrauchten Sinn genau die gleiche Menge an Information enthalten. Dies meint Shannon zweifellos, wenn er sagt, daß "die semantischen Aspekte der Kommunikation unabhängig sind von den technischen Aspekten". Dies bedeutet aber nicht, daß die technischen Aspekte unabhängig sind von den semantischen.

Anders ausgedrückt: Information in der Kommunikationstheorie bezieht sich nicht so sehr auf das, was gesagt *wird*, sondern mehr auf das, was gesagt werden *könnte*. Das heißt, Information ist ein Maß für die Freiheit der Wahl, wenn man eine Nachricht aus anderen aussucht. Für den sehr einfachen Fall, daß man nur zwischen zwei möglichen Nachrichten zu wählen hat, legt man willkürlich fest, daß die Information, die mit dieser Situation verbunden ist, gleich Eins ist. Beachten Sie, daß es irreführend (wenn auch oft bequem) ist, zu sagen, daß die eine oder andere Nachricht eine Informationseinheit übermittelt. Der Begriff der Information läßt sich nicht auf eine einzelne Nachricht anwenden (wie es bei dem Begriff der Bedeutung möglich wäre), eher auf eine Situation als Ganzes, wobei eine Informationseinheit andeutet, daß man in dieser Situation einen Freiheitsgrad in der Wahl der Nachricht hat, den man vorteilhafterweise als Standard- oder Einheitsgröße betrachtet.

Die beiden Nachrichten, zwischen denen man sich bei einer solchen Wahl entscheiden muß, können vollkommen beliebig sein. Die eine könnte die König-Jakob-Bibelübersetzung sein und die andere könnte "Ja" lauten. Der Sender könnte diese beiden Nachrichten so verschlüsseln, daß "Null" das Signal für die erste, und "Eins" das Signal für die zweite Nachricht ist, oder so, daß ein geschlossener Kreis (Strom fließt) das Signal für die erste ist und ein offener Kreis (kein Strom fließt) das Signal für die zweite. So können die beiden Stellungen, offen und geschlossen, eines einfachen Relais den beiden Nachrichten entsprechen.

Um etwas genauer zu sein: Der Betrag der Information ist im einfachsten Fall definiert als der Logarithmus der Anzahl der Wahl-

möglichkeiten. Da es vorteilhafter ist, den Logarithmus [3]) zur Basis 2 zu verwenden, als den normalen oder Briggschen Logarithmus zur Basis 10, ist die Information, wenn es nur zwei Wahlmöglichkeiten gibt, proportional dem Logarithmus von 2 zur Basis 2. Dieser aber ist gerade Eins, so daß die Situation, in der zwei Wahlmöglichkeiten bestehen, charakterisiert ist durch die Informationseinheit, wie es bereits weiter oben erwähnt wurde. Diese Informationseinheit wird ein "bit" genannt;, wobei dieses Wort, das zuerst von John W. Tukey vorgeschlagen wurde, eine Abkürzung für "binary digit" ist. Zur Darstellung von Zahlen im Binärsystem benötigt man nur zwei Ziffern, nämlich 0 und 1; geradeso, wie man im Dezimalsystem, das als Basis die Zahl 10 verwendet, zehn Ziffern, von 0 bis 9, benötigt. Null und Eins können symbolisch benutzt werden, je eine von zwei Wahlen darzustellen, wie oben erwähnt, so daß das "binary digit" oder "bit" direkt verbunden ist mit der Situation der Binärwahl, die genau eine Informationseinheit darstellt.

Wenn man z.B. 16 alternative Nachrichten zur Verfügung hat, unter denen man frei wählen kann, dann sagt man, da $16 = 2^4$ und damit $\log_2 16 = 4$, daß diese Situation charakterisiert ist durch einen Informationsgehalt von 4 bit.

Anfänglich erscheint es zweifellos seltsam, daß eine Information durch den Logarithmus der Anzahl der Wahlmöglichkeiten definiert ist. Aber während der Darstellung der Theorie wird es mehr und mehr offensichtlich, daß das logarithmische Maß tatsächlich das Natürliche ist. Im Augenblick werden wir dafür nur eine Andeutung geben. Es wurde bereits erwähnt, daß ein simples Zweipunkt-Relais mit seinen beiden kennzeichnenden Stellungen, 0 und 1, die Situation für die Informationseinheit darstellen kann, in der nur zwei Nachrichten gewählt werden können. Wenn nun ein Relais die Informationseinheit darstellt, wieviel Information kann dann z.B. durch drei Relais dargestellt werden? Es erscheint vernünftig, wenn man sagen möchte, daß drei Relais dreimal soviel Information darstellen können wie eines. Eben dieses erreicht man aber, wenn man die logarithmische Definition der Information benutzt. Denn für drei Relais ist es möglich, 2^3 oder 8 Wahlmöglichkeiten darzustellen, die symbolisch geschrieben werden können als 000, 001, 010, 011, 100, 101, 110, 111. Im ersten Fall sind alle drei Relais geöffnet, im letzten sind alle drei Relais geschlossen. Der duale Logarithmus von 2^3 ist 3, so daß das logarithmische Maß für diese Situation drei

[3]) Wenn gilt $m^x = y$, dann sagt man, x sei der Logarithmus von y zur Basis m.

Informationseinheiten anzeigt, genau wie es sein soll. Ebenso wird durch die Verdopplung der verfügbaren Zeit die Anzahl der möglichen Nachrichten quadriert und der Logarithmus verdoppelt und folglich wird die Information nur dann auch verdoppelt, wenn sie logarithmisch gemessen wird.

Die bisherigen Bemerkungen beziehen sich auf künstlich vereinfachte Situationen, bei denen die Nachrichtenquelle nur zwischen verschiedenen festgelegten Nachrichten wählen kann - wie ein Mensch, der aus einer Anzahl von Schmuckblatt-Telegrammen eines aussucht. In einer natürlicheren und wichtigeren Situation nimmt die Nachrichtenquelle eine Folge von Auswahlvorgängen aus einer Menge von elementaren Zeichen vor, wobei die ausgewählte Folge dann die Nachricht darstellt. So kann ein Mensch sich ein Wort nach dem anderen aussuchen und diese einzeln ausgewählten Worte dann aneinanderreihen, um eine Nachricht zu bilden.

An diesem Punkt zieht eine wichtige Betrachtung, die bisher im Hintergrund geblieben ist, die Aufmerksamkeit auf sich, nämlich die Rolle, die die Wahrscheinlichkeit in der Erzeugung einer Nachricht spielt. Denn wenn die aufeinanderfolgenden Zeichen ausgewählt werden, ist diese Auswahl, zumindest vom Standpunkt des Kommunikationssystems aus, von Wahrscheinlichkeiten bestimmt und diese Wahrscheinlichkeiten sind nicht einmal unabhängig voneinander, sondern hängen in jedem Stadium des Vorgangs von den vorangegangenen Auswahlergebnissen ab. Wenn z.B. in einem englischen Text das zuletzt gewählte Zeichen ein "the" ist, ist die Wahrscheinlichkeit, daß das nächste Wort auch ein Artikel oder ein anderes Wort als ein Hauptwort ist, sehr gering. Dieser Einfluß der Wahrscheinlichkeit erstreckt sich sogar über mehr als zwei Worte. Nach den drei Worten "in the event" ist die Wahrscheinlichkeit für "that" (deutsch: für den Fall - daß) als nächstes Wort verhältnismäßig groß und für "elephant" als nächstes Wort sehr klein.

Daß es Wahrscheinlichkeiten gibt, die einen gewissen Einfluß auf die englische Sprache ausüben, wird auch offensichtlich, wenn man z.B. bedenkt, daß ein englisches Wörterbuch kein einziges Wort enthält, in welchem dem Anfangsbuchstaben "j" ein b, c, d, f, g, j, k, l, q, r, t, v, w, x oder z folgt; damit ist die Wahrscheinlichkeit, daß nach dem Anfangsbuchstaben j einer dieser Buchstaben folgt, praktisch Null. Ebenso wird jeder zustimmen, daß die Wahrscheinlichkeit für eine Wortfolge wie "Constantinople fishing nasty pink" (= Konstantinopel angeln widerlich rosa) sehr gering ist. Sie ist zwar

klein, aber nicht gleich Null, denn es ist durchaus ein Text vorstellbar, in dem ein Satz schließt mit "Constantinople fishing" und der nächste beginnt mit "Nasty pink". Und nebenbei bemerkt, dieser unwahrscheinliche Satz aus vier Worten *ist* in einem normalen englischen Text vorgekommen, nämlich in dem hier geschriebenen.

Ein System, das eine Folge von Zeichen (die natürlich genausogut Buchstaben oder Noten, eher noch als Worte, sein können) erzeugt, denen gewisse Wahrscheinlichkeiten zukommen, wird ein *stochastischer Prozeß* genannt, und der spezielle Fall eines stochastischen Prozesses, in dem die Wahrscheinlichkeiten von den vorhergehenden Ereignissen abhängig sind, bezeichnet man als *Markoff-Prozeß* oder Markoff-Kette. Unter den Markoff-Prozessen, die nach unserer Vorstellung Nachrichten erzeugen könnten, gibt es eine besondere Klasse, die für die Kommunikationstheorie in erster Linie von Bedeutung ist; man bezeichnet sie als *ergodische Prozesse*. Die analytisch-mathematischen Einzelheiten dafür sind so kompliziert, die Beweisführung so tiefreichend und umfassend, daß es der größten Anstrengungen der besten Mathematiker bedurfte, um dafür eine geschlossene Theorie zu schaffen; allerdings ist die grobe Struktur eines ergodischen Prozesses ohne Mühe zu verstehen. Es handelt sich um einen Vorgang, in dem eine Zeichenfolge erzeugt wird, die vielleicht der Traum eines Demoskopen wäre, weil jede einigermaßen große Probe danach strebt, repräsentativ für die Folge als Ganzes zu sein. Nehmen wir an, daß zwei Personen auf verschiedene Art und Weise Proben aussuchen und überprüfen, welcher Trend sich in ihren statistischen Eigenschaften ergibt, sobald die Proben größer werden. Wenn es eine ergodische Situation ist, stimmen jene beiden Personen, wie immer sie ihre Proben auch gewählt haben mögen, in ihren Schätzungen über die Eigenschaften des Ganzen überein. Ergodische Systeme zeigen, anders ausgedrückt, eine besonders sichere und bequeme Art von statistischer Regelmäßigkeit.

Kehren wir nun wieder zurück zum Begriff der *Information*. Wenn wir eine Nachrichtenquelle haben, die eine Nachricht erzeugt, indem sie fortlaufend einzelne diskrete Zeichen auswählt (Buchstaben, Worte, Noten, Zeichen bestimmter Größe usw.), wobei die Wahrscheinlichkeit für die Wahl der verschiedenen Zeichen in jedem Stadium des Prozesses abhängig ist von der vorhergegangenen Auswahl (d.h. ein Markoff-Prozeß), was läßt sich dann über die Information sagen, die bei diesem Vorgang entsteht?

Die Größe, die in einzigartiger Weise den natürlichen Anforderun-

gen genügt, die man an die "Information" stellt, ist genau jene, die in der Thermodynamik als *Entropie* bekannt ist. Sie wird als Funktion der verschiedenen hier vorkommenden Wahrscheinlichkeiten ausgedrückt - der Wahrscheinlichkeit, einen bestimmten Zustand in einem Nachrichten erzeugenden Prozeß zu erreichen, und der Wahrscheinlichkeiten, daß, wenn dieser Zustand erreicht ist, bestimmte Zeichen als nächste ausgewählt werden. Außerdem enthält der Ausdruck den *Logarithmus* der Wahrscheinlichkeiten, so daß die Entropie eine natürliche Verallgemeinerung des logarithmischen Maßes ist, von dem wir weiter oben im Zusammenhang mit einfachen Fällen gesprochen haben.

Für jene, die Physik studiert haben, wird es besonders bemerkenswert sein, daß ein der Entropie ähnlicher Ausdruck in der Theorie als ein Maß der Information erscheint. Vor mehr als 100 Jahren von Clausius eingeführt, eng verbunden mit dem Namen Boltzmanns, und dadurch, daß Gibbs in seinem klassischen Werk über statistische Mechanik ihr eine tiefe Bedeutung gegeben hat, ist "Entropie" zu einem so grundsätzlichen und beherrschenden Begriff geworden, daß Eddington bemerkt: "Das Gesetz, daß die Entropie immer wächst - der zweite Hauptsatz der Thermodynamik - besitzt, glaube ich, die überragende Stellung unter den Naturgesetzen."

In der Physik ist die Entropie ein Maß für die Zufälligkeit oder "Vermischtheit" einer Situation; und die Tendenz der physikalischen Systeme, weniger und weniger organisiert, immer perfekter "vermischt" zu werden, ist so grundsätzlich, daß Eddington behauptet, daß in erster Linie diese Tendenz der Zeit ihre Richtung gibt - uns also zeigen würde, ob ein Film der physikalischen Welt vorwärts oder rückwärts läuft.

Daher hat jeder, der den Begriff der Entropie in der Kommunikationstheorie kennenlernt, das Recht, ziemlich aufgeregt zu sein - das Recht, anzunehmen, daß er auf etwas gestoßen sei, was sich als grundsätzlich und wichtig herausstellen könnte. Daß die Information durch die Entropie gemessen wird, ist letzten Endes natürlich, wenn wir uns erinnern, daß Information in der Kommunikationstheorie mit der Größe der Wahlfreiheit, die wir bei der Konstruktion der Nachricht haben, zusammenhängt. Daher kann man auch für eine Nachrichtenquelle sagen, ebenso wie man es von einem thermodynamischen System feststellen würde: "Diese Situation ist klar organisiert, sie ist nicht durch ein großes Maß an Zufälligkeit oder Auswahlmöglichkeit charakterisiert - das heißt, die Information (oder die Entropie) ist niedrig." Wir werden später darauf zurückkommen, denn wenn ich micht nicht sehr irre, ist dies ein wichtiger Aspekt der allgemeineren Bedeutung dieser Theorie.

Nachdem man die Entropie (oder die Information oder die Wahlfreiheit) einer bestimmten Nachrichtenquelle berechnet hat, kann man diese mit dem maximalen Wert vergleichen, den die Entropie unter der einzigen Bedingung, daß die Quelle weiterhin dieselben Zeichen verwendet, haben könnte. Das Verhältnis der tatsächlichen zur maximalen Entropie wird *relative Entropie* der Quelle genannt. Wenn die relative Entropie einer bestimmten Quelle zum Beispiel 0,8 ist, heißt dies, grob gesagt, daß diese Quelle in der Wahl ihrer Zeichen, um eine Nachricht zu bilden, etwa zu 80% so frei ist, wie es mit diesen bestimmten Zeichen überhaupt möglich ist. Zieht man die relative Entropie von 1 ab, so erhält man die *Redundanz* (Überfluß). Dies ist der Teil des Aufbaus der Nachricht, der nicht durch die Wahlfreiheit der Quelle bestimmt wird, sondern eher von angenommenen statistischen Regeln, die den Gebrauch der fraglichen Zeichen bestimmen. Es ist vernünftig, ihn Redundanz zu nennen, denn dieser Teil der Nachricht ist tatsächlich im üblichen Sinn des Wortes überflüssig; das heißt, dieser Teil der Nachricht ist unnötig (Wiederholung, überflüssig, eben redundant) in jenem Sinne, daß, wenn er nicht vorhanden wäre, die Nachricht immer noch im wesentlichen vollständig wäre oder doch vervollständigt werden könnte.

Es ist sehr interessant festzustellen, daß die Redundanz der englischen Sprache etwa 50% beträgt [4]), so daß etwa die Hälfte der in Sprache oder Schrift ausgewählten Buchstaben oder Wörter von uns frei gewählt worden ist, und die andere Hälfte (obwohl wir uns dessen gewöhnlich nicht bewußt sind) in Wirklichkeit durch die statistische Struktur der Sprache bestimmt ist. Abgesehen von wichtigeren Feststellungen, die wir wieder bis zu unserer Abschlußdiskussion zurückstellen wollen, ist es interessant zu erwähnen, daß eine Sprache in der Wahl ihrer Buchstaben mindestens zu 50% wirkliche Freiheit besitzen muß, wenn es möglich sein soll, zufriedenstellende Kreuzworträtsel zu konstruieren. Wenn vollständige Freiheit bestünde, wäre jede Anordnung von Buchstaben ein Kreuzworträtsel. Wenn nur eine Freiheit von 20% bestünde, wäre es unmöglich, so viele und so komplizierte Kreuzworträtsel herzustellen, daß das Spiel populär würde. Shannon hat geschätzt, daß es möglich wäre,

[4]) Die Schätzung von etwa 50 % gilt nur für die statistischen Strukturen aus bis zu acht Buchstaben, so daß der endgültige Wert wahrscheinlich etwas höher liegt.

dreidimensionale Kreuzworträtsel zu konstruieren, wenn die englische Sprache eine Redundanz von nur 30% besäße.

Bevor wir diesen Abschnitt über die Information abschließen, ist folgendes festzustellen: Auf der Ebene A wird Information durch alle statistischen Eigenschaften charakterisiert und betrifft nicht einzelne Nachrichten (und schon gar nicht die Bedeutung individueller Nachrichten). Der wahre Grund dafür ist, daß vom technischen Standpunkt aus ein Kommunikationssystem jede Nachricht verarbeiten können soll, die die Quelle erzeugen kann. Wenn es nicht möglich oder nicht praktikabel ist, ein System zu entwerfen, das alles perfekt verarbeiten kann, dann sollte das System so entworfen werden, daß es alle die Arbeiten erledigen kann, die am wahrscheinlichsten anfallen, und man sollte sich damit abfinden, daß es nicht effizient genug für seltenere Aufgaben ist. Diese Überlegungsweise führt sofort zu der Notwendigkeit, die statistische Natur der Gesamtmenge von Nachrichten zu charakterisieren, die eine gegebene Klasse von Quellen erzeugen kann und wird. Der Begriff der *Information*, wie er in der Kommunikationstheorie verwendet wird, berücksichtigt genau dieses Problem.

Obwohl es ganz und gar nicht die Aufgabe dieser Abhandlung ist, sich mit mathematischen Details zu befassen, scheint es dennoch notwendig zu sein, ein möglichst gutes Verständnis der der Entropie ähnlichen Ausdrücke zur Informationsbestimmung zu erreichen. Betrachtet man als einen einfachen Fall eine Menge von n unabhängigen Zeichen oder eine Menge von n unabhängigen ganzen Nachrichten, deren Auswahlwahrscheinlichkeiten $p_1, p_2 \cdots p_n$ sind, so lautet der Ausdruck für die Information

$$H = -[p_1 \log p_1 + p_2 \log p_2 + \cdots + p_n \log p_n]$$

oder

$$H = -\Sigma p_i \log p_i.$$

Das Symbol Σ, wie es in der Mathematik gebraucht wird, deutet an, daß über alle Glieder wie dem typischen, als Definition gebrauchten, $p_i \log p_i$, aufsummiert werden muß [5]).

Dies erscheint etwas kompliziert, aber sehen wir uns an, wie sich dieser Ausdruck in einigen einfachen Fällen verhält.

Nehmen wir zunächst an, daß wir nur zwischen zwei möglichen Nachrichten wählen können, deren Wahrscheinlichkeiten dann für die eine p_1 und für die andere $p_2 = 1 - p_1$ sind. Wenn man für diesen Fall den numerischen Wert von H berechnet, so stellt sich heraus, daß H dann seinen größten Wert besitzt, nämlich Eins, wenn beide Nachrichten gleich wahrscheinlich sind, d.h. wenn $p_1 = p_2 = \frac{1}{2}$; wenn man also vollkommen frei ist in der Wahl der beiden Nachrichten. Sobald eine Nachricht wahrscheinlicher wird als die andere (z.B. p_1 größer als p_2), nimmt der Wert von H ab. Und wenn eine Nachricht sehr wahrscheinlich ist (z.B. p_1 fast Eins und p_2 fast Null), ist der Wert von H sehr klein (fast Null).

Im Grenzfall, wenn eine Wahrscheinlichkeit gleich Eins ist (Gewißheit) und alle anderen gleich Null sind (Unmöglichkeit), ist H gleich Null (überhaupt keine Ungewißheit, keine Wahlfreiheit und keine Information).

H ist also dann am größten, wenn die beiden Wahrscheinlichkeiten gleich sind (d.h. wenn man vollkommen frei und unbeeinflußt wählen kann), und wird zu Null, wenn die Wahlfreiheit verschwunden ist.

Die soeben beschriebene Situation ist tatsächlich typisch. Wenn es mehr als zwei Wahlmöglichkeiten gibt, ist H dann am größten, wenn die Wahrscheinlichkeiten der verschiedenen Möglichkeiten sich so weit gleichen, wie es die Umstände erlauben, d.h. wenn man so frei wie möglich wählen kann und so wenig wie möglich zu bestimmten Entscheidungen gezwungen wird, deren Wahrscheinlichkeiten dann überdurchschnittlich sind. Nehmen wir andererseits an, daß für eine Wahl die Wahrscheinlichkeit fast Eins ist und damit die Wahrscheinlichkeiten für die anderen Möglichkeiten fast Null sind. Dies ist ganz offensichtlich eine Situation, in der man stark zugunsten einer bestimmten Entscheidung beeinflußt wird und die daher eine geringe Wahlfreiheit aufweist. Der Wert von H ist in einem solchen Fall sehr klein - die Information (die Wahlfreiheit, die Ungewißheit) gering.

Wenn die Anzahl der Möglichkeiten feststeht, wie es bisher der Fall war, dann ist die Information um so größer, je mehr sich die Wahrscheinlichkeiten für die verschiedenen Möglichkeiten gleichen. Es gibt noch einen anderen wichtigen Weg, den Wert von H zu erhöhen, nämlich die Anzahl der Wahlmöglichkeiten zu erhöhen. Genauer gesagt, wenn sich die Wahrscheinlichkeiten in etwa gleichen, dann ist H um so größer, je mehr Möglichkeiten es gibt. Es entsteht mehr "Information", wenn man frei aus einer Menge von 50 Standard-Nachrichten auswählt, als wenn man frei aus einer Menge von 25 wählt.

[5]) Machen Sie sich keine Sorgen wegen des Minuszeichens. Jede Wahrscheinlichkeit ist eine Zahl kleiner als oder gleich 1, und die Logarithmen der Zahlen, die kleiner als Eins sind, sind negativ. Daher ist das Minuszeichen notwendig, damit H positiv wird.

2.3 Die Kapazität eines Übertragungskanals

Nach der Diskussion des vorhergehenden Abschnitts ist man nicht überrascht, daß die Kapazität eines Kanals nicht durch die Anzahl der *Zeichen*, die er übertragen kann, sondern eher durch die *Information*, die er überträgt, beschrieben wird. Oder besser, da der letzte Satz sehr schnell zu einer Fehlinterpretation des Wortes Information führen kann, die Kapazität eines Kanals wird beschrieben, durch seine Fähigkeit, das zu übertragen, was eine Quelle an Information erzeugt.

Wenn es eine einfache Quelle ist, bei der alle Zeichen die gleiche Zeitdauer haben (dies ist z.B. beim Fernschreiber der Fall), wenn die Quelle so beschaffen ist, daß jedes ausgewählte Zeichen einen Informationsgehalt von s bit hat (frei gewählt aus einer Zahl von 2^s Zeichen), und wenn der Kanal beispielsweise n Zeichen pro Sekunde übertragen kann, dann wird die Kapazität C des Kanals definiert zu ns bit pro Sekunde.

In einem allgemeineren Fall muß man die unterschiedlichen Längen der verschiedenen Zeichen berücksichtigen. Dadurch enthält der allgemeine Ausdruck der Kapazität eines Kanals den Logarithmus der Anzahl der Zeichen einer bestimmten Zeitdauer (dies führt natürlich die Idee der *Information* ein und entspricht für den einfachen Fall des vorhergehenden Abschnitts dem Faktor s) und außerdem enthält er noch die Zahl solcher Zeichen, die verarbeit werden (dies entspricht dem Faktor n im vorhergehenden Abschnitt). Daher gilt für den allgemeinen Fall, daß die Kapazität nicht die Anzahl der Zeichen mißt, die pro Sekunde übertragen werden, sondern eher die Informationsmenge, die pro Sekunde übertragen wird, wobei als Einheit bit pro Sekunde benutzt wird.

2.4 Codierung

Am Anfang wurde betont, daß der *Sender* die *Nachricht* annimmt und sie in etwas, das *Signal* genannt wird, umwandelt, wobei das letztere tatsächlich über den Kanal zum *Empfänger* gelangt. Im Fall der Telefonie wandelt der Sender lediglich die akustischen Signale in etwas um (in den sich ändernden elektrischen Strom im Telefondraht), das zwar eindeutig verschieden, aber ebenso eindeutig dem Ursprünglichen gleichwertig ist. Aber der Sender kann auch eine viel komplexere Operation an der Nachricht ausführen, um das Signal zu erzeugen. Er könnte z.B. eine geschriebene Nachricht aufnehmen und mit Hilfe eines Codes diese Nachricht in eine Folge von Zahlen verschlüsseln; diese Zahlen werden dann als das Signal über den Kanal gesendet.

Deshalb sagt man im allgemeinen, daß es die Aufgabe des Senders ist, die Nachricht zu *codieren*, und die Aufgabe des Empfängers, sie zu *decodieren*. Die Theorie gilt auch für sehr komplizierte Sender und Empfänger, z.B. solche mit "Gedächtnis", so daß die Art, wie ein bestimmtes Nachrichtenzeichen codiert wird, nicht nur von dem Zeichen selbst abhängt, sondern auch von den vorhergegangenen Nachrichtenzeichen und davon, wie diese codiert worden sind.

Wir sind nun in der Lage, den fundamentalen Lehrsatz, den diese Theorie entwickelt hat, für einen störungsfreien Kanal, der diskrete Zeichen überträgt, aufzustellen. Dieser Lehrsatz bezieht sich auf einen Übertragungskanal mit der Kapazität von C bit pro Sekunde, der Signale einer Quelle mit der Entropie von H bit pro Zeichen annimmt. Der Lehrsatz besagt, daß es mit einem gut ausgedachten Codiervorgang für den Sender möglich ist, Zeichen über den Kanal mit einer mittleren Übertragungsrate [6]) zu übertragen, die nahezu den Wert C/H hat, die aber nie, wie geschickt die Codierung auch sein mag, den Wert C/H vollständig annehmen kann.

Es ist nützlicher, die Bedeutung dieses Lehrsatzes etwas später zu besprechen, wenn wir den allgemeineren Fall, der Störungen mit einschließt, behandeln. Im Augenblick jedoch ist es wichtig, die kritische Rolle, die die Codierung spielt, zu erkennen.

Erinnern wir uns, daß die Entropie eines Vorgangs, der Nachrichten oder Signale erzeugt, durch den statistischen Charakter dieses Vorgangs bestimmt ist - durch die verschiedenen Wahrscheinlichkeiten für bestimmte Nachrichten sowie die durch diese Nachrichten bedingten Wahrscheinlichkeiten für die nächsten Zeichen. Die statistischen Eigenschaften von *Nachrichten* sind voll und ganz von dem Charakter der Quelle bestimmt. Aber der statistische Charakter des *Signals*, wie es dann tatsächlich durch den Kanal übertragen wird, und damit die Entropie im Kanal, ist sowohl von dem abhängig, was man in den Kanal einspeist, als auch von den Möglichkeiten des Kanals, verschiedene Signalsituationen zu verarbeiten. Zum

[6]) Wir erinnern uns, daß die Kapazität C sich auf die pro Sekunde übertragene Information bezieht und in bit pro Sekunde gemessen wird. Die Entropie H mißt hier die Information pro Zeichen, so daß der Quotient von C und H die Zeichen pro Sekunde mißt.

Beispiel sind in der Telegrafie Zwischenräume zwischen Punkten und Punkten, zwischen Strichen und Strichen und zwischen Punkten und Strichen notwendig, sonst könnte man die Punkte und Striche nicht erkennen.

Es stellt sich nun heraus, wenn ein Kanal bestimmten Bedingungen dieser Art unterliegt, die die Signalfreiheit beschränken, so gibt es bestimmte statistische Signaleigenschaften, die zu einer Signalentropie führen, die größer ist, als sie es für irgendeine andere statistische Signalstruktur wäre; in diesem wichtigen Fall ist die Signalentropie genauso groß wie die Kapazität des Kanals.

Ausgehend von diesen Überlegungen ist es nun möglich, die wirkungsvollste Art der Codierung genau zu charakterisieren. Es ist tatsächlich der Sender am besten, der die Nachricht in einer solchen Weise codiert, daß das Signal eben jene optimalen statistischen Eigenschaften besitzt, die dem verwendeten Kanal am besten entsprechen - die also tatsächlich die Signal- (oder man kann auch sagen Kanal-) Entropie maximiert und der Kapazität C des Kanals angleicht.

Diese Art der Codierung führt nach dem oben beschriebenen fundamentalen Lehrsatz zum größtmöglichen Wert der Übertragungsrate C/H. Aber für diesen Gewinn in der Übertragungsrate zahlt man einen Preis. Es stellt sich nämlich die vertrackte Sache heraus, daß man je idealer man die Codierung gestaltet, desto mehr Zeit für den *Codierprozeß* benötigt. Dem Dilemma wird einmal dadurch begegnet, daß in der Elektronik "lang" nur einen sehr kleinen Teil einer Sekunde bedeuten kann, zum anderen dadurch, daß man einen Vergleich zwischen dem Gewinn an Übertragungsrate und dem Zeitverlust durch die Codierung anstellt und sich daraus eine optimale Codierung bestimmt.

2.5 Störungen

Wie beeinflussen Störungen die Information? Information ist, daran sollten wir uns ständig erinnern, ein Maß für die Freiheit der Entscheidung, eine Nachricht auszuwählen. Je größer diese Wahlfreiheit und damit auch die Information ist, desto größer ist die Unsicherheit, ob die Nachricht, die wirklich gewählt wird, eine ganz bestimmte Nachricht ist. So gehen größere Wahlfreiheit, größere Unsicherheit, größere Information Hand in Hand.

Falls Störungen auftreten, enthält die empfangene Nachricht gewisse Verzerrungen, bestimmte Fehler oder fremde Bestandteile, die sicherlich die Behauptung rechtfertigen, die erhaltene Nachricht zeige wegen des Störeffekts eine größere Unsicherheit. Aber vermehrte Unsicherheit bedeutet vermehrte Information; dies klingt, als seien Störungen von Vorteil!

Es ist allgemein richtig, daß, wenn Störungen auftreten, das empfangene Signal eine größere Information aufweist - oder besser, das empfangene Signal ist aus einer größeren Menge von Möglichkeiten ausgewählt, als es beim Senden des Signals der Fall war. Dies zeigt deutlich die semantische Falle, in die man hineintappen kann, wenn man vergißt, daß "Information" hier in einer speziellen Bedeutung benutzt wird, die die Wahlfreiheit und damit die Ungewißheit mißt, welche Wahl man getroffen hat. Es ist deswegen möglich, daß das Wort Information eine gute oder eine schlechte Nebenbedeutung hat. Unsicherheit, die der Entscheidungsfreiheit des Senders entspricht, ist erwünschte Unsicherheit. Unsicherheit, die aus Fehlern oder durch den Einfluß von Störungen entsteht, ist unerwünschte Unsicherheit.

So zeigt sich der Spaßvogel, wenn einer behauptet, daß das empfangene Signal mehr Information enthalte. Ein Teil dieser Information ist unecht und unerwünscht und durch die Störungen hineingekommen. Um die nützliche Information aus dem empfangenen Signal zu erhalten, müssen wir diesen unechten Teil ausfiltern.

Bevor wir diesen Punkt ganz klären können, müssen wir einen kleinen Umweg gehen. Angenommen, man hat zwei Gruppen von Zeichen, einmal die Zeichen, die von einer Nachrichtenquelle zur Erzeugung einer Nachricht verwandt wurden, zum anderen die Signalzeichen, die man tatsächlich empfangen hat. Die Wahrscheinlichkeiten dieser beiden Zeichengruppen hängen zusammen, da die Wahrscheinlichkeit, ein bestimmtes Zeichen zu erhalten, offensichtlich davon abhängt, welches Zeichen gesendet wurde. Ohne jeden Fehler durch Störungen oder andere Ursachen würden die empfangenen Signale genau den gesendeten Nachrichtenzeichen entsprechen; sind aber Fehler vorhanden, so sind die Wahrscheinlichkeiten für empfangene Zeichen immer noch stark beeinflußt durch jene Wahrscheinlichkeiten, die genau, oder fast genau, den gesendeten Zeichen entsprechen.

In einem solchen Fall kann man nun eine Entropie berechnen, die das Verhältnis der einen Zeichengruppe zur anderen ausdrückt. Be-

trachten wir z. B. die Entropie einer Nachricht relativ zu dem Signal. Leider sind die damit verbundenen Ergebnisse nicht zu verstehen, ohne daß man etwas mehr ins Detail geht. Nehmen wir für den Augenblick an, daß ein bestimmtes Signalzeichen tatsächlich empfangen worden ist. Dann erhält jedes *Nachrichten*zeichen eine bestimmte Wahrscheinlichkeit - relativ groß für das Zeichen, das dem Empfangenen gleicht und für die, die dem empfangenen Zeichen ähnlich sind, und relativ klein für alle anderen. Unter Verwendung dieser Wahrscheinlichkeiten errechnet man einen vorläufigen Entropiewert. Dies ist die Nachrichtenentropie in bezug auf ein genau bekanntes empfangenes Signalzeichen. Unter einigermaßen guten Bedingungen ist ihr Wert niedrig, da die Wahrscheinlichkeiten sich nicht gleichmäßig auf die verschiedenen Fälle verteilen, sondern sich auf einen oder einige wenige Fälle konzentrieren. Ihr Wert wäre in jedem Fall Null, in dem es keinerlei Störungen gibt (siehe Seite 13), denn wenn das Signalzeichen bekannt ist, wären alle Wahrscheinlichkeiten der Zeichen Null, außer von einem (nämlich dem empfangenen Zeichen), dessen Wahrscheinlichkeit gleich Eins wäre.

Für jedes der empfangenen Zeichen kann eine solche vorläufige Entropie der Nachricht berechnet werden. Nachdem man sie alle bestimmt hat, bildet man einen Mittelwert, indem man jede dieser Entropien mit der Wahrscheinlichkeit des Signalzeichens, die bei der Berechnung angenommen wurde, wichtet. Entropien, die auf diese Weise errechnet werden, indem man zwei Zeichenmengen betrachtet, werden *relative Entropien* genannt. Die spezielle, soeben beschriebene, ist die Entropie der Nachricht relativ zu ihrem Signal, und Shannon hat diese auch *Äquivokation*[7]) genannt.

Aus dem Weg, auf dem diese Äquivokation berechnet wird, können wir ihre Bedeutung erkennen. Sie mißt die *mittlere Unsicherheit in der Nachricht, wenn das Signal bekannt ist*. Wenn es keine Störungen gäbe, dann bestünde keine Unsicherheit in bezug auf die Nachricht, wenn das Signal bekannt ist. Falls aber noch irgendeine Unsicherheit über die Nachricht bleibt, nachdem das Signal bekannt ist, dann muß diese unerwünschte Unsicherheit auf Störungen zurückzuführen sein.

Die Diskussion der letzten Abschnitte bezieht sich auf den Wert für "die mittlere Unsicherheit in der Nachrichtenquelle, wenn das Signal bekannt ist". Man würde den gleichen Wert erhalten, wenn man

[7]) Äquivokation: engl. *equivocation* = Zweideutigkeit

sich auf "die mittlere Unsicherheit im empfangenen Signal, wenn die Nachricht bekannt ist", bezieht. Diese letztgenannte Unsicherheit wäre natürlich ebenfalls Null, falls es keine Störungen gäbe.

Für die Beziehung zwischen diesen beiden Größen ist leicht zu beweisen, daß

$$H(x) - H_y(x) = H(y) - H_x(y)\,,$$

wobei $H(x)$ die Entropie oder Information der Nachrichtenquelle ist, $H(y)$ die Entropie oder Information des empfangenen Signals, $H_y(x)$ die Äquivokation oder die Unsicherheit in der Nachrichtenquelle, wenn das Signal bekannt ist, $H_x(y)$ die Unsicherheit im empfangenen Signal, wenn die gesendete Nachricht bekannt ist, bzw. der Teil der Information des empfangenen Signals, der auf Störungen zurückzuführen ist. Die rechte Seite dieser Gleichung ist die Nutzinformation, die trotz des Störeffekts ankommt.

Nun kann erklärt werden, was man unter der Kapazität C eines gestörten Kanals versteht. Sie ist so definiert, daß ihr Wert der maximalen Übertragungsrate (in bit pro Sekunde) entspricht, mit der Nutzinformation (d. h. totale Unsicherheit minus Störunsicherheit) über den Kanal übertragen werden kann.

Warum wird hier von einer "maximalen" Rate gesprochen? Was kann man tun, um diesen Wert kleiner oder größer zu machen? Die Antwort ist, daß man diesen Wert beeinflussen kann, indem man eine Quelle wählt, deren statistische Eigenschaften den natürlichen Beschränkungen des Kanals angepaßt sind. Das bedeutet, daß man die Übertragungsrate der Nutzinformation durch die richtige Codierung optimieren kann (siehe Seite 16-17).

Und nun wollen wir zum Schluß den fundamentalen Lehrsatz für einen gestörten Kanal betrachten. Nehmen wir an, der gestörte Kanal hat im eben erwähnten Sinne eine Kapazität C und wird von einer Nachrichtenquelle der Entropie $H(x)$ bit pro Sekunde gespeist; das empfangene Signal habe die Entropie $H(y)$ bit pro Sekunde. Wenn die Kanalkapazität C gleich oder größer als $H(x)$ ist, dann läßt sich, indem man entsprechende Codiersysteme entwirft, jede Nachricht der Quelle mit einem beliebig kleinen Fehler durch den Kanal übertragen. Wie klein Sie auch die Fehlerhäufigkeit festlegen mögen, es gibt immer einen Code, der die Forderungen erfüllt. Ist jedoch die Kanalkapazität C kleiner als $H(x)$, die Entropie der den Kanal speisenden Quelle, dann ist es unmöglich, Codierungen zu finden, die die Fehlerhäufigkeit beliebig klein machen.

Wie geschickt die Codierung auch sein mag, es wird immer nach Empfang des Signals eine unerwünschte (Stör-) Unsicherheit bleiben, wie die Nachricht tatsächlich lautet; und diese unerwünschte Unsicherheit - diese Äquivokation - wird immer gleich oder größer als $H(x) - C$ sein. Darüber hinaus gibt es immer mindestens einen Code, der es ermöglicht, die unerwünschte Unsicherheit über die Nachricht dem Wert $H(x) - C$ beliebig gut anzunähern.

Der wichtigste Aspekt ist natürlich der, daß das Minimum an unerwünschter oder falscher Unsicherheit nicht verringert werden kann, wie kompliziert oder angepaßt der Codierprozeß auch sein mag. Dieser wichtige Lehrsatz gibt eine genaue und fast aufregend einfache Beschreibung der größtmöglichen Verläßlichkeit, die ein gestörter Kommunikationskanal erreichen kann.

Eine praktische Konsequenz, die von Shannon hervorgehoben wird, sollte angemerkt werden. Da die englische Sprache zu etwa 50% redundant ist, wäre es möglich, bei der normalen Telegrafie etwa die Hälfte der Zeit durch eine passende Codierung einzusparen, *vorausgesetzt*, man könnte über einen ungestörten Kanal senden. Wenn Störungen im Kanal vorhanden sind, ist es jedoch ein echter Vorteil, Codierungen zu verwenden, die nicht die ganze Redundanz auslöschen, da die noch vorhandene Redundanz bei der Bekämpfung der Störungen hilft. Dies ist sehr einfach einzusehen, denn gerade durch die hohe Redundanz in der englischen Sprache hat man z. B. keine Mühe, eventuelle Fehler zu korrigieren, die bei der Übertragung der einzelnen Buchstaben entstanden sind.

2.6 Kontinuierliche Nachrichten

Bis jetzt haben wir uns mit Nachrichten beschäftigt, die sich aus diskreten Zeichen zusammensetzen, wie Worte aus Buchstaben bestehen, Sätze aus Worten, eine Melodie aus Tönen oder ein Rasterbild aus einer endlichen Zahl von Bildpunkten. Wie verändert sich die Theorie, wenn man kontinuierliche Nachrichten berücksichtigt, wie z. B. die Stimme mit ihrer sich kontinuierlich ändernden Tonhöhe und Lautstärke?

Grob gesagt wird die Theorie schwieriger und mathematisch komplizierter, aber nicht grundsätzlich anders. Viele der oben gegebenen Darstellungen für den diskreten Fall bedürfen keiner Korrektur, andere müssen nur leicht verändert werden.

Ein recht hilfreicher Umstand ist der folgende. In der Praxis ist man immer an einem kontinuierlichen Signal interessiert, das aus einfachen, harmonischen Bestandteilen *nicht aller Frequenzen* aufgebaut ist, sondern aus Frequenzen, die innerhalb eines Frequenzbandes von null bis, sagen wir, W Perioden pro Sekunde liegen. So kann, obwohl die menschliche Stimme höhere Frequenzen enthält, eine sehr befriedigende Verständigung über einen Telefonkanal erreicht werden, der nur Frequenzen bis etwa viertausend Hertz übertragen kann. Mit Frequenzen bis zu zehn- oder zwölftausend Hertz ist eine "high fidelity" Radioübertragung von sinfonischer Musik möglich usw.

Es gibt einen sehr nützlichen mathematischen Lehrsatz, der besagt, daß ein kontinuierliches Signal mit einer Dauer von T Sekunden und einer Bandbreite zwischen den Frequenzen von 0 bis W *vollständig beschrieben werden* kann, wenn man $2TW$ diskrete Werte angibt. Dies ist wirklich ein bemerkenswerter Lehrsatz. Normalerweise kann eine kontinuierliche Kurve durch eine endliche Anzahl von Punkten nur angenähert beschrieben werden, und für die vollständige Information über die Kurve ist im allgemeinen eine unendliche Anzahl von Punkten notwendig. Wenn aber die Kurve aus einfachen, harmonischen Bestandteilen einer begrenzten Anzahl von Frequenzen aufgebaut ist, wie ein komplexer Klang aus einer begrenzten Anzahl von reinen Tönen besteht, dann benötigt man nur eine endliche Anzahl von Parametern. Dies hat den wichtigen Vorteil, das Kommunikationsproblem für kontinuierliche Signale von einer komplizierten Situation, in der man mit einer unendlichen Anzahl von Variablen zu tun hätte, auf eine beträchtlich einfachere Situation zu reduzieren, in der man mit einer endlichen (allerdings großen) Anzahl von Variablen zu tun hat.

In der Theorie für den kontinuierlichen Fall wurden Formeln entwickelt, die die maximale Kapazität C eines Kanals mit der Frequenz-Bandbreite W beschreiben, wenn die mittlere *Sendeleistung* P ist, wenn außerdem der Kanal einer Störung mit der Leistung N unterliegt, wobei diese Störung ein "weißes thermisches Rauschen" einer bestimmten, von Shannon genau umschriebenen Art ist. Dieses weiße Rauschen hat selbst eine begrenzte Frequenz-Bandbreite, und die Amplituden der verschiedenen Frequenzbestandteile zeigen eine Normal- (oder Gauß-) Verteilung. Unter diesen Umständen erhält Shannon den bemerkenswert einfachen und allgemeinen Lehrsatz, daß es bei optimaler Codierung möglich ist, binäre Zeichen mit der Übertragungsrate von

$$W \log_2 \frac{P+N}{N}$$

bit pro Sekunde zu senden und dabei eine beliebig niedrige Fehlerhäufigkeit zu erreichen. Diese Rate aber kann unmöglich überschritten werden, unabhängig davon, wie gut die Codierung ist, ohne eine bestimmte Fehlerhäufigkeit in Kauf zu nehmen. Für den Fall von beliebigem Rauschen, anstatt des speziellen oben angenommenen "weißen Rauschens", kann Shannon keine eindeutige Formel für die Kanalkapazität herleiten; er erhält jedoch brauchbare obere und untere Grenzen für die Kanalkapazität. Und er leitet auch Grenzen für die Kanalkapazität ab, wenn man nicht von der durchschnittlichen Leistung des Senders ausgeht, sondern von einer kurzzeitigen Leistungsspitze.

Schließlich sollte bemerkt werden, daß Shannon Resultate erzielt, die notwendigerweise etwas weniger spezifisch, jedoch von offensichtlicher Tiefe und weitreichender Bedeutung sind; sie beschreiben in allgemeiner Weise für kontinuierliche Nachrichten oder Signale die Genauigkeit einer empfangenen Nachricht, den Begriff der Geschwindigkeit, mit der eine Quelle Nachrichten erzeugt, die Übertragungsrate und die Kanalkapazität, wobei diese alle von bestimmten Anforderungen an die Genauigkeit abhängen.

3. Die Wechselbeziehung zwischen den drei Ebenen der Kommunikationsprobleme

3.1 Einleitung

Im ersten Kapitel dieses Buches wurde vorgeschlagen, drei Ebenen anzunehmen, auf denen man das allgemeine Kommunikationsproblem betrachten kann. Man kann nämlich fragen:

EBENE A. Wie genau können die Zeichen der Kommunikation übertragen werden?

EBENE B. Wie genau entsprechen die übertragenen Zeichen der gewünschten Bedeutung?

EBENE C. Wie effektiv beeinflußt die empfangene Nachricht das Verhalten in der gewünschten Weise?

Es wurde angedeutet, daß die mathematische Theorie der Kommunikation, wie sie von Shannon, Wiener und anderen entwickelt wurde, insbesondere die von Shannon behandelte, zweifellos auf die Technik bezogene Theorie, obwohl diese scheinbar nur auf die Probleme der Ebene A anwendbar ist, tatsächlich hilfreich und anregend auch für die Probleme der Ebenen B und C ist.

Im 2. Kapitel betrachteten wir dann, wozu diese mathematische Theorie entwickelt wurde, welche Vorstellungen sie prägen, welche Ergebnisse sie gebracht hat. Es ist der Zweck dieses abschließenden 3. Kapitels, die Gedanken zusammenzufassen und zu sehen, bis zu welchem Grad und in welchen Grenzen die ursprüngliche Aussage gerechtfertigt war, wonach der auf der Ebene A erreichte Fortschritt auch zur Lösung der Probleme der Ebenen B und C beiträgt, und ob in der Tat die Wechselbeziehungen zwischen den drei Ebenen so beträchtlich sind, daß die Trennung in die drei Ebenen letzten Endes als künstlich und unerwünscht erscheint.

3.2 Grundzüge der Theorie auf der Ebene A

Die offensichtliche, erste Feststellung, und in der Tat die Feststellung, die die Hauptlast der Argumente trägt, ist die, daß die ma-

thematische Theorie äußerst allgemein in ihrer Reichweite ist, grundlegende Probleme behandelt und durch die Ergebnisse, die sie erzielt, von klassischer Einfachheit und Überzeugungskraft ist.

Es handelt sich um eine so allgemeine Theorie, daß man nicht zu sagen braucht, welche Art von Zeichen betrachtet werden - ob geschriebene Buchstaben oder Worte, ob Noten oder gesprochene Worte oder sinfonische Musik oder Bilder. Die Theorie ist genügend tiefgründig, daß die Beziehungen, die sie darlegt, unterschiedslos für all diese sowie für andere Formen der Kommunikation anwendbar sind. Das bedeutet natürlich, daß die Theorie so einfallsreich durchdacht ist, daß sie sich wirklich mit dem innersten Kern des Kommunikationsproblems befaßt - mit jenen Grundbeziehungen, die allgemein gelten, ohne Rücksicht darauf, welche speziellen Formen im aktuellen Fall vorkommen könnten.

Ein Nachweis für diese Allgemeingültigkeit ergibt sich daraus, daß zur Geheimsprache, die natürlich auch nur eine Form der Codierung ist, diese Theorie nicht nur einen wesentlichen Beitrag liefert, sondern tatsächlich ihre Grundlage ist. Auf ähnliche Art und Weise trägt die Theorie zum Problem der Übersetzung von einer Sprache in eine andere bei, obwohl hier nicht nur Probleme der Informationstheorie, sondern auch semantische Probleme auftreten. Ebenso sind die in diesem Werk entwickelten Gedanken so eng mit dem Problem der logischen Konstruktion großer Computer verbunden, daß es nicht überrascht, daß Shannon gerade jetzt eine Abhandlung über die Konstruktion eines Computers geschrieben hat, der dazu fähig wäre, vortrefflich Schach zu spielen. Außerdem bezieht sich der Schluß dieser Abhandlung direkt auf einen gegenwärtigen Wortstreit, indem er die Frage aufwirft, ob man entweder sagen müsse, ein solcher Computer "denkt", oder aber die übliche Bedeutung des Wortes "denken" grundlegend ändern müsse.

Zum zweiten erscheint es offensichtlich, daß durch die Struktur, auf der die gegenwärtige Theorie basiert, ein wichtiger Beitrag zu jeder möglichen allgemeinen Kommunikationstheorie geleistet wurde. Es erscheint zunächst selbstverständlich, ein Kommunikationssystem so darzustellen, wie es am Ausgangspunkt dieser Theorie geschehen ist; diese Beschränkung der Sachlage scheint jedoch sehr vernünftig und angemessen zu sein, überzeugt man sich, wie reibungslos und allgemein dieser Gesichtspunkt zu zentralen Ergebnissen führt. Mit hoher Wahrscheinlichkeit werden Überlegungen zur Kommunikation auf den Ebenen B und C Ergänzungen zu dem schematischen Diagramm auf Seite 7 erforderlich machen, aber es ist ebenso wahrscheinlich, daß die geforderten Ergänzungen minimal sind und keine wirkliche Korrektur darstellen.

Daher kann es, wenn man zu den Ebenen B und C übergeht, notwendig werden, die statistischen Eigenschaften des Nachrichtenziels zu berücksichtigen. Man kann sich als eine Ergänzung des Diagramms einen weiteren Block vorstellen, der die Aufschrift "semantischer Empfänger" trägt und der zwischen dem technischen Empfänger (der die Signale in Nachrichten umwandelt) und dem Ziel aufgestellt ist. Dieser semantische Empfänger unterwirft die Nachricht einer zweiten Decodierung, welche die statistisch-*semantischen* Eigenschaften den statistisch-semantischen Fähigkeiten der Gesamtheit der Empfänger anpassen soll oder jener Untermenge von Empfängern, die den Hörerkreis darstellen, den man beeinflussen will.

Auf ähnliche Weise kann man sich einen Block in dem Diagramm vorstellen, der, zwischen der Nachrichtenquelle und dem Sender eingebaut, die Aufschrift "semantische Störung" tragen würde, wobei der Block, der vorher einfach als Störung beschrieben wurde, jetzt die Aufschrift "technische Störung" trägt. Von dieser Quelle wird die Störung oder Entstellung der Bedeutung dem Signal aufgeprägt, welche von der Nachrichtenquelle nicht beabsichtigt ist, die jedoch unvermeidlich das Ziel beeinflußt. Das Problem der semantischen Decodierung muß eine solche semantische Störung in Betracht ziehen. Man könnte ebenso an eine Anpassung der ursprünglichen Nachricht denken, so daß die Summe von Nachrichtenbedeutung und semantischer Störung gleich der gewünschten Gesamtbedeutung der Nachricht am Ziel ist.

Drittens erscheint es für die Probleme auf allen drei Ebenen bedeutsam, daß Irrtümer und Verwirrungen entstehen und die Übertragungstreue abnimmt, wenn man, wie gut auch immer die Codierung sei, versucht, einen Kanal zu überlasten (d.h. $H > C$). Hier wieder sollte eine allgemeine Theorie auf allen drei Ebenen sicherlich nicht nur die Kapazität des Kanals, sondern auch (sogar die Worte sind richtig!) die Kapazität des Hörerkreises in Betracht ziehen. Versucht man die Kapazität des Publikums zu überlasten, so ist es wahrscheinlich - dies ist eine unmittelbare Analogie -, daß man sozusagen die Zuhörer überschwemmt und die Überschußmenge an Nachrichten verschwendet. Fast sicher fördert man bei einer Überforderung der Zuhörerschaft Irrtümer und eine allgemeine und unvermeidliche Verwirrung.

Viertens ist anzunehmen, daß die Beziehung zwischen Entropie und Information auch für die Betrachtungen auf den Ebenen B und C hilfreich ist und eine nützliche Orientierung für den Zugang zu jenen Problemen bietet.

Die Vorstellung von der Information, wie sie in dieser Theorie entwickelt wird, erscheint anfänglich enttäuschend und seltsam - enttäuschend, weil sie nichts mit Bedeutung zu tun hat, seltsam, weil sie sich nicht auf eine einzelne Nachricht bezieht, sondern eher auf die statistischen Eigenschaften einer Gesamtheit von Nachrichten, und seltsam auch, weil in den statistischen Ausdrücken die beiden Worte *Information* und *Unsicherheit* die gleiche Bedeutung haben.

Ich denke jedoch, dies sollten nur erste Reaktionen sein, und am Schluß sollte man sagen, daß diese Analyse die Sachlage so weit geklärt hat, daß man nun, vielleicht zum ersten Mal, für eine wirkliche Theorie der Bedeutung bereit ist. Eine technische Kommunikationstheorie ist geradeso wie eine gute und diskrete Postangestellte, die Ihre Telegramme annimmt. Sie achtet nicht auf die Bedeutung, ob sie nun traurig oder fröhlich oder unangenehm ist. Aber sie muß bereit sein, sich um alles zu kümmern, was auf ihren Schreibtisch kommt. Dieser Gedanke, daß ein Kommunikationssystem versuchen sollte, alle möglichen Nachrichten verarbeiten zu können, und daß man dazu die statistischen Eigenschaften der Quelle als Grundlage zu berücksichtigen hat, das ist sicherlich nicht ohne Bedeutung für die Kommunikation im allgemeinen. Eine Sprache muß ausgedacht (oder entwickelt) werden mit einem Blick auf die Gesamtheit der Dinge, die man sagen möchte; wenn sie auch nicht fähig ist, alle Anforderungen zu erfüllen, sollte sie sie doch so oft wie möglich und so gut wie möglich erfüllen. Das heißt, sie sollte ihre Aufgabe auf statistische Art lösen.

Der Begriff der Information, der mit einer Quelle verbunden ist, führt direkt, wie wir bereits gesehen haben, zu einer Studie der statistischen Struktur der Sprache; und diese Studie verschafft zum Beispiel einen Einblick in die englische Sprache, der für Studierende jeder Erscheinungsform von Sprache und Kommunikation sicher bedeutungsvoll sein kann. Die Idee, als wichtiges Hilfsmittel die Theorie des Markoff-Prozesses zu benützen, scheint besonders vielversprechend für semantische Studien, da diese Theorie besonders an die Behandlung des bezeichnendsten, aber auch schwierigsten Aspektes der Bedeutung angepaßt worden ist, nämlich dem Einfluß des Zusammenhangs. Man hat das unbestimmte Gefühl, daß Information und Bedeutung sich wie zwei kanonisch-konjugierte Variablen in der Quantentheorie verhalten; diese unterliegen einer gemeinsamen Beschränkung, wonach man von dem einen opfern muß, wenn man von dem anderen viel haben möchte.

Man kann die Bedeutung auch als Analogon zu einer der Größen ansehen, von der die Entropie eines thermodynamischen Ensembles abhängt. Das Auftreten der Entropie in der Theorie ist, wie bereits vorher erwähnt wurde, sicherlich sehr interessant und bedeutungsvoll. Eddington wurde in diesem Zusammenhang bereits zitiert, aber es gibt einen anderen Abschnitt in "The Nature of the Physical World", der besonders anregend und geeignet erscheint:

> "Angenommen wir würden gebeten, folgendes in zwei Kategorien einzuordnen - *Entfernung*, *Masse*, *elektrische Stromstärke*, *Entropie*, *Schönheit*, *Melodie*.
>
> Ich glaube, es gibt zwingende Gründe, Entropie Seite an Seite mit Schönheit und Melodie zu setzen und nicht mit den ersten drei. Entropie wird nur gefunden, wenn die Teile zusammenhängend betrachtet werden, und durch das Sehen oder Hören der Teile in ihrem Zusammenhang werden Schönheit und Melodie erkannt. Alle drei sind Formen der Ordnung. Es ist ein bedeutender Gedanke, daß einer dieser drei Ordnungsbegriffe in der Lage sein sollte, sich als eine allgemeingültige Größe der Wissenschaft darzustellen. Der Grund, warum dieser Unbekannte sich bei den Ureinwohnern der physikalischen Welt als einer der ihren ausgeben kann, ist der, daß er fähig ist, ihre Sprache zu sprechen, nämlich die Sprache der Arithmetik."

Ich bin sicher, daß Eddington gewillt gewesen wäre, das Wort *Bedeutung* mit Schönheit und Melodie zusammenzubringen; und ich nehme an, er wäre begeistert gewesen zu sehen, daß in seiner Theorie Entropie nicht nur die Sprache der Arithmetik spricht; sie spricht auch die Sprache der Sprache.

Die mathematische Theorie der Kommunikation

von Claude E. Shannon

Einführung

Die neueste Entwicklung verschiedener Modulationsmethoden wie PCM und PPM, die die Bandbreite gegen das Verhältnis Signal/Störung austauschen, hat das Interesse an einer allgemeinen Theorie der Kommunikation intensiviert. Eine Basis für solch eine Theorie ist in den Schriften von Nyquist [1] und Hartley [2], die dieses Gebiet betreffen, enthalten. In diesem Aufsatz werden wir die Theorie erweitern, um eine Anzahl neuer Faktoren einzuschließen, insbesondere die Wirkung von Störung im Kanal und die Einsparungen, die sowohl durch die statistische Struktur der Originalnachricht als auch durch die Art des Endzieles der Information möglich sind.

Das grundlegende Problem der Kommunikation besteht darin, an einer Stelle entweder genau oder angenähert eine Nachricht wiederzugeben, die an einer anderen Stelle ausgewählt wurde. Oft haben die Nachrichten *Bedeutung*, das heißt, sie beziehen sich auf gewisse physikalische oder begriffliche Größen oder sie befinden sich nach irgendeinem System mit diesen in Wechselwirkung. Diese semantischen Aspekte der Kommunikation stehen nicht im Zusammenhang mit den technischen Problemen. Der technisch bedeutungsvolle Aspekt ist, daß die tatsächliche Nachricht *aus einem Vorrat von möglichen Nachrichten ausgewählt* worden ist. Das System muß so entworfen werden, daß es für jede mögliche Nachricht funktioniert, nicht nur für die eine, die tatsächlich ausgewählt wird, da diese zum Zeitpunkt der Konstruktion noch unbekannt ist.

---

[1] H. Nyquist, "Certain Factors Affecting Telegraph Speed", *Bell System Technical Journal*, April 1924, S. 324; "Certain Topics in Telegraph Transmission Theory", A.I.E.E. Trans., Bd. 47, April 1928, S. 616.

[2] R. V. L. Hartley, "Transmission of Information", *Bell System Technical Journal*, Juli 1928, S. 535.

Falls die Anzahl der Nachrichten im Vorrat begrenzt ist, kann diese Anzahl oder jegliche monotone Funktion dieser Anzahl als ein Maß der Information angesehen werden, die erzeugt wird, wenn eine Nachricht aus dem Vorrat ausgewählt wird und die Wahrscheinlichkeit für jede Auswahl die gleiche ist. Wie von Hartley betont wurde, ist der natürlichste Maßstab dafür die logarithmische Funktion. Obwohl die Definition beträchtlich verallgemeinert werden muß, wenn wir den Einfluß der Statistik auf das Vorkommen der einzelnen Nachrichten berücksichtigen und wenn wir eine fortlaufende Reihe von Nachrichten haben, werden wir in allen Fällen ein im wesentlichen logarithmisches Maß benutzen.

Der logarithmische Maßstab ist aus folgenden Gründen vorteilhaft:

1. Er ist praktisch nützlicher. Parameter aus dem technischen Bereich wie zum Beispiel Zeit, Bandbreite, Anzahl der Relais usw. neigen dazu, sich mit dem Logarithmus der Anzahl von Möglichkeiten linear zu verändern. Wenn man z. B. einer Gruppe von Relais ein Relais hinzufügt, vordoppelt sich die Anzahl der insgesamt möglichen Zustände der Relais: Zum dualen Logarithmus dieser Zahl wird der Wert 1 addiert. Durch Verdopplung der Zeit wird die Anzahl der möglichen Nachrichten quadriert oder der Logarithmus verdoppelt usw.

2. Er ist unserem intuitiven Gefühl näher als dem eigentlichen Maß. Dies hängt eng mit (1) zusammen, da wir intuitiv Größen durch linearen Vergleich mit allgemeinen Standards messen. Man hat z. B. das Gefühl, daß zwei Lochkarten für die Informationsspeicherung die doppelte Kapazität von einer besitzen und daß zwei identische Kanäle für die Sendung von Information zusammen die doppelte Kapazität von einem besitzen.

3. Er ist mathematisch besser geeignet. Viele der Grenzwertprozesse sind in logarithmischer Ausdrucksweise einfach, würden jedoch, ausgedrückt durch die Anzahl der Möglichkeiten, eine schwerfällige Neuformulierung verlangen.

Die Wahl einer logarithmischen Basis entspricht der Wahl einer Maßeinheit für die Information. Wenn die Basis 2 benützt wird, können die sich ergebenden Einheiten "binary digits" (Binärziffern) genannt werden oder kürzer *bit*, ein Wort, das von J. W. Tukey vorgeschlagen wurde. Eine Vorrichtung mit zwei stabilen Positionen, wie ein Relais oder ein Flip-Flop, kann ein Informationsbit speichern. N solcher Vorrichtungen können N bit speichern, da die

Gesamtzahl von möglichen Zuständen 2^N und $\log_2 2^N = N$ beträgt. Wenn die Basis 10 benutzt würde, könnten die Einheiten Dezimaldigits genannt werden. Da nun

$$\log_2 M = \log_{10} M / \log_{10} 2 = 3{,}32 \log_{10} M,$$

ist ein Dezimaldigit etwa $3\frac{1}{3}$ bit. Eine digitale Wählscheibe an einem Tischcomputer hat 10 feste Positionen und daher eine Speicherkapazität von einem Dezimaldigit. Bei analytischer Arbeit, zu der Integration und Differentiation gehören, ist die Basis e manchmal nützlich. Die so entstehenden Einheiten der Information werden dann natürliche Einheiten genannt. Ein Wechsel von Basis a zur Basis b erfordert nur eine Multiplikation mit dem $\log_b a$.

Mit einem Kommunikationssystem bezeichnen wir ein System der Art, wie es schematisch in Abb. 1 aufgezeichnet ist. Es besteht aus fünf notwendigen Teilen:

1. Einer *Nachrichtenquelle*, die eine Nachricht oder eine Nachrichtenfolge produziert, die dem Empfänger mitgeteilt wird. Die Nachricht kann verschiedener Art sein: (a) Eine Folge von Buchstaben wie in einem Telegrafie- oder Fernschreibsystem; (b) eine reine Funktion der Zeit $f(t)$ wie bei Radio oder Telefon; (c) eine Funktion der Zeit und anderer Variabler wie im Schwarz-Weiß-Fernsehen - hier kann man von der Nachricht als einer Funktion $f(x, y, t)$ reden, von zwei Raumkoordinaten und der Zeit, von der Lichtintensität an einem Bildpunkt (x, y) auf einem Oszillografenschirm zur Zeit t; (d) zwei oder mehr Funktionen der Zeit, sagen wir $f(t)$, $g(t)$, $h(t)$ - dies ist der Fall in der "dreidimensionalen" Klangübertragung oder wenn das System erweitert wird, um mehrere individuelle Kanäle im Multiplexverfahren zu betreiben; (e) verschiedene Funktionen verschiedener Variabler - im Farbfernsehen besteht die Nachricht aus drei Funktionen $f(x, y, t)$, $g(x, y, t)$, $h(x, y, t)$ in einem dreidimensionalen Kontinuum - wir können uns diese drei Funktionen auch als Komponenten eines Vektorfeldes, das im entsprechenden Bereich definiert ist, vorstellen - ähnlich würden verschiedene Schwarz-Weiß-Fernsehquellen "Nachrichten" herstellen, die aus einer Anzahl von Funktionen von drei Variablen bestehen; (f) auch verschiedene Kombinationen kommen vor, z. B. ist das Fernsehen mit einem Hörfunkkanal gekoppelt.

2. Einem *Sender*, der eine Nachricht auf irgendeine Weise umformt, um ein Signal zu erzeugen, das für die Übertragung über

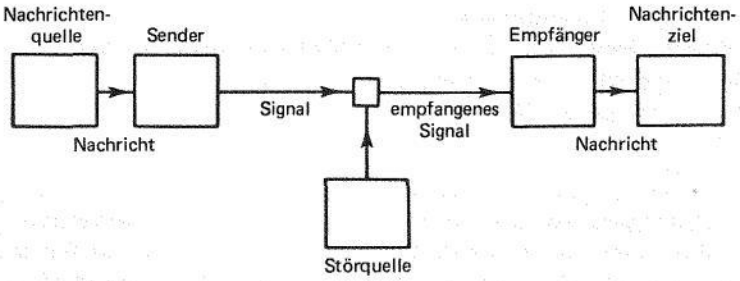


Abb. 1: Schema eines allgemeinen Kommunikationssystems

den Kanal geeignet ist. In der Telefonie besteht diese Operation einfach darin, den Schalldruck in einen proportionalen elektrischen Strom umzuwandeln. In der Telegrafie haben wir eine Codieroperation, die eine Folge von Punkten, Strichen und Leerstellen auf dem Kanal erzeugt, die der Nachricht entsprechen. In einem Multiplex PCM-System müssen die verschiedenen Sprachfunktionen vor der eigentlichen Übertragung registriert, komprimiert, quantisiert, codiert und schließlich in geeignete Abschnitte unterteilt werden, um das Signal herzustellen. Vocoder-Systeme, Fernsehen und Frequenzmodulation sind andere Beispiele von komplexen Arbeitsgängen, die auf die Nachricht angewandt werden, um ein Signal zu erhalten.

3. Der *Kanal* ist nur das Mittel, das man benützt, um das Signal vom Sender zum Empfänger zu übertragen. Es können ein paar Drähte sein, ein Koaxialkabel, ein Frequenzband, ein Lichtstrahl usw. Während der Übertragung oder an einem der "Terminals" kann das Signal gestört werden. Dies wird in Abb. 1 schematisch dargestellt durch die Störquelle, die auf das übertragene Signal einwirkt, so daß das empfangene Signal entsteht.

4. Der *Empfänger* führt normalerweise den entgegengesetzten Arbeitsgang des Senders durch, indem er die Nachricht wieder aus dem Signal rekonstruiert.

5. Das *Nachrichtenziel* ist die Person (oder Sache), für die die Nachricht bestimmt ist.

Wir möchten in diesem Zusammenhang gewisse allgemeine Probleme erwähnen, die mit Kommunikationssystemen verbunden sind.

Dafür ist es zuerst notwendig, die verschiedenen Bestandteile des Kommunikationssystems mathematisch auszudrücken, passend idealisiert gegenüber ihren physikalischen Ebenbildern. Wir können Kommunikationssysteme grob in drei Hauptkategorien einteilen: diskrete, kontinuierliche und gemischte. In einem diskreten System ist die Nachricht sowie das Signal eine Folge von einzelnen Zeichen. Ein typischer Fall ist wieder die Telegrafie, in der die Nachricht eine Folge von Buchstaben und das Signal eine Folge von Punkten, Strichen und Zwischenräumen bilden. In einem kontinuierlichen System werden Nachricht und Signal als kontinuierliche Funktionen behandelt, wie bei Radio oder Fernsehen. In einem gemischten System erscheinen sowohl diskrete als auch kontinuierliche Variable, etwa bei der PCM-Übertragung der Sprache.

Zunächst betrachten wir den diskreten Fall. Dieser Fall kommt nicht nur in der Kommunikationstheorie vor, er wird auch in der Theorie der Computer, der Darstellung von Telefonvermittlungen und auf anderen Gebieten angewandt. Zusätzlich bildet der diskrete Fall eine Grundlage für die kontinuierlichen und gemischten Methoden, die in der zweiten Hälfte dieser Darstellung behandelt werden.

I. Diskrete ungestörte Systeme

1. Der diskrete ungestörte Kanal

Fernschreiber und Telegraf sind zwei einfache Beispiele für einen diskreten Kanal zur Nachrichtenübertragung. Allgemein versteht man unter einem diskreten Kanal ein System, mit dem Folgen von Zeichen, die aus einer endlichen Menge elementarer Zeichen $S_1 \cdots, S_n$ ausgewählt sind, von einer Stelle zu einer anderen übertragen werden können. Für jedes der Zeichen S_i nimmt man eine bestimmte Zeitdauer von t_i Sekunden an (nicht notwendigerweise die gleiche Dauer für verschiedene S_i, z.B. die Punkte und Striche in der Telegrafie). Es wird nicht verlangt, daß alle möglichen Folgen von S_i mit dem System übertragen werden können; nur bestimmte Folgen mögen erlaubt sein. Diese werden dann die für den Kanal möglichen Signale sein. So nehmen wir in der Telegrafie folgende Zeichen an: (1) Einen Punkt, der für eine Zeiteinheit aus einer geschlossenen, für die nachfolgende Zeiteinheit aus einer geöffneten Verbindung besteht; (2) einen Strich, der aus drei Zeiteinheiten geschlossener und einer Zeiteinheit geöffneter Verbindung besteht; (3) einen Buchstabenzwischenraum, bestehend aus - sagen wir - drei Zeiteinheiten geöffneter Verbindung; (4) einen Wortzwischenraum aus sechs Zeiteinheiten geöffneter Verbindung. Wir können eine Einschränkung für erlaubte Folgen festlegen, wonach zwei Zwischenräume nicht aufeinanderfolgen dürfen (denn wenn zwei Buchstabenzwischenräume aufeinanderfolgen, sind sie identisch mit einem Wortzwischenraum). Im folgenden werden wir die Frage behandeln: Wie kann man die Kapazität eines solchen Kanals zur Nachrichtenübertragung messen?

Im Falle des Fernschreibers, bei dem alle Zeichen die gleiche Zeitdauer haben und jede Folge der 32 Zeichen erlaubt ist, ist die Antwort leicht. Jedes Zeichen stellt fünf bit Information dar. Wenn das System n Zeichen pro Sekunde überträgt, ist es naheliegend zu sagen, daß der Kanal eine Kapazität von $5n$ bit pro Sekunde besitzt. Das bedeutet nicht, daß der Fernschreibkanal in jeder Sekunde eine so große Informationsmenge überträgt - dies ist vielmehr die größtmögliche Rate. Wie sich später zeigen wird, hängt es von der den Kanal speisenden Nachrichtenquelle ab, ob die tatsächliche Übertragungsrate dieses Maximum erreicht.

Für den allgemeinen Fall mit Zeichen verschiedener Dauer und mit Beschränkungen für erlaubte Folgen führen wir folgende Definition ein: Die Kapazität C eines diskreten Kanals ist gegeben durch

$$C = \lim_{T\to\infty} \frac{\log N(T)}{T},$$

wobei $N(T)$ die Anzahl der erlaubten Signale von der Dauer T ist. Man kann leicht nachprüfen, daß dies im Fall des Fernschreibers auf das oben genannte Ergebnis führt. Außerdem läßt sich zeigen, daß der fragliche Grenzwert in den meisten interessierenden Fällen als endliche Zahl existiert. Angenommen, alle Folgen der Zeichen $S_1, \cdots, S_n$ seien zulässig und diese Zeichen seien von der Dauer $t_1, \cdots, t_n$. Welches ist dann die Kanalkapazität? Wenn $N(t)$ die Anzahl der Folgen der Dauer t darstellt, so gilt

$$N(t) = N(t-t_1) + N(t-t_2) + \cdots + N(t-t_n).$$

Die Anzahl der Folgen der Dauer t ist gleich der Summe der Anzahlen der Folgen, die mit $S_1, \cdots, S_n$ enden, diese aber sind gerade $N(t-t_1), \cdots, N(t-t_n)$. Nach einem wohlbekannten Ergebnis für endliche Differenzen nähert sich $N(t)$ für große t asymptotisch dem Wert AX_0^t, wobei A eine Konstante ist und X_0 die größte reelle Lösung der charakteristischen Gleichung

$$X^{-t_1} + X^{-t_2} + \cdots + X^{-t_n} = 1.$$

Deshalb gilt

$$C = \lim_{T\to\infty} \frac{\log AX_0^T}{T} = \log X_0.$$

Auch mit Beschränkungen für die erlaubten Folgen können wir oft eine Differenzengleichung obigen Typs erhalten und C aus der charakteristischen Gleichung finden. Im obenerwähnten Fall der Telegrafie gilt

$$N(t) = N(t-2) + N(t-4) + N(t-5) + N(t-7) + N(t-8) + N(t-10),$$

wie wir ersehen können, wenn wir die Folgen der Zeichen zählen, zusammengefaßt nach dem letzten bzw. vorletzten vorkommenden Zeichen. Also gilt $C = -\log \mu_0$, wobei μ_0 die positive Wurzel der Gleichung $1 = \mu^2 + \mu^4 + \mu^5 + \mu^7 + \mu^8 + \mu^{10}$ ist. Die Lösung ergibt $C = 0,539$.

Eine sehr allgemeine Art der Einschränkung, die erlaubten Folgen auferlegt werden kann, ist folgende: Stellen wir uns eine Anzahl möglicher Zustände $a_1, a_2, \cdots, a_m$ vor. In jedem Zustand können nur bestimmte Zeichen aus der Menge $S_1, \cdots, S_n$ übertragen werden (verschiedene Untermengen für die verschiedenen Zustände). Wenn eines der Zeichen übertragen worden ist, wechselt jedesmal der Zustand in einen neuen, der sowohl vom alten Zustand als auch vom speziellen übertragenen Zeichen abhängig ist. Die Telegrafie ist dafür ein einfaches Beispiel. Es gibt zwei Zustände, je nachdem, ob als letztes Zeichen ein Zwischenraum oder kein Zwischenraum übertragen worden ist. Im ersten Fall kann als nächstes Zeichen nur ein Punkt oder ein Strich gesendet werden, und beidemal ändert sich der Zustand. Im zweiten Fall kann jedes Zeichen folgen, und der Zustand ändert sich nur, wenn ein Zwischenraum gesendet wird, andernfalls bleibt er erhalten. Die Bedingungen können in einem linearen Graphen dargestellt werden, wie in Abb. 2 gezeigt.

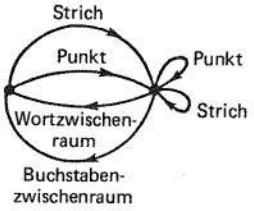


Abb. 2:
Graphische Darstellung der Bedingungen für telegrafische Zeichenfolgen

Die Knotenpunkte entsprechen den Zuständen, und die Pfeile zeigen die für einen Zustand möglichen Zeichen an und führen zu dem Folgezustand. Im Anhang 1 wird gezeigt, daß, wenn die Bedingungen für erlaubte Folgen auf diese Weise dargestellt werden können, C existiert und berechnet werden kann unter Verwendung des folgenden Lehrsatzes:

Lehrsatz 1: *Wenn $b_{ij}^{(s)}$ die Zeitdauer des Zeichen S_s ist, das im Zustand i erlaubt ist und in den Zustand j führt, dann ist die Kanalkapazität gleich* log W, *wobei W die größte reelle Wurzel folgender Determinantengleichungen ist:*

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0$$

$\delta_{ij} = 1$ *wenn* $i = j$, *andernfalls* $\delta_{ij} = 0$.

Zum Beispiel ist im Falle der Telegrafie (Abb. 2) die Determinante

$$\begin{vmatrix} -1 & (W^{-2} + W^{-4}) \\ (W^{-3} + W^{-6}) & (W^{-2} + W^{-4} - 1) \end{vmatrix} = 0.$$

In einer Reihe ausgeschrieben, führt dies zu der oben angegebenen Gleichung für diesen Satz von Bedingungen.

2. Die diskrete Nachrichtenquelle

Wir haben gesehen, daß unter sehr allgemeinen Bedingungen der Logarithmus der Anzahl von möglichen Signalen in einem diskreten Kanal linear mit der Zeit wächst. Die Kapazität, Information zu übertragen, kann durch die Wachstumsrate dargestellt werden, d.h. durch die Anzahl der bit pro Sekunde, die benötigt wird um das bestimmte verwendete Signal zu erzeugen.

Wir betrachten nun die Nachrichtenquelle. Wie soll eine Nachrichtenquelle mathematisch beschrieben werden und wieviel Information, gemessen in bit pro Sekunde, wird in einer gegebenen Quelle erzeugt? Ein wesentlicher Ausgangspunkt ist der Effekt, daß durch die Kenntnis der statistischen Eigenschaften der Quelle die benötigte Kanalkapazität reduziert werden kann, indem man die Nachricht auf die günstigste Weise codiert. In der Telegrafie zum Beispiel bestehen die zu übertragenden Nachrichten aus Buchstabenfolgen. Die Folgen sind jedoch nicht vollständig zufällig. Im allgemeinen bilden sie Sätze und haben die statistische Struktur etwa der englischen Sprache. Der Buchstabe E kommt häufiger vor als Q, die Folge TH häufiger als XP usw. Die Existenz dieser Struktur erlaubt eine Zeiteinsparung (oder die Reduzierung der Kanalkapzität) dadurch, daß die Nachrichtenfolgen mit der richtigen Codierung in Signalfolgen überführt werden. Dies ist bis zu einem begrenzten Grad in der Telegrafie bereits dadurch geschehen, daß das kürzeste Kanalzeichen, ein Punkt, für den am meisten gebrauchten englischen Buchstaben E benutzt wird, während die seltenen Buchstaben Q, X, Z durch längere Folgen von Punkten und Strichen dargestellt werden. Diese Idee ist noch weiter verwendet worden in gewissen kommerziellen Codes, in denen gebräuchliche Worte und Sätze durch Codegruppen von vier oder fünf Buchstaben dargestellt werden; dabei entsteht ein beträchtlicher Zeitgewinn. Die standardisierten Gruß- und Glückwunschtelegramme, die man jetzt benutzt, erweitern dies bis zu dem Punkt, daß ein oder zwei Sätze durch eine verhältnismäßig kleine Folge von Zahlen codiert werden*).

*) gilt nicht in Deutschland (Anm. d. Übers.)

Wir können dann von einer diskreten Quelle sprechen, wenn die Nachricht durch einzelne aufeinanderfolgende Zeichen erzeugt wird. Sie wird die aufeinanderfolgenden Zeichen nach gewissen Wahrscheinlichkeiten auswählen, die im allgemeinen ebenso von vorhergegangenen Auswahlergebnissen wie von den jeweiligen Zeichen selbst abhängen. Wenn ein physikalisches System oder ein mathematisches Modell eines Systems eine solche Folge von Zeichen, die von einem Satz von Wahrscheinlichkeiten bestimmt wird, erzeugt, dann wird dieser Vorgang als stochastischer Prozeß bezeichnet [3]). Wir können also eine diskrete Quelle betrachten, die durch einen stochastischen Prozeß beschrieben wird. Umgekehrt kann man jeden stochastischen Prozeß, der eine diskrete Zeichenfolge - aus einem bestimmten Vorrat ausgewählt - erzeugt, als eine diskrete Quelle betrachten.

Das wird auch solche Fälle einschließen wie:

1. Natürliche geschriebene Sprachen wie Englisch, Deutsch oder Chinesisch.
2. Kontinuierliche Nachrichtenquellen, die nach irgendeinem Quantisierungsprozeß auf diskrete zurückgeführt worden sind. Zum Beispiel die quantisierte Sprache eines PCM-Senders oder ein quantisiertes Fernsehsignal.
3. Mathematische Fälle, in denen wir einfach einen stochastischen Prozeß abstrakt definieren, der eine Zeichenfolge erzeugt. Anschließend finden Sie Beispiele für diese dritte Art von Quellen.

(A) Angenommen, wir haben fünf Buchstaben A, B, C, D, E, die jeweils mit der Wahrscheinlichkeit 0,2 gewählt worden sind, wobei aufeinanderfolgende Auswahlen unabhängig voneinander sind. Dies würde zu einer Folge führen, für welche diese ein typisches Beispiel ist:

B D C B C E C C C A D C B D D A A E C E E A A B B D A E E C
A C E E B A E E C B C E A D.

Sie wurde nach einer Tabelle von Zufallszahlen [4]) konstruiert.

(B) Nun werden dieselben fünf Buchstaben benutzt mit den Wahrscheinlichkeiten 0,4 , 0,1 , 0,2 , 0,1 bei voneinander unabhängigen, aufeinanderfolgenden Auswahlvorgängen. Eine typische Nachricht von dieser Quelle ist dann:

[3]) Siehe z. B. von S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy", erschienen in *Reviews of Modern Physics*, Bd. 15, Nr. 1, Januar 1943, S. 1.

[4]) Kendall und Smith, *Tables of Random Sampling Numbers*, Cambridge, 1939.

A A A C D C B D C E A A D A D A C E D A E A D C A B E D A D
D C E C A A A A A D.

(C) Eine komplizierte Struktur erhält man, wenn aufeinanderfolgende Zeichen nicht unabhängig ausgewählt werden, ihre Wahrscheinlichkeiten vielmehr von vorhergehenden Zeichen abhängig sind. Im einfachsten Fall dieser Art hängt eine Auswahl nur von dem gerade vorhergehenden Zeichen und nicht von früheren ab. Der statistische Aufbau kann dann durch einen Satz von Übergangswahrscheinlichkeiten $p_i(j)$ beschrieben werden, die Wahrscheinlichkeit, daß dem Buchstaben i der Buchstabe j folgt. Die Indizes i und j gelten für alle hier möglichen Zeichen. Ein zweiter äquivalenter Weg, den Aufbau genau darzustellen, ist, die "Digramm"-Wahrscheinlichkeiten (zweidimensionale Verbundwahrscheinlichkeit) $p(i,j)$ anzugeben, d. h. die relative Häufigkeit des Digramms $i\,j$. Die Buchstabenhäufigkeiten $p(i)$ (die Wahrscheinlichkeit des Buchstabens i), die Übergangswahrscheinlichkeiten $p_i(j)$, und die Verbundwahrscheinlichkeiten $p(i,j)$ hängen durch die folgenden Formeln zusammen:

$$p(i) = \sum_j p(i,j) = \sum_j p(j,i) = \sum_j p(j)p_j(i)$$
$$p(i,j) = p(i)p_i(j)$$
$$\sum_j p_i(j) = \sum_i p(i) = \sum_{i,j} p(i,j) = 1.$$

Als ein spezifisches Beispiel nehmen wir an, es gibt drei Buchstaben A, B und C mit den Wahrscheinlichkeitstabellen:

$p_i(j)$		j A	B	C
i	A	0	$\frac{4}{5}$	$\frac{1}{5}$
	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

i	$p(i)$
A	$\frac{9}{27}$
B	$\frac{16}{27}$
C	$\frac{2}{27}$

$p(i,j)$		j A	B	C
i	A	0	$\frac{4}{15}$	$\frac{1}{15}$
	B	$\frac{8}{27}$	$\frac{8}{27}$	0
	C	$\frac{1}{27}$	$\frac{4}{135}$	$\frac{1}{135}$

Eine typische Nachricht dieser Quelle wäre die folgende:

A B B A B A B A B A B A B A B B B A B B B B B A B A B A B A
B A B B B A C A C A B B A B B B B A B B A B A C B B B A B A.

Die nächste Erhöhung des Schwierigkeitsgrades würde Trigramm-Häufigkeiten mit sich bringen, jedoch nicht mehr. Die Auswahl eines Buchstabens würde von den zwei vorhergehen-

den Buchstaben abhängen, aber nicht von der Nachricht vor dieser Stelle. Ein Satz von Trigramm-Häufigkeiten $p(i, j, k)$ oder, gleichbedeutend, ein Satz von Übergangswahrscheinlichkeiten $p_{ij}(k)$ wäre notwendig. Wenn man auf diese Art und Weise fortfährt, erhält man aufeinanderfolgend kompliziertere stochastische Prozesse. In dem allgemeinen n-gramm-Fall wird ein Satz von n-gramm-Wahrscheinlichkeiten $p(i_1, i_2, \cdots, i_n)$ oder von Übergangswahrscheinlichkeiten $p_{i_1, i_2, \cdots, i_{n-1}}(i_n)$ verlangt, um die statistische Struktur genau darzulegen.

(D) Es können auch solche stochastischen Prozesse definiert werden, die als Text eine Folge von "Wörtern" hervorbringen. Angenommen, es gibt fünf Buchstaben A, B, C, D, E und 16 "Wörter" in der Sprache, mit den zugehörigen Wahrscheinlichkeiten:

0,10 A	0,16 BEBE	0,11 CABED	0,04 DEB
0,04 ADEB	0,04 BED	0,05 CEED	0,15 DEED
0,05 ADEE	0,02 BEED	0,08 DAB	0,01 EAB
0,01 BADD	0,05 CA	0,04 DAD	0,05 EE

Angenommen, aufeinanderfolgende "Wörter" werden unabhängig voneinander ausgewählt und sind durch einen Zwischenraum getrennt. Eine typische Nachricht könnte sein:

DAB EE A BEBE DEED DEB ADEE ADEE EE DEB BEBE BEBE ADEE BED DEED DEED CEED ADEE A DEED DEED BEBE CABED BEBE BED DAB DEED ADEB.

Wenn alle Wörter eine endliche Länge haben, ist dieser Prozeß äquivalent zu einem der vorangegangenen Art, aber die Beschreibung kann mit Hilfe der Wortstruktur und den Wortwahrscheinlichkeiten einfacher sein. Wir können auch hier verallgemeinern und die Übergangswahrscheinlichkeiten zwischen Wörtern einführen usw.

Diese künstlichen Sprachen sind nützlich, um einfache Probleme und Beispiele zu konstruieren und um verschiedene Möglichkeiten zu illustrieren. Wir können auch eine natürliche Sprache approximieren durch Anwendung einer Serie von einfachen künstlichen Sprachen. Die Näherung nullter Ordnung wird erreicht, indem man alle Buchstaben mit denselben Wahrscheinlichkeiten und unabhängig voneinander auswählt. Bei der Näherung erster Ordnung werden aufeinanderfolgende Buchstaben unabhängig voneinander gewählt, wobei jedoch jeder Buchstabe die Wahrscheinlichkeit besitzt, die er in der

natürlichen Sprache hat [5]). Auf diese Art wird in der Näherung erster Ordnung zur englischen Sprache der Buchstabe E mit der Wahrscheinlichkeit 0,12 ausgewählt (entsprechend seiner Häufigkeit im normalen Englisch) und W mit der Wahrscheinlichkeit 0,02; es besteht bei dieser Näherung jedoch kein Einfluß zwischen aufeinanderfolgenden Buchstaben und keine Tendenz, die bevorzugten Digramme wie z.B. TH, ED usw. zu erzeugen. In der Näherung zweiter Ordnung wird der Digramm-Aufbau eingeführt. Nachdem ein Buchstabe gewählt wurde, wird der nächste in Übereinstimmung mit den Häufigkeiten ausgewählt, mit denen die verschiedenen Buchstaben dem ersten folgen. Dies erfordert eine Tabelle von Digramm-Häufigkeiten $p(i,j)$. In der Näherung dritter Ordnung wird die Trigramm-Struktur verwendet. Jeder Buchstabe wird mit Wahrscheinlichkeiten, die von den beiden vorangehenden Buchstaben abhängen, ausgewählt.

3. Die Serien der Näherungen zur englischen Sprache

Um eine anschauliche Vorstellung davon zu vermitteln, wie diese Serie von Prozessen eine Sprache annähert, wurden typische Folgen in der Näherung zum Englischen konstruiert, wie sie unten angegeben werden. In allen Fällen haben wir ein Alphabet von 27 Zeichen vorausgesetzt, die 26 Buchstaben und einen Zwischenraum.

1. Näherung nullter Ordnung (Zeichen voneinander unabhängig und gleich wahrscheinlich)

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYV-KCQSGHYD QPAAMKBZAACIBZLHJQD.

2. Näherung erster Ordnung (Zeichen voneinander unabhängig, jedoch mit Häufigkeiten eines englischen Textes)

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Näherung zweiter Ordnung (Digramm-Struktur wie im Englischen)

[5]) Buchstaben-, Digramm- und Trigramm-Häufigkeiten werden in *Secret and Urgent* von Fletcher Pratt, Blue Ribbon Books, 1939, angegeben. Worthäufigkeiten sind tabelliert in *Relative Frequency of English Speech Sounds* von G. Dewey, Harvard University Press, 1923.

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Näherung dritter Ordnung (Trigramm-Struktur wie im Englischen)

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. Näherung erster Ordnung, auf Wörter bezogen.
Es ist einfacher und besser, an diesem Punkt zu Worteinheiten überzugehen, als mit Tetragramm ... bis n-gramm-Strukturen weiterzuarbeiten. Hier werden die Wörter unabhängig voneinander ausgewählt, jedoch mit den ihnen eigenen Häufigkeiten.

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Wort-Näherung zweiter Ordnung. Die Wort-Übergangswahrscheinlichkeiten sind korrekt, jedoch ist keine weitere Strukturierung inbegriffen.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Die Ähnlichkeit gegenüber einem gewöhnlichen englischen Text vergrößert sich erkennbar bei jedem der vorangegangenen Schritte. Beachtenswert ist, daß diese Beispiele eine ziemlich gute Struktur aufweisen, etwa bis zum Doppelten der Buchstaben- bzw. Wortanzahl, deren Übergangswahrscheinlichkeit bei der Konstruktion berücksichtigt wird. So garantiert in (3) der statistische Prozeß eine mögliche Folge von jeweils zwei Buchstaben; vier-Buchstaben-Folgen aus dem Beispiel können jedoch normalerweise in gute Sätze eingefügt werden. In (6) können Folgen von vier oder mehr Wörtern leicht in Sätzen ohne ungewöhnliche oder überspannte Konstruktion einen Platz finden. Die besondere Folge von zehn Wörtern "attack on an English writer that the character of this" (deutsch etwa: Angriff auf einen englischen Schriftsteller, daß der Charakter dieses

.....) ist überhaupt nicht unnatürlich. Es stellt sich heraus, daß ein genügend komplexer stochastischer Prozeß eine befriedigende Vorstellung einer diskreten Quelle gibt.

Die ersten beiden Beispiele wurden mit Hilfe eines Buches von zufälligen Zahlen in Verbindung (für Beispiel 2) mit einer Tabelle von Buchstabenhäufigkeiten erzeugt. Diese Methode hätte für (3), (4) und (5) weitergeführt werden können, da Digramm-, Trigramm- und Worthäufigkeitstabellen zur Verfügung stehen, es wurde jedoch eine einfachere äquivalente Methode verwendet. Um etwa die Folge (3) zu konstruieren, schlägt man ein Buch an irgendeiner Stelle auf und tippt auf dieser Seite blind auf irgendeinen Buchstaben. Dieser Buchstabe wird niedergeschrieben. Dann wird das Buch auf einer anderen Seite geöffnet und man liest so lange, bis man denselben Buchstaben wieder trifft. Nun wird der unmittelbar nachfolgende Buchstabe aufgezeichnet. Man schlägt eine weitere Seite auf, sucht nun nach diesem Buchstaben und schreibt den jetzt folgenden Buchstaben auf usw. Ein ähnlicher Vorgang wurde für (4) angewendet sowie für (5) und (6). Es wäre interessant, ob weitere Näherungen konstruiert werden könnten, jedoch wird der Aufwand dafür bei der nächsten Stufe außerordentlich groß.

4. Graphische Darstellung eines Markoff-Prozesses

Stochastische Prozesse der beschriebenen Art sind in der Mathematik bekannt als diskrete Markoff-Prozesse und sind in der Literatur [6]) ausführlich behandelt worden. Der allgemeine Fall kann wie folgt beschrieben werden: Es existiert eine bestimmte Anzahl von möglichen Zuständen eines Systems; $S_1, S_2, \cdots, S_n$. Zusätzlich gibt es einen Satz von Übergangswahrscheinlichkeiten, $p_i(j)$, die Wahrscheinlichkeit, mit der das System vom Zustand S_i in den Zustand S_j übergeht. Um aus diesem Markoff-Prozeß eine Informationsquelle herzustellen, müssen wir annehmen, daß bei jedem Übergang von einem Zustand in einen anderen ein Zeichen erzeugt wird. Die Zustände werden dem "noch verbleibenden Einfluß" von vorhergehenden Buchstaben entsprechen.

Die Situation kann grafisch dargestellt werden, wie es in den Abb. 3, 4 und 5 gezeigt ist. Die Zustände sind die Knotenpunkte in dem

[6]) Für eine detaillierte Behandlung siehe M. Frechet, *Methods des fonctions arbitraires. Theorie des énénements en chaine dans le cas d'un nombre fini d'états possibles*, Paris, Gauthier Villars, 1938.

Diagramm, und die Wahrscheinlichkeiten und Buchstaben, die bei einem Übergang erzeugt werden, sind neben der entsprechenden Linie angegeben. Abb. 3 entspricht dem Beispiel B in Kapitel 2, während Abb. 4 das Beispiel C abbildet. In Abb. 3 gibt es nur einen Zustand, da die aufeinanderfolgenden Buchstaben unabhängig voneinander sind. In Abb. 4 gibt es soviele Zustände wie Buchstaben.

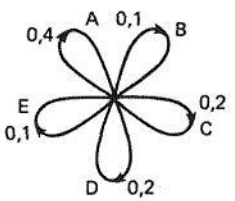


Abb. 3:
Ein Diagramm, das der Quelle in Beispiel B entspricht

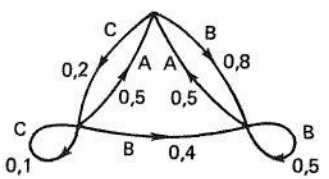


Abb. 4:
Ein Diagramm, das der Quelle in Beispiel C entspricht

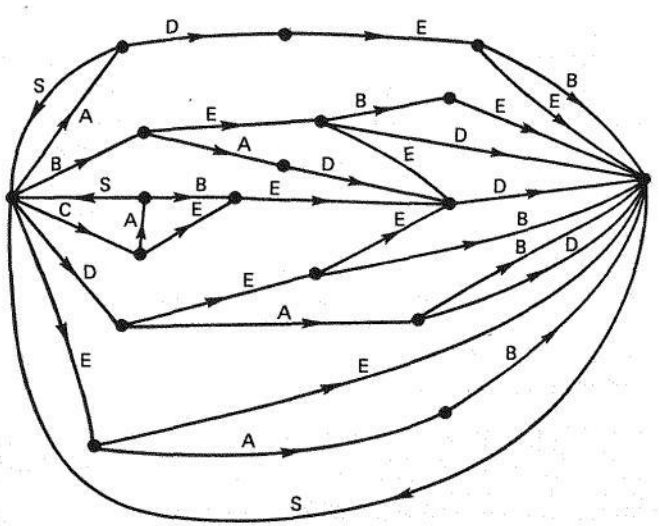


Abb. 5: Ein Diagramm, das der Quelle in Beispiel D entspricht

Wenn ein Trigramm-Beispiel konstruiert würde, gäbe es höchstens n^2 Zustände, entsprechend den möglichen Buchstabenpaaren, die einem einzelnen gewählten Buchstaben vorangehen. Abb. 5 ist ein Diagramm für den Fall der Wortstruktur in Beispiel D. Hier entspricht "s" dem Zeichen "Zwischenraum".

5. Ergodische und gemischte Quellen

Wie wir bereits beschrieben haben, kann eine diskrete Quelle für unsere Zwecke so betrachtet werden, als werde sie von einem Markoff-Prozeß dargestellt. Unter den möglichen diskreten Markoff-Prozessen gibt es eine Gruppe mit besonderen, bedeutungsvollen Merkmalen für die Kommunikationstheorie. Diese besondere Klasse sind die ergodischen Prozesse, und wir werden die entsprechenden Quellen ergodische Quellen nennen. Obwohl eine strenge Definition eines ergodischen Prozesses etwas verwickelt ist, ist der allgemeine Gedanke einfach. Alle in einem ergodischen Prozeß erzeugten Folgen haben dieselben statistischen Eigenschaften. Daher werden sich die von bestimmten Folgen erhaltenen Buchstabenhäufigkeiten, Digrammhäufigkeiten usw., wenn sich die Längen der Folgen vergrößern, festgelegten Grenzen nähern, unabhängig von den gewählten Folgen. Tatsächlich ist dies nicht bei jeder Folge der Fall, aber die Menge, für die das nicht der Fall ist, hat die Wahrscheinlichkeit null. Vereinfacht bedeutet die ergodische Eigenschaft statistische Homogenität.

All die oben erwähnten Beispiele von künstlichen Sprachen sind ergodischer Natur. Diese Besonderheit hängt mit der Struktur des zugehörigen Diagramms zusammen. Falls das Diagramm die beiden folgenden Eigenschaften[7]) hat, werden die entsprechenden Prozesse ergodisch sein:

1. Das Diagramm besteht nicht aus zwei getrennten Teilen A und B, so daß es unmöglich wäre, von Knotenpunkten im Teil A über Linien des Diagramms in Pfeilrichtung zu Knotenpunkten im Teil B zu gelangen, und ebenso unmöglich, von Knotenpunkten im Teil B Knoten im Teil A zu erreichen.

2. Eine geschlossene Folge von Linien im Diagramm mit Pfeilen, die in die gleiche Umlaufrichtung zeigen, wird ein "Umlauf" ge-

[7]) Dies sind Behauptungen im Sinne der Aufstellung von Bedingungen, die in FRECHET angegeben werden.

nannt. Die "Länge" eines Umlaufs ist die Anzahl der in ihm vorkommenden Linien. Daher ist die Folge BEBES in Abb. 5 ein Umlauf mit der Länge 5. Die zweite notwendige Eigenschaft ist, daß der größte gemeinsame Teiler der Längen aller Umläufe im Diagramm Eins ist.

Falls nur die erste Bedingung erfüllt ist, die zweite jedoch nicht, da der größte gemeinsame Teiler $d > 1$ ist, besitzen die Folgen eine gewisse periodische Struktur. Die diversen Folgen sind in d verschiedene Klassen einzuordnen, welche statistisch dieselben sind, abgesehen von einer Veränderung des Ursprungspunktes (z. B. welcher Buchstabe in der Folge wird Buchstabe 1 genannt). Durch eine Veränderung von 0 bis zu $d-1$ kann jede Folge statistisch äquivalent zu jeder anderen gemacht werden. Ein einfaches Beispiel mit $d = 2$ ist das folgende: Es gibt drei mögliche Buchstaben: a, b und c. Dem Buchstaben a folgt entweder b oder c mit den Wahrscheinlichkeiten $\frac{1}{2}$ bzw. $\frac{1}{2}$. Den Zeichen b oder c folgt immer a. Eine typische Folge wäre:

$$a\,b\,a\,c\,a\,c\,a\,c\,a\,b\,a\,c\,a\,b\,a\,b\,a\,c\,a\,c.$$

Diese Situation hat für unseren Zweck keine große Bedeutung.

Wenn die erste Bedingung nicht erfüllt ist, kann das Diagramm in einen Satz Sub-Diagramme unterteilt werden, von welchen jedes die erste Bedingung erfüllt. Wir werden voraussetzen, daß die zweite Bedingung ebenfalls für jedes Sub-Diagramm erfüllt ist. In diesem Fall haben wir das, was man eine "gemischte" Quelle nennen kann, die aus einer Anzahl von reinen Komponenten gebildet worden ist. Die Komponenten entsprechen den verschiedenen Sub-Diagrammen. Wenn $L_1, L_2, L_3, \ldots$ die Komponentenquellen sind, können wir schreiben

$$L = p_1 L_1 + p_2 L_2 + p_3 L_3 + \cdots,$$

wobei p_i die Wahrscheinlichkeit der Komponentenquelle L_i ist. Physikalisch ist die dargestellte Situation so: Es gibt einige verschiedene Quellen $L_1, L_2, L_3, \ldots$, die alle von einer homogenen statistischen Struktur sind (d. h. sie sind ergodisch). Wir wissen *a priori* nicht, welche von ihnen Verwendung finden soll, aber wenn die Folge einmal mit einer reinen Komponente L_i beginnt, wird sie laufend fortgesetzt, gemäß der statistischen Struktur dieser Komponente.

Als ein Beispiel kann man zwei der schon erklärten Prozesse nehmen und festlegen, daß $p_1 = 0{,}2$ und $p_2 = 0{,}8$ seien. Eine Folge

von der gemischten Quelle

$$L = 0{,}2\,L_1 + 0{,}8\,L_2$$

würde man erhalten, indem man zuerst L_1 oder L_2 mit den Wahrscheinlichkeiten 0,2 und 0,8 wählte und nach dieser Wahl eine Folge entsprechend der gewählten Quelle erzeugen würde.

Außer wenn ausdrücklich das Gegenteil behauptet wird, nehmen wir an, daß eine Quelle ergodisch sei. Diese Annahme versetzt uns in die Lage, Mittelwerte über eine Folge mit Mittelwerten über die Klasse von möglichen Folgen gleichzusetzen (wobei die Wahrscheinlichkeit einer Abweichung gleich Null ist).

Zum Beispiel wird die relative Häufigkeit des Buchstabens A in einer einzelnen unendlichen Folge mit der Wahrscheinlichkeit "Eins" gleich seiner relativen Häufigkeit in der Gesamtheit von Folgen sein. Wenn P_i die Wahrscheinlichkeit des Zustandes i und $p_i(j)$ die Übergangswahrscheinlichkeit zum Zustand j ist, dann müssen in einem stationären Prozeß die verschiedenen P_i die Gleichgewichtsbedingungen erfüllen:

$$P_j = \sum_i P_i p_i(j).$$

Im ergodischen Fall kann gezeigt werden, daß bei beliebigen Anfangsbedingungen die Wahrscheinlichkeiten $P_j(N)$, nach N Zeichen im Zustand j zu sein, sich den Gleichgewichtswerten nähern für $N \to \infty$.

6. Auswahl, Unsicherheit und Entropie

Wir haben eine diskrete Nachrichtenquelle als einen Markoff-Prozeß vorgestellt. Können wir eine Größe definieren, die in einem gewissen Sinn mißt, wieviel Information durch einen solchen Vorgang "erzeugt" wird, oder besser, mit welcher Rate Information erzeugt wird?

Angenommen, wir haben einen Satz von möglichen Ereignissen, deren Wahrscheinlichkeiten, daß sie auftreten, $p_1, p_2, \cdots, p_n$ sind. Diese Wahrscheinlichkeiten sind bekannt, das ist jedoch alles, was wir darüber wissen, welches Ereignis auftreten wird. Können wir einen Maßstab dafür finden, wieviel Wahlfreiheit in die Auswahl des Ereignisses einbezogen ist oder wie ungewiß wir bezüglich des Ergebnisses sind?

Wenn ein solches Maß vorhanden ist, sagen wir $H(p_1, p_2, \cdots, p_n)$,

dann ist es angebracht, von ihm folgende Eigenschaften zu verlangen:

1. H sollte kontinuierlich in den p_i sein.

2. Wenn alle p_i gleich sind, $p_i = \frac{1}{n}$, dann sollte H eine monoton wachsende Funktion von n sein. Bei Ereignissen, welche dieselbe Wahrscheinlichkeit haben, besteht eine um so größere Wahlfreiheit oder Ungewißheit, je mehr mögliche Ereignisse es gibt.

3. Wenn eine Auswahl in zwei aufeinanderfolgende Wahlvorgänge aufgeteilt wird, sollte das ursprüngliche H die gewichtete Summe der individuellen H-Werte sein. Die Bedeutung davon ist in Abb. 6 aufgezeichnet. Links haben wir drei Möglichkeiten $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. Rechts wählen wir zuerst zwischen zwei Möglichkeiten, von denen jede die Wahrscheinlichkeit $\frac{1}{2}$ besitzt, und wenn die zweite auftritt, wählt man noch einmal nach den Wahrscheinlichkeiten $\frac{2}{3}$, $\frac{1}{3}$. Die Endergebnisse haben die gleichen Wahrscheinlichkeiten wie vorher. Wir fordern in diesem besonderen Fall, daß

$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2}\, H(\tfrac{2}{3}, \tfrac{1}{3}).$$

Der Koeffizient $\frac{1}{2}$ ist der wichtende Faktor, der eingeführt wird, weil diese zweite Wahl nur halb so oft vorkommt.

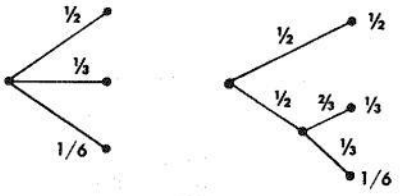


Abb. 6:
Zerlegung einer Wahl aus drei Möglichkeiten

Im Anhang 2 wird das folgende Ergebnis abgeleitet:

Lehrsatz 2: *Der einzige Wert H, der die drei obigen Voraussetzungen erfüllt, hat die Form:*

$$H = -K \sum_{i=1}^{n} p_i \log p_i ,$$

wobei K eine positive Konstante ist.

Dieser Lehrsatz und die für seinen Beweis verlangten Voraussetzungen sind überhaupt nicht notwendig für die behandelte Theorie.

Er wird in erster Linie deswegen angegeben, um eine gewisse Plausibilität für einige unserer späteren Definitionen zu erreichen. Die wirkliche Rechtfertigung für diese Definitionen wird allerdings erst durch ihre Auswirkungen zu erreichen sein.

Größen von der Form $H = -\Sigma\, p_i \log p_i$ (die Konstante K wird nur für die Wahl einer Maßeinheit gebraucht) spielen eine zentrale Rolle in der Informationstheorie als Maßeinheiten für Information, Wahlfreiheit und Ungewißheit. Die Form von H entspricht der der

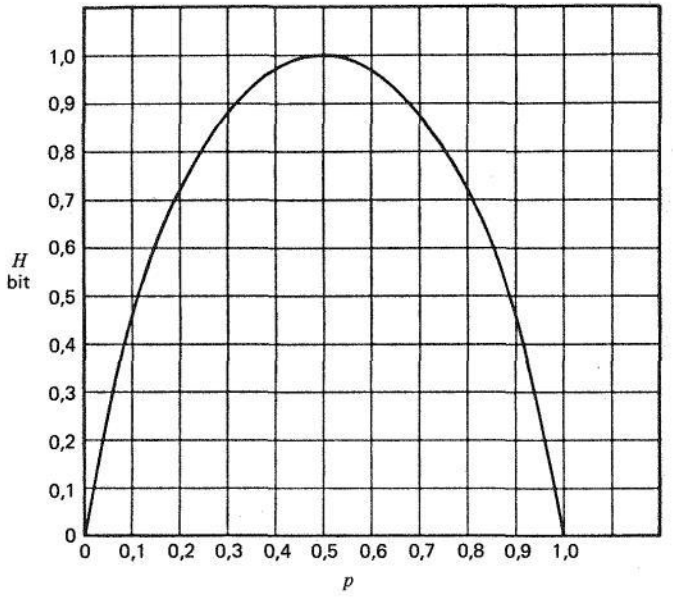


Abb. 7: Entropie bei zwei Ereignissen mit den Wahrscheinlichkeiten p und $(1-p)$

Entropie, wie sie in bestimmten Formulierungen der statistischen Mechanik [8]) erklärt ist, wobei p_i die Wahrscheinlichkeit ist, daß sich ein System im Volumenelement i seines Phasenraumes befindet. Damit ist H zum Beispiel der Wert H in Boltzmanns berühmtem Lehrsatz zur Entropie. Wir werden $H = -\Sigma\, p_i \log p_i$ die Entropie des Wahrscheinlichkeitssatzes $p_1, \cdots, p_n$ nennen. Wenn x eine Zufallsvariable ist, werden wir $H(x)$ für seine Entropie schrei-

[8]) Siehe z. B. R. LC. Tolman, *Prinziples of Statistical Mechanics*, Oxford, Clarendon, 1938.

ben; dadurch ist x nicht das Argument einer Funktion, sondern ein "Etikett" für eine Zahl, um sie von $H(y)$ zu unterscheiden, wie etwa von der Entropie der Zufallsvariablen y.

Die Entropie ist im Fall von zwei Möglichkeiten mit den Wahrscheinlichkeiten p und $q = 1 - p$, also

$$H = -(p \log p + q \log q)$$

in Abb. 7 als eine Funktion von p dargestellt.

Die Größe H hat eine Anzahl interessanter Eigenschaften, die sie weiterhin als ein vortreffliches Maß für Wahlfreiheit oder Information erscheinen lassen.

1. $H = 0$ dann und nur dann, wenn alle p_i außer einem einzigen Null sind, wobei dieses eine den Wert Eins hat. Danach verschwindet H nur dann, wenn wir über das Ergebnis sicher sind. Sonst ist H positiv.

2. Für einen gegebenen Wert n ist H ein Maximum und gleich dem $\log n$, wenn alle Werte p_i gleich sind, d.h. $\frac{1}{n}$. Dies erscheint uns auch intuitiv als die unsicherste Lage.

3. Angenommen, es gäbe zwei mögliche Ereignisse, x und y, mit m Möglichkeiten für das erste und n Möglichkeiten für das zweite. Es sei $p(i,j)$ die Wahrscheinlichkeit für das Verbundereignis, wobei i für das erste und j für das zweite Ereignis steht. Die Gesamtentropie ist dann

$$H(x,y) = -\sum_{i,j} p(i,j) \log p(i,j),$$

wobei

$$H(x) = -\sum_{i,j} p(i,j) \log \sum_j p(i,j)$$

$$H(y) = -\sum_{i,j} p(i,j) \log \sum_i p(i,j).$$

Es kann leicht gezeigt werden, daß

$$H(x,y) \leq H(x) + H(y),$$

mit Gleichheit nur in dem Falle, daß die Einzelereignisse unabhängig voneinander sind (d.h. $p(i,j) = p(i)\,p(j)$). Die Unsicherheit für ein Verbundereignis ist kleiner oder höchstens gleich der Summe der einzelnen Unsicherheiten.

4. Jeder Veränderung in Richtung auf Gleichheit der Wahrscheinlichkeiten $p_1, p_2, \cdots, p_n$ läßt H wachsen. Wenn also $p_1 < p_2$ ist und

wir p_1 vergrößern und p_2 um den gleichen Betrag verringern, damit p_1 und p_2 sich annähern, wächst die Entropie H. Allgemeiner gesagt, wenn wir irgendeine "mittelnde" Operation an p_i durchführen in der Form

$$p_i' = \sum_j a_{ij} p_j,$$

wobei $\sum_i a_{ij} = \sum_j a_{ij} = 1$ und alle $a_{ij} \geq 0$ sind, dann wächst H (außer in dem Spezialfall, in dem diese Umwandlung nur eine Permutation der p_i bedeutet, wobei H natürlich gleich bleibt).

5. Angenommen, es gibt zwei Ereignisse x und y wie in 3, die nicht unbedingt unabhängig voneinander sein müssen. Für jeden einzelnen Wert i, den x annehmen kann, gibt es eine bedingte Wahrscheinlichkeit $p_i(j)$, daß y den Wert j hat. Dies ist gegeben durch

$$p_i(j) = \frac{p(i,j)}{\sum_j p(i,j)}.$$

Wir definieren die *bedingte Entropie* von y, $H_x(y)$ als die mittlere Entropie von y für jeden Wert von x und gewichtet entsprechend der Wahrscheinlichkeit, gerade dieses spezielle x zu erhalten. Das bedeutet

$$H_x(y) = -\sum_{i,j} p(i,j) \log p_i(j).$$

Diese Größe ist ein Maß dafür, wie unsicher wir im Durchschnitt über y sind, wenn wir x kennen. Indem wir den Wert von $p_i(j)$ einsetzen, erhalten wir

$$H_x(y) = -\sum_{i,j} p(i,j) \log p(i,j) + \sum_{i,j} p(i,j) \log \sum_j p(i,j)$$
$$= H(x,y) - H(x)$$

oder

$$H(x,y) = H(x) + H_x(y).$$

Die Unsicherheit (oder Entropie) des Verbundereignisses x, y ist die Unsicherheit von x plus der Unsicherheit von y, wenn x bekannt ist.

6. Aus 3 und 5 erhalten wir

$$H(x) + H(y) \geq H(x,y) = H(x) + H_x(y),$$

wonach

$$H(y) \geq H_x(y) \qquad \text{folgt.}$$

Die Unsicherheit von y wird mit der Kenntnis von x niemals vergrößert. Sie wird stets verringert, es sei denn, x und y sind unabhängige Ereignisse; in diesem Fall ändert sie sich nicht.

7. Die Entropie einer Nachrichtenquelle

Betrachten wir eine diskrete Quelle mit einer endlichen Anzahl von Zuständen, wie sie weiter oben schon erwähnt wurde. Für jeden möglichen Zustand i wird es einen Satz von Wahrscheinlichkeiten $p_i(j)$ für die Erzeugung der Zeichen j geben. Dadurch gibt es für jeden Zustand eine Entropie H_i. Die Entropie der Quelle wird als der Mittelwert dieser Werte H_i definiert, gewichtet mit der Wahrscheinlichkeit für das Auftreten des fraglichen Zustands:

$$H = \sum_i P_i H_i$$
$$= -\sum_{i,j} P_i p_i(j) \log p_i(j).$$

Dies ist die Entropie der Quelle pro Zeichen des Textes. Wenn der Markoff-Prozeß mit einer bestimmten Geschwindigkeit abläuft, gibt es auch eine Entropie pro Sekunde.

$$H' = \sum_i f_i H_i ,$$

wobei f_i die mittlere Frequenz (Auftreten pro Sekunde) des Zustandes i darstellt. Damit wird

$$H' = m \cdot H ,$$

wobei m die durchschnittliche Anzahl von Zeichen ist, die pro Sekunde erzeugt werden, H oder H' mißt den Betrag der Information, die von der Quelle pro Zeichen oder pro Sekunde erzeugt wird. Wenn die Basis des Logarithmus 2 ist, stellen sie bit pro Zeichen oder bit pro Sekunde dar.

Wenn aufeinanderfolgende Zeichen unabhängig voneinander sind, ist H einfach nur $-\Sigma p_i \log p_i$, wobei p_i die Wahrscheinlichkeit des Zeichens i ist. Angenommen, wir betrachten in diesem Fall eine lange Nachricht von N Zeichen; sie wird mit einer hohen Wahrscheinlichkeit etwa $p_1 N$ mal das erste Zeichen enthalten, $p_2 N$ mal das zweite usw. Dadurch wird die Wahrscheinlichkeit dieser besonderen Nachricht ungefähr

$$p = p_1^{p_1 N} \, p_2^{p_2 N} \cdots p_n^{p_n N}$$

sein oder

$$\log p \cong N \sum_i p_i \log p_i$$

$$\log p \cong -NH$$

$$H \cong \frac{\log 1/p}{N} .$$

H ist daher näherungsweise der Logarithmus der reziproken Wahrscheinlichkeit einer typisch langen Folge, dividiert durch die Anzahl der Zeichen in dieser Folge. Dieses Ergebnis gilt für jede Quelle. Um es genauer darzustellen, haben wir (siehe Anhang 3):

Lehrsatz 3: *Wenn ein beliebiges $\epsilon > 0$ und $\delta > 0$ gegeben ist, können wir ein N_0 finden, so daß die Folgen einer jeden Länge $N \geq N_0$ in zwei Kategorien fallen:*

1. *eine Menge, deren Gesamtwahrscheinlichkeit kleiner als ϵ ist,*
2. *den Rest, dessen sämtliche Elemente Wahrscheinlichkeiten besitzen, die der Ungleichung*

$$\left| \frac{\log p^{-1}}{N} - H \right| < \delta$$

genügen.

Mit anderen Worten: Wir sind fast sicher, daß $\frac{\log p^{-1}}{N}$ sehr nahe an H herankommt, wenn N groß ist.

Ein ähnliches Ergebnis behandelt die Anzahl von Folgen verschiedener Wahrscheinlichkeiten. Betrachten wir nochmals die Folgen der Länge N und ordnen wir sie in der Reihenfolge sinkender Wahrscheinlichkeiten. Wir definieren nun $n(q)$ als die Anzahl, die wir dieser Menge entnehmen müssen, wobei wir mit der wahrscheinlichsten beginnen, um für diejenigen, die genommen wurden, eine Gesamtwahrscheinlichkeit q zu ermitteln.

Lehrsatz 4:

$$\operatorname{Lim}_{N \to \infty} \frac{\log n(q)}{N} = H ,$$

wenn q nicht gleich 0 oder 1 ist.

Wir können $\log n(q)$ als die Anzahl der bit interpretieren, die erforderlich ist, um eine Folge darzustellen, wenn wir nur die wahr-

scheinlichsten Folgen mit einer Gesamtwahrscheinlichkeit q berücksichtigen. Dann ist $\frac{\log n(q)}{N}$ die Anzahl der bit pro Zeichen für diese Darstellung. Der Lehrsatz besagt, daß für großes N der Wert unabhängig von q und gleich H sein wird. Die Rate, mit der der Logarithmus der Anzahl von ziemlich wahrscheinlichen Folgen wächst, ist durch H gegeben, ungeachtet unserer Interpretation von "ziemlich wahrscheinlich". Dank dieser Ergebnisse, die in Anhang 3 bewiesen werden, ist es für die meisten Anwendungen möglich, lange Folgen so zu behandeln, als gäbe es nur 2^{HN} davon, jede mit einer Wahrscheinlichkeit von 2^{-HN}.

Die folgenden zwei Lehrsätze zeigen, daß H und H' durch Grenzwertprozesse bestimmt sind, die direkt aus der Statistik der Nachrichtenfolgen hervorgehen, ohne Verbindung zu den Zuständen und den Übergangswahrscheinlichkeiten zwischen diesen Zuständen.

Lehrsatz 5: *$p(B_i)$ sei die Wahrscheinlichkeit einer Folge B_i von Zeichen der Quelle. Es sei*

$$G_N = -\frac{1}{N}\sum_i p(B_i)\log p(B_i)\,,$$

wobei sich die Summe über alle Folgen B_i erstreckt, die N Zeichen enthalten. Dann ist G_N eine monoton fallende Funktion von N und

$$\lim_{N\to\infty} G_N = H.$$

Lehrsatz 6: *$p(B_iS_j)$ sei die Wahrscheinlichkeit der Folge B_i, gefolgt von dem Zeichen S_j, und $p_{B_i}(S_j) = p(B_i, S_j)/p(B_i)$ sei die bedingte Wahrscheinlichkeit, mit der S_j der Folge B_i folgt. Es sei*

$$F_N = -\sum_{i,j} p(B_i, S_j)\log p_{B_i}(S_j)\,,$$

wobei sich die Summe über alle Blöcke B_i von $N-1$ Zeichen und alle Zeichen S_j erstreckt. Dann ist auch F_n eine monoton fallende Funktion von N,

$$F_N = NG_N - (N-1)\,G_{N-1},$$

$$G_N = \frac{1}{N}\sum_1^N F_N,$$

$$F_N \leq G_N,$$

mit $\lim_{N\to\infty} F_N = H.$

Diese Ergebnisse sind im Anhang 3 abgeleitet. Sie zeigen, daß eine Serie von Näherungen an H dadurch erreicht werden kann, daß man nur die statistische Struktur der Folgen, die 1, 2 ... , N Zeichen enthalten, berücksichtigt. F_N ist die bessere Näherung. In der Tat ist F_N die Entropie der N-ten Näherung an die Quelle des oben erwähnten Types. Wenn keine statistischen Einflüsse vorhanden sind, die sich über mehr als N Zeichen erstrecken, d.h. wenn die bedingte Wahrscheinlichkeit des nächsten Zeichens, falls man die vorhergehenden $(N-1)$ kennt, durch die Kenntnis eines noch weiter vorher aufgetretenen Zeichens nicht geändert wird, dann ist $F_N = H$. F_N ist natürlich die bedingte Entropie des nächsten Zeichens, falls die $(N-1)$ vorausgegangenen bekannt sind, während G_N die Entropie pro Zeichen der Blöcke mit N Zeichen ist.

Das Verhältnis der tatsächlichen Entropie einer Quelle zum Maximalwert, den sie haben könnte, wobei sie noch auf dieselben Zeichen beschränkt ist, wird ihre *relative Entropie* genannt. Das ist, wie später gezeigt wird, die höchstmögliche Informationsdichte, wenn wir mit demselben Alphabet codieren. "Eins" minus die relative Entropie ist die *Redundanz*. Die Redundanz des gewöhnlichen Englisch, ohne daß dabei die statistische Struktur über Zeichenfolgen von mehr als acht Buchstaben berücksichtigt wird, ist etwa 50%. Wenn wir also Englisch schreiben, ist die Hälfte des Geschriebenen durch die Struktur der Sprache bestimmt, und die andere Hälfte ist frei gewählt. Der Wert von 50% wurde durch verschiedene, voneinander unabhängige Methoden herausgefunden, die alle Ergebnisse in diesem Bereich erzielten. Eine davon war die Abschätzung der Entropie der Näherungen an das Englische. Eine zweite Methode ist, einen bestimmten Teil der Buchstaben aus einem englischen Text herauszunehmen, und dann jemanden zu bitten, sie wieder einzusetzen. Falls sie wieder eingesetzt werden können, wenn 50% herausgenommen wurden, muß die Redundanz größer als 50% sein. Eine dritte Methode ergibt sich aus gewissen bekannten Resultaten bei der Entschlüsselung von Geheimschriften.

Zwei Extreme von Redundanz in der englischen Prosa sind durch das "Basic English" und durch das Buch *Finnegans Wake* von James Joyce repräsentiert. Das "Basic English"-Wörterverzeichnis ist auf 850 Worte begrenzt, und die Redundanz ist darin sehr hoch. Dies zeigt sich in der Aufweitung, die sich ergibt, wenn ein Textabschnitt in "Basic English" übertragen wird. Dagegen wird von Joyce behauptet, er habe dadurch, daß er sein Vokabular erweiterte, eine Verdichtung des semantischen Gehalts erreicht.

Mit der Redundanz einer Sprache hängt die Existenz von Kreuzworträtseln zusammen. Wenn die Redundanz gleich Null ist, ergibt jede Folge von Buchstaben einen vernünftigen Text in dieser Sprache, und jede zweidimensionale Anordnung von Buchstaben ergibt ein Kreuzworträtsel. Wenn die Redundanz zu hoch ist, werden durch die Sprache zu viele Einschränkungen auferlegt, um ein größeres Kreuzworträtsel zu ermöglichen. Eine detailliertere Analyse zeigt, daß, wenn wir die von der Sprache auferlegten Beschränkungen als ziemlich verworren und zufällig voraussetzen, große Kreuzworträtsel gerade noch möglich sind bei einer Redundanz von 50%. Ist die Redundanz 33%, müßten eigentlich dreidimensionale Kreuzworträtsel möglich sein, usw.

8. Beschreibung der Codierung und der Decodierung

Wir müssen mathematisch noch die Operationen darstellen, die von Sender und Empfänger beim Codieren und Decodieren der Information durchgeführt werden. Zu beidem benötigt man einen diskreten Wandler. Die Eingabe in den Wandler (Codierer/Decodierer) ist eine Folge von Eingabezeichen und seine Ausgabe eine Folge von Ausgabezeichen. Der Wandler kann einen internen Speicher haben, so daß seine Ausgabe nicht nur vom gegenwärtigen Eingabezeichen abhängt, sondern auch von der Vorgeschichte. Wir setzen voraus, daß der interne Speicher beschränkt ist, d.h. es existiert nur eine begrenzte Anzahl m möglicher Zustände des Wandlers und seine Ausgabe ist jedesmal eine Funktion des momentanen Zustandes und des gegenwärtigen Eingabezeichens. Der nächste Zustand wird eine zweite Funktion dieser beiden Größen sein. Daher kann ein Wandler durch zwei Funktionen beschrieben werden:

$$y_n = f(x_n, \alpha_n)$$
$$\alpha_{n+1} = g(x_n, \alpha_n),$$

wobei: x_n das n-te Eingabezeichen ist,

α_n der Zustand des Wandlers ist, wenn das n-te Eingabezeichen ankommt,

y_n das Ausgabezeichen (oder die Folge von Ausgabezeichen) ist, das erzeugt wird, wenn x_n im Zustand α_n auftritt.

Wenn die Ausgabezeichen eines Wandlers zugleich die Eingabezeichen eines zweiten Wandlers sind, dürfen diese als "Tandem" verbunden werden und ergeben wiederum einen Wandler. Wenn ein zweiter Wandler vorhanden ist, der mit den Ausgaben des ersten arbeitet und die ursprüngliche Eingabe reproduziert, wird der erste Wandler als "nicht-singulär" bezeichnet und der zweite wird seine Umkehrung genannt.

Lehrsatz 7: *Die Ausgabe eines begrenzten Zustandswandlers, der von einer statistischen Quelle mit begrenzten Zuständen gesteuert wird, stellt selbst eine statistische Quelle mit begrenzten Zuständen dar, und zwar mit einer Entropie (pro Zeiteinheit), die kleiner oder gleich jener der Eingabe ist. Wenn der Wandler nicht-singulär ist, dann sind beide Entropien gleich.*

Nehmen wir an, daß α den Zustand einer Quelle darstellt, die eine Folge von Zeichen x_i erzeugt, und daß β der Zustand des Wandlers ist, der in seiner Ausgabe Blöcke von Zeichen y_j erzeugt. Das kombinierte System kann durch Erzeugung von Paaren (α, β) im Zustandsraum dargestellt werden. Zwei Punkte im Raum, (α_1, β_1) und (α_2, β_2), sind durch eine Linie verbunden, falls α_1 ein Zeichen x herstellen kann, das β_1 in β_2 verwandelt. Dieser Verbindungslinie wird in diesem Fall die Wahrscheinlichkeit des Zeichens x zugeordnet. Sie ist außerdem mit einem Block von y_1 Zeichen, die vom Wandler erzeugt werden, gekennzeichnet. Die Entropie der Ausgabe kann nun als die gewichtete Summe der Zustände berechnet werden. Wenn wir zuerst über β summieren, so ist jeder erhaltene Ausdruck kleiner oder gleich dem entsprechenden Ausdruck für α, also vergrößert sich die Entropie nicht. Wenn der Wandler nicht-singulär ist, läßt man seine Ausgabe mit dem inversen Wandler verbinden. Wenn H_1', H_2' und H_3' die Ausgabeentropien der Quelle bzw. des ersten und zweiten Wandlers sind, dann gilt $H_1' \geq H_2' \geq H_3' = H_1'$ und damit $H_1' = H_2'$.

Wir setzen nun voraus, wir hätten ein System von Bedingungen für mögliche Folgen der Art, daß die Bedingungen durch einen linearen Graphen nach Abb. 2 dargestellt werden können. Wenn die Wahrscheinlichkeiten $p_{ij}^{(s)}$ den verschiedenen Linien zugewiesen würden, die den Zustand i mit dem Zustand j verbinden, dann wäre dies eine Quelle. Es gibt dann eine ganz bestimmte Zuweisung, die die sich ergebende Entropie maximiert (siehe Anhang 4).

Lehrsatz 8: *Ein System von Bedingungen, das als Kanal angesehen wird, möge eine Kapazität $C = \log W$ haben. Wenn wir festsetzen, daß*

$$p_{ij}^{(s)} = \frac{B_j}{B_i} W^{-l_{ij}^{(s)}},$$

wobei $l_{ij}^{(s)}$ die Dauer für das Zeichen s ist, das vom Zustand i zum Zustand j führt, und B_i folgende Bedingung erfüllt:

$$B_i = \sum_{s,j} B_i W^{-l_{ij}^{(s)}},$$

dann erreicht H den Maximalwert, der gleich ist der Kanalkapazität C.

Durch die richtige Zuweisung der Übergangswahrscheinlichkeiten kann die Entropie der Zeichen auf einem Kanal dem Maximalwert, der Kanalkapazität, angenähert werden.

9. Der fundamentale Lehrsatz für einen störungsfreien Kanal

Wir werden nun unsere Interpretation von H als der erzeugten Informationsrate rechtfertigen, indem wir beweisen, daß H zusammen mit der effektivsten Codierung die geforderte Kanalkapazität bestimmt.

Lehrsatz 9: *Angenommen, die Quelle habe die Entropie H (bit pro Zeichen) und der Kanal eine Kapazität C (bit pro Sekunde). Dann ist es möglich, die Ausgabe der Quelle so zu codieren, daß man durchschnittlich $\frac{C}{H} - \epsilon$ Zeichen pro Sekunde über den Kanal überträgt, wobei ϵ beliebig klein ist. Es ist nicht möglich, eine größere Durchschnittsrate als $\frac{C}{H}$ zu übertragen.*

Die Aussage des Lehrsatzes, daß $\frac{C}{H}$ nicht überschritten werden kann, läßt sich beweisen, wenn man beachtet, daß die Entropie der Kanaleingabe pro Sekunde gleich derjenigen der Quelle ist, da der Sender nicht-singulär sein muß und auch diese Entropie die Kanalkapazität nicht übersteigen kann. Daher ist $H' \leq C$ und die Anzahl von Zeichen pro Sekunde $= H'/H \leq C/H$.

Der erste Teil des Lehrsatzes wird auf zwei verschiedene Arten bewiesen. Die erste Methode ist die, die Menge aller Folgen von N Zeichen, die von der Quelle erzeugt werden, zu berücksichtigen. Wenn N groß ist, kann man diese in zwei Gruppen aufteilen, wobei eine weniger als $2^{(H+\eta)N}$ Elemente und die zweite weniger als 2^{RN} Elemente enthält (wobei R der Logarithmus der Anzahl der verschiedenen Zeichen ist) und eine Gesamtwahrscheinlichkeit kleiner als μ hat. Wenn N größer wird, nähern sich η und μ dem Wert

Null. Die Anzahl von Signalen der Dauer T im Kanal ist größer als $2^{(C-\theta)T}$ mit kleinem θ, wenn T groß ist.
Wenn wir

$$T = \left(\frac{H}{C} + \lambda\right) N$$

wählen, dann wird es, wenn N und T genügend groß sind (λ jedoch klein ist), eine ausreichende Anzahl an Folgen von Kanalzeichen für die hohe Wahrscheinlichkeitsgruppe geben und auch einige zusätzliche. Die hohe Wahrscheinlichkeitsgruppe ist willkürlich eins-zu-eins in diesem Satz codiert. Die Elemente der 2. Gruppe werden von längeren Folgen dargestellt, die mit einer der Folgen, die für die hohe Wahrscheinlichkeitsgruppe nicht benützt werden, beginnen und enden. Diese besondere Folge bewirkt ein Anfangs- und End-Signal für einen anderen Code. Dazwischen wird ausreichend Zeit für genügend verschiedene Folgen für alle die Nachrichten mit niedriger Wahrscheinlichkeit gewährt. Das macht es erforderlich, daß

$$T_1 = \left(\frac{R}{C} + \varphi\right) N,$$

wobei φ klein ist. Die mittlere Übertragungsrate in Zeichen pro Sekunde wird dann größer sein als

$$\left[(1-\delta)\frac{T}{N} + \delta\frac{T_1}{N}\right]^{-1} = \left[(1-\delta)\left(\frac{H}{C}+\lambda\right) + \delta\left(\frac{R}{C}+\varphi\right)\right]^{-1}.$$

Mit wachsendem N nähern sich δ, λ und φ dem Wert Null, und die Rate nähert sich $\frac{C}{H}$.

Eine andere Methode, diese Codierung durchzuführen und dabei den Lehrsatz zu beweisen, kann wie folgt beschrieben werden: Man ordne die Nachrichten mit der Länge N nach sinkenden Wahrscheinlichkeiten und lege ihre Wahrscheinlichkeiten als $p_1 \geq p_2 \geq p_3 \cdots \geq p_n$ fest.
Es sei

$$P_s = \sum_{1}^{s-1} p_i.$$

Das heißt, P_s ist die aufsummierte Wahrscheinlichkeit bis, aber nicht inklusive p_s. Zuerst codieren wir in ein Binärsystem um. Der Binärcode für die Nachricht s wird erreicht, indem P_s in eine binäre Zahl entwickelt wird. Die Entwicklung wird bis m_s Stellen durchgeführt, wobei m_s die ganze Zahl ist, die der folgenden Ungleichung genügt:

$$\log_2 \frac{1}{p_s} \leq m_s < 1 + \log_2 \frac{1}{p_s}.$$

Dadurch werden die Nachrichten mit hoher Wahrscheinlichkeit durch kurze Codierung repräsentiert und jene mit niedriger Wahrscheinlichkeit durch lange Codes. Durch diese Ungleichheiten erhalten wir

$$\frac{1}{2^{m_s}} \leq p_s < \frac{1}{2^{m_s-1}}.$$

Der Code für P_s wird sich von allen nachfolgenden in einer oder mehrerer seiner m_s Stellen unterscheiden, da alle verbleibenden kumulierten Wahrscheinlichkeiten P_i mindestens $\frac{1}{2^{m_s}}$ größer sind und ihre binäre Entwicklung sich daher in den ersten m_s Stellen unterscheidet. Daraus folgt, daß alle Codes verschieden sind, und es ist möglich, die Nachricht aus ihrem Code wiederzugewinnen. Wenn die Kanalfolgen nicht bereits Folgen von Binärziffern sind, können ihnen irgendwelche binäre Zahlen willkürlich zugeschrieben werden, und der Binärcode kann damit in für den Kanal passende Signale übersetzt werden. Die Durchschnittszahl H_1 von Binärstellen, die pro Zeichen der Originalnachricht gebraucht werden, ist leicht abzuschätzen. Wir haben

$$H_1 = \frac{1}{N} \Sigma m_s p_s.$$

Aber es ist auch

$$\frac{1}{N} \Sigma \left(\log_2 \frac{1}{p_s}\right) p_s \leq \frac{1}{N} \Sigma m_s p_s < \frac{1}{N} \Sigma \left(1 + \log_2 \frac{1}{p_s}\right) p_s$$

und daher

$$G_N \leq H_1 < G_N + \frac{1}{N}.$$

Mit wachsendem N nähert sich G_N und H_1 der Entropie H der Quelle.

Daraus ersehen wir, daß die Unwirksamkeit in der Codierung, falls nur eine begrenzte Verzögerung der N Zeichen benutzt wird, nicht größer als $\frac{1}{N}$ zu sein braucht, plus der Differenz zwischen der wirklichen Entropie H und der Entropie G_N, welche für die Folgen der Länge N berechnet wurde. Der prozentuale Zeitzuschlag, der über das Ideal hinaus benötigt wird, ist daher geringer als

$$\frac{G_N}{H} + \frac{1}{HN} - 1.$$

Diese Methode der Codierung ist im wesentlichen dieselbe, wie sie unabhängig davon durch R. M. Fano [9]) herausgefunden wurde. Bei seiner Methode werden die Nachrichten der Länge N nach sinkender Wahrscheinlichkeit geordnet. Dann teilt man die Folgen in zwei Gruppen, die eine möglichst gleiche Wahrscheinlichkeit besitzen sollen. Wenn eine Nachricht in der ersten Gruppe ist, wird ihre erste Binärziffer 0 sein, sonst 1. Die Gruppen werden nun ähnlich in Untergruppen wiederum mit nahezu gleicher Wahrscheinlichkeit aufgeteilt, und die einzelne Untergruppe bestimmt die zweite Binärziffer. Dieser Prozeß wird fortgesetzt, bis jede Untergruppe nur eine einzige Nachricht enthält. Es kann leicht durchschaut werden, daß, abgesehen von kleineren Unterschieden (allgemein für die letzte Stelle), dies auf dasselbe Ergebnis hinausläuft, wie der arithmetische Prozeß, der oben beschrieben wurde.

10. Diskussion und Beispiele

Um eine maximale Energieübertragung von einem Generator zu einem Verbraucher zu erhalten, muß ein Umformer im allgemeinen so eingefügt werden, daß der Innenwiderstand des Generators die Größe des Lastwiderstandes hat. Die Situation hier ist ähnlich dazu. Der Wandler, der die Codierung vornimmt, sollte im statistischen Sinn die Quelle dem Kanal anpassen. Die Quelle, wie sie vom Kanal über den Wandler gesehen wird, sollte dieselbe statistische Struktur besitzen wie jene Quelle, die die Entropie im Kanal maximiert. Der Inhalt von Lehrsatz 9 ist der, daß man eine Quelle im allgemeinen nicht exakt, jedoch beliebig gut anpassen kann. Das Verhältnis der wirklichen Übertragungsrate zu der Kapazität C kann als Effizienz des Codiersystems bezeichnet werden. Dies ist natürlich dasselbe wie das Verhältnis der tatsächlichen Entropie der Kanalzeichen zur maximal möglichen Entropie.

Im allgemeinen verursacht die ideale oder nahezu ideale Codierung eine lange Verzögerung im Sender und im Empfänger. Im störungsfreien Fall, den wir hier behandeln, ist die Hauptaufgabe dieser Verzögerung, eine einigermaßen gute Anpassung der Wahrscheinlichkeiten an die entsprechenden Längen der Folgen zu erlauben. Bei einem guten Code muß der Logarithmus der reziproken Wahrscheinlichkeit einer langen Nachricht proportional zu der Dauer des

[9]) Technical Report No. 65, The Research Laboratory of Electronics, M. I. T., 17. März 1949.