Klaus Nagel

Sprache, Information, Verschlüsselung

Übersicht

In diesem Vortrag wird dargestellt, wie Sprache Information zwischen Menschen überträgt, wo und wie die Informationen verschlüsselt werden und wie sich Information messen läßt.

Verschlüsselung bedeutet hier Umsetzen einer Nachricht in eine andere Form, die eine Übertragung überhaupt erst ermöglicht, die Nachricht gegen Fehler und Störungen beim Übermitteln sichert oder auch gegen unerwünschtes Abhören schützt. Bei der Sicherung wird die Bedeutung der Redundanz gezeigt, die dann zum Informationsmaß Entropie führt. Bei der Verschlüsselung wird das Verfahren öffentlicher Schlüssel erläutert. Die Wahl dieses Themas führt zur Einschränkung auf objektiv meßbare Größen der Sprache, wie etwa Anzahl und Häufigkeitsverteilung von Buchstaben und Zeichen; subjektive Aspekte wie etwa Ästhetik, Betonung und Stil werden nicht berücksichtigt.

Sprache

Gedankenübertragung, nämlich Gedanken aus dem Gehirn eines Menschen in das eines anderen zu übertragen, ist unmittelbar nicht möglich. Wir bedienen uns deshalb unserer Organe, um andere über unsere Gehirnvorgänge zu unterrichten: Der Sendende macht Gesten, die der Empfänger mit dem Auge erkennt, oder – und das ist der häufigere Fall – er spricht und sein Partner hört ihn. Schon Sprechen und Hören erfordert mehrere Verschlüsselungen:

Das Hirn setzt die Gedanken in Wörter und Sätze um.
Die Sprechwerkzeuge – Lunge, Kehlkopf, Rachenraum, Zunge, Lippen - wandeln die – Wörter zu Schall.
Der Schall durcheilt die Luft und trifft auf das Ohr des Hörers; das erkennt die Wörter und Sätze.
Sein Hirn ordnet dem Gehörten wieder Bedeutung zu, der Gedanke ist übertragen.
So trivial dieser Vorgang auch erscheinen mag, er läßt uns etwas Wesentliches erkennen:
Die Übertragung mit Schall arbeitet bei allen Menschen gleich, Sprechapparat und Ohr sind angeboren und aufeinander abgestimmt. Das Umsetzen der Gedanken in Wörter und deren Interpretation beim Hörer aber setzt

voraus, daß beide Partner den Schlüssel kennen, daß sie die gleiche Sprache sprechen! Die Existenz der unterschiedlichen Sprachen auf der Erde zeigt, daß unzählige Verschlüsselungen möglich sind, die Zuordnung von Begriffen und Wörtern ist willkürlich. In der Kindheit wird ein bestimmter Schlüssel dadurch festgelegt, daß Gegenstände und Begriffe wiederholt mit Worten bezeichnet werden; durch Assoziation erlernt man so die Muttersprache.

Statt Wörter lassen sich den Begriffen auch Bilder oder Zeichen zuordnen, so kommt man zu einer Bilderschrift, die unabhängig von der Sprache verständlich ist. Die chinesische Schrift ist vor langer Zeit in das Japanische übernommen worden und daher für Japaner großteils noch lesbar, obwohl sich die Sprachen verschieden weiterentwickelt haben. Auch die modernen Verkehrszeichen und Piktogramme sind ohne Sprachkenntnisse zu begreifen. Die mündliche Sprache war für den Augenblick bestimmt, Sprecher und Hörer müssen am gleichen Ort sein; erst die technische Entwicklung der letzten hundert Jahre ermöglichte eine Aufzeichnung und weltweite Übertragung von Schall. Symbole und Bilder sind etwas Bleibendes, als Inschrift oder Plakat kann eine Botschaft an beliebig viele Menschen gebracht werden, als Brief überallhin verschickt werden. Andererseits verschwindet beim Schreiben vieles, was dem Hörer beim direkten Gespräch zusätzliche Information gibt: Aus Betonung, Mimik und Gestik, aus zögernder oder schneller Sprechweise erfährt der anwesende Gesprächspartner manches Zusätzliche über Hintergründe oder Glaubwürdigkeit des Gesagten und des Sprechers.

Ein Nachteil einer Bilderschrift ist, daß nun zwei Schlüssel nötig sind Gedanken umzusetzen, einer für das Bild, ein anderer für das gesprochene Wort. Eine große kulturelle Leistung war die Erfindung der Lautschrift vor etwa dreitausend Jahren als die Akkader begannen, sumerische Bildzeichen als Lautzeichen zu gebrauchen (von Weizsäcker, 1959, 42). Damit entfiel – oder vereinfachte sich zumindest – einer der beiden Schlüssel: Das gesprochene Wort konnte jetzt ganz schematisch in Zeichen umgesetzt werden. Dieses Verfahren lernen wir heute in den ersten Schuljahren und ein zehnjähriges Kind ist dadurch in der Lage, auch nie gehörte Wörter zu lesen, und es kann sie nach dem Klang niederschreiben, ohne ihre Bedeutung zu kennen. Probleme ergeben sich dadurch, daß sich die geschriebene Sprache langsamer als die gesprochene an Veränderungen anpaßt. So hat zum Beispiel das gesprochene Englisch nach der Festlegung der Schreibweise noch starke Wandlungen durchgemacht, so daß es weit schwieriger ist als im Deutschen, aus geschriebenem Text auf die Aussprache zu schließen. Die Aufnahme von Wörtern aus anderen

Sprachen mindert auch im Deutschen die klare Zuordnung zwischen Aussprache und Schreibweise.

Physikalische Übertragungsmöglichkeiten:

Mechanisch:
Brief, Plakat, Buchdruck

Akustisch:
Morsezeichen, Glocken, Buschtrommeln, Sirenen, Hupen

Optisch:
Lichtsignale, Flaggen, Semaphore, Rauchzeichen, Handzeichen, Glasfasertechnik.

Elektrisch:
Fernschreiber, Fax, Klingel, Telegraph, Telefon, Rundfunk

Für die Übertragung von Zeichen bieten sich, wie oben angeführt, viele Möglichkeiten an. Für die Schallübertragung gab es lange nur das unmittelbare Sprechen und Hören, bis dann in den letzten 150 Jahren durch die Erfindung von Telefon und Übertragung durch elektrische Wellen akustische Verständigung über große Entfernungen möglich wurde. Die Verschlüsselung der Schallsignale ist beim Telefon sehr einfach: Ein Mikrofon erzeugt einen Strom, dessen Stärke dem wechselnden Schalldruck folgt. Bei der Funkübertragung kommt eine weitere Verschlüsselungsstufe hinzu, das elektrische Sprachsignal beeinflußt - moduliert - eine elektrische Welle hoher Frequenz, den sogenannten Träger, der große Entfernungen überbrückt. Der Empfänger muß die Verschlüsselungen wieder rückgängig machen. Dazu muß er die Schlüssel kennen, in diesem Fall die Frequenz des Senders und die Art der Modulation. Damit gewinnt er aus den empfangenen Wellen das elektrische Signal zurück, dieses wird im Lautsprecher wieder in Schall umgesetzt, der dann – genau wie im direkten Gespräch – durch Ohr und Gehirn erkannt wird.

Digitalisierung

Die vielfältigen Möglichkeiten bei der Übertragung und Speicherung von Zeichen lassen sich auch beim Schall nutzen, wenn man diesen als eine Zeichenfolge verschlüsselt. Solche Verfahren setzen sich in jüngster Zeit mehr und mehr in der Digitaltechnik durch. Das dem Schalldruck analoge elektrische Signal wird dazu in regelmäßigen Zeitabständen abgetastet, die augenblickliche Stärke gemessen und als Zahlenwert gespeichert. Die

Abtastrate muß mindestens doppelt so groß sein wie die höchste vorkommende Frequenz. Bei Telefonübertragungen reichen 8000 Abtastungen pro Sekunde aus, bei den heute gebräuchlichen Compact Discs (CDs) für Musikaufnahmen hoher Qualität geschieht dieses 44100 mal in jeder Sekunde, denn das Ohr junger Menschen reicht bis zu Frequenzen von etwa 20000 Schwingungen pro Sekunde. Die Signalstärke wird dabei auf fünf Dezimalstellen genau gemessen und das für den rechten und den linken Stereokanal. Diese Genauigkeit ist so groß, als gäbe man Entfernungen bis zu sechzig Metern auf Millimeter genau an. Die Speicherung und Bearbeitung solch ungeheurer Datenmengen wurde erst möglich mit der Miniaturisierung und der damit verbundenen Geschwindigkeitssteigerung in der modernen Elektronik.

Binäre Kodierung

Üblicherweise benutzen wir für Zahlenangaben das Dezimalsystem; es verwendet nur zehn verschiedene Ziffern, die den Zahlen von Null bis Neun entsprechen. Größere Zahlen werden durch die Angabe von Einern, Zehnern, Hundertern und so fort ausgedrückt, Werte die sich jeweils um einen Faktor Zehn unterscheiden. Daß wir gerade die Zehn als Basis unseres Zahlensystems benutzen, kommt von unseren zehn Fingern, diese Wahl ist nicht zwingend. Für Maschinen ist Zwei als Basis viel günstiger: Bei diesen Binärzahlen gibt es nur noch zwei verschiedenen Ziffern - Null und Eins - und die können technisch auf viele Weisen repräsentiert werden:

Loch oder Papier bei Lochkarte und Lochstreifen,
Nord- oder Südpol bei magnetischen Datenträgern,
Punkt oder Strich beim Morsen und bei optischen Speichermedien wie zum Beispiel der Compact Disc,
entladene oder geladene Kondensatoren in Computerspeichern.

Die Wertigkeiten der Stellen einer Binärzahl verdoppeln sich jeweils, haben also die Werte 1, 2, 4, 8, 16 ... Die Dezimalzahl Dreizehn etwa hat die binäre Darstellung 1101, denn es gilt 13 = 1*8 + 1*4 + 0* 2 + 1*1. Den Vorteil mit nur zwei Ziffern auszukommen, bezahlt man mit einer größeren Stellenzahl gegenüber dem Dezimalsystem. Für eine Binärstelle, die entweder Eins oder Null ist , wurde das Kunstwort „Bit" (Binary Digit) eingeführt.

Redundanz

Den verstümmelten Text des Bildes 1 konnten die Teilnehmer in Matrei vollständig rekonstruieren. Das war nur möglich, weil Redundanz vorhan-

den ist, „überflüssige“ Wörter, die den Text gegen Zerstörung und Verfälschung schützen.

„He, Kyklop, wenn einer ■■■■■■■■gt von den sterbliche■■■■■■hen,
wer denn gar so greulich das einzige■■■■■■ir zerrüttet.
Sag: Es blende■■■■■■ch der Städtezerstö■■■■■■■■■■■■■■
des Lae■■■■■■■■■■, der in ■■■■■■■■■■imat und Haus hat.
(Odyssee, Neunter Gesang)

Bild 1

Vielfältiges Wissen mußte eingebracht werden, die verdeckten Buchstaben zu bestimmen. Der Text stammt aus dem neunten Gesang der Odyssee in der Übersetzung von Rudolf Alexander Schröder. Wer die ersten acht Gesänge gelesen hat, kennt die Beinamen der Helden, weiß, daß der Städtezerstörer Odysseus ist und daß sein Vater Laertes auf Ithaka herrscht. Aus den Wörtern „Kyklop“ und „blendete“ läßt sich auf das „einzige Auge“ in der zweiten Zeile schließen, vom Versmaß her aber kann es nur „Aug“ heißen. Der vollständige Text lautet:

„He, Kyklop, wenn einer dich fragt von den sterblichen Menschen,
wer denn gar so greulich das einzige Aug dir zerrüttet.
Sag ihm, es blendete mich der Städtezerstörer Odysseus,
des Laertes Sohn, der in Ithaka Heimat und Haus hat.“

Ein technisches Verfahren zum Sichern der Nachrichtenübertragung kann man nicht auf solch subjektives Wissen aufbauen, schließlich wäre man schon ohne Kenntnis der deutschen Sprache in diesem Fall völlig hilflos, und ein Verständnis für den Sprachrhythmus wäre einer Maschine kaum zu vermitteln. Besser geeignet ist da schon Redundanz, so wie sie sich im zweiten Beispiel zeigt. Jeder kann leicht den unleserlichen Preis für das Brot ausrechnen, schließlich besteht eine eindeutige und wohlbekannte Beziehung zwischen Summanden und Summe.

Mehl	3,70
Zucker	2,50
Brot	■■■
Butter	3,20
Summe	14,20

(Redundanz bei der Addition)

Bild 2

Nach ähnlicher Methode lassen sich binäre Nachrichten sichern. Typischerweise verschlüsselt man ein Zeichen als ein „Byte“ von acht Bits; das

bietet 256 Möglichkeiten, genug um große und kleine Buchstaben, Ziffern und Satzzeichen darzustellen. Zur Sicherung fügt man ein neuntes Bit an, derart daß das Zeichen jetzt eine ungerade Anzahl von Einsen enthält. Man spricht von ungerader Parität und nennt das angefügte Bit Paritätsbit. Tritt bei der Übertragung ein Fehler auf, der genau ein Bit verfälscht, so wird ein Zeichen mit einer geraden Anzahl von Einsen empfangen, das bedeutet: Der Fehler wird erkannt. Er kann aber nicht korrigiert werden, weil nicht festzustellen ist, welches Bit umgedreht wurde. Dem kann man abhelfen, indem man für die einzelnen Bitstellen mitzählt, wie oft Einsen vorkommen und am Ende der Nachricht ein Zeichen anfügt, so daß jetzt auch in jeder Spalte eine ungerade Anzahl von Einsen erscheint. Ein falsches Zeichen äußert sich dann in falscher Zeilenparität und falscher Spaltenparität (Bild 3). Solange höchstens ein Bit falsch übertragen wird, läßt es sich lokalisieren und korrigieren. Es gibt heute aufwendige Verfahren, auch Fehler in vielen Bits zu korrigieren. Das ist etwa bei den Compact Discs erforderlich, denn ein Kratzer oder Fingerabdruck kann viele der nur mikrometergroßen Bits verfälschen, ein Viertel aller Bits dient daher der Fehlererkennung und Korrektur.

	8 Informationsbits	1 Paritätsbit	Zahl der 1-Bits
1. Zeichen	00101100	0	3
2. Zeichen	11010111	1	7
3. Zeichen	00000000	1	1
4. Zeichen	01101100	1	5
Spaltenparität	01101000	0	3

(Sicherung durch Paritätsbits)

Bild 3

„Ich freue mich, daß ich am Montag, dem 19. Januar , nachmittags 5 Uhr in München eintreffen kann." Dieser Brief gibt dem Empfänger, der nur wissen will, wann er seinen Gast am Bahnhof abholen soll, nicht mehr Information als das Telegramm: „Eintreffe Montag 17 Uhr". Dabei sei stillschweigend angenommen, daß der Empfänger in München wohnt und daß mit Montag ohne weitere Datumsangabe der nächste Montag gemeint ist. Dieses Beispiel (von Weizsäcker,1959, 43ff.) zeigt, daß die Anzahl der Zeichen kein geeignetes Maß für die Information einer Nachricht sein kann.

Eine sinnvolle Nachricht verringert die Unsicherheit über einen Tatbestand. Diese Verringerung bildet die Grundlage für das Messen von Informationen. Ein Zahlenwert läßt sich nur angeben, wenn sich Tatbestand und Unsicherheit in Zahlen ausdrücken lassen. Bei dem Beispiel aus der Odys-

see kann man feststellen, daß der volle Text und der verstümmelte die gleiche Information enthalten, aus den geschwärzten Stellen erfährt man nichts Neues, sie ließen sich aus dem Rest wiedergewinnen. Diese Rekonstruktion hängt von den subjektiven Kenntnissen des Lesers ab, eignen sich nicht für eine objektive Bewertung, und es macht keinen Sinn, ein Maß für die Information dieser Verse anzugeben. Wir beschränken uns deshalb auf Informationen, die durch objektiv zu bestimmende Merkmale charakterisiert sind, wie zum Beispiel der Häufigkeitsverteilung von Zeichen oder Zeichengruppen. Der Informationsgewinn ist größer, wenn zuvor eine hohe Unsicherheit bestand. Der Ausgang einer Lottoziehung mit über 13 Millionen möglichen Ausgängen ist schwieriger vorauszusagen, als die Zahl der Augen beim einmaligen Würfeln mit nur sechs Möglichkeiten; noch weniger überrascht das Werfen einer Münze mit den beiden Ausgängen Kopf oder Zahl. Diese Beispiele legen es nahe, im Falle gleichwahrscheinlicher Ausgänge das Informationsmaß als eine Funktion f(a) zu wählen, wobei a die Anzahl der möglichen Ausgänge ist:

a = 2	Münzwurf
a = 6	Würfel
a =13 983 816	Lotto, 6 aus 49.

Betrachten wir ein Experiment, das aus zwei unabhängigen Versuchen zusammengesetzt ist, von denen der erste a_1, der zweite a_2 mögliche Ausgänge hat. Dann bietet der zusammengesetzte Versuch $a = a_1 * a_2$ Möglichkeiten, und es ist sinnvoll zu fordern, daß sein Informationsmaß gleich der Summe der beiden Teilinformationen ist:

$$f(a) = f(a_1 * a_2) = f(a_1) + f(a_2).$$

Beispiel: Das Experiment bestehe aus dem Werfen einer Münze (zwei Ausgänge) und anschließendem Würfeln (sechs Ausgänge). Dieses zusammengesetzte Experiment hat die zwölf möglichen Ausgänge:

(Kopf, 1) (Kopf, 2) (Kopf, 3) (Kopf, 4) (Kopf, 5) (Kopf, 6)
(Zahl, 1) (Zahl, 2) (Zahl, 3) (Zahl, 4) (Zahl, 5) (Zahl, 6)

Die Forderung an das Maß lautet in diesem Fall f(12) = f(2) + f(6).

Wir suchen also eine Funktion, die das Multiplizieren auf Addieren zurückführt. Das ist gerade die Eigenschaft, die beim logarithmischen Rechnen ausgenutzt wird, der Logarithmus ist das gesuchte Maß.:

$$f(a) = \log_b(a)$$

Die Basis b des Logarithmus kann frei gewählt werden. Sinnvoll ist die Basis Zwei, denn der Münzwurf, das einfachste Experiment der Wahrscheinlich-

keitsrechnung und das gerade einem Bit entspricht, liefert dann mit dem Wert f(2) = 1 die Einheit für das Informationsmaß. Den Logarithmus zur Basis zwei nennt man Logarithmus Dualis, abgekürzt ld. Ein Experiment mit nur einem möglichen Ausgang liefert keine Information, in Übereinstimmung mit der Definition des Maßes:

$f(1) = \log_2(1) = \mathrm{ld}(1) = 0.$

Zu Experimenten mit 2, 4, 8, 16, 32... gleichwahrscheinlichen Ausgängen gehören die Maßzahlen 1, 2, 3, 4, 5 usw.

Diese Definition mit Zwei als Basis läßt noch eine andere Interpretation zu. Das Maß ist die Mindestanzahl der Alternativfragen, die nötig sind, den Ausgang zu ermitteln. So läßt sich etwa aus einem Kartenspiel von 32 Karten eine ausgesuchte Karte mit fünf Fragen erraten, indem man die verbliebenen Möglichkeiten jeweils halbiert:

1. Frage: „Ist die Karte rot ?"
2. Frage, wenn die erste Frage bejaht wurde: „Ist es eine Herzkarte?" Sonst: „Ist es eine Kreuzkarte?"

Von den ursprünglich 32 Möglichkeiten sind jetzt nur noch acht übrig. Mit drei weiteren Fragen identifiziert man die gesuchte Karte.

Bisher gilt die Definition des Informationsmaßes nur für gleichwahrscheinliche Ereignisse, eine Ausdehnung auf den allgemeinen Fall ergibt sich ganz natürlich, wenn man die bisherige Definition in folgender Form schreibt:

$f(a) = \mathrm{ld}(\frac{a}{1}) = 1/a*\mathrm{ld}(a) + 1/a*\mathrm{ld}(a) + \ldots + 1/a*\mathrm{ld}(a)$,

$1/a*\mathrm{ld}(a) = -1/a*\mathrm{ld}(1/a)$ ist der Beitrag eines der a Ausgänge zum Gesamtmaß, 1/a seine Wahrscheinlichkeit oder relative Häufigkeit. Dies legt folgende Erweiterung der Definition nahe:

$H = f(p_1, p_2, \ldots, p_a) = -\{ p_1* \mathrm{ld}(p_1) + p_2* \mathrm{ld}(p_2) + p_3* \mathrm{ld}(p_3) + \ldots. p_a* \mathrm{ld}(p_a) \}$

Dabei bezeichnen p_1, p_2, ...pa die Wahrscheinlichkeiten der a verschiedenen Ausgänge. Der so definierte Ausdruck H wird wegen seiner Analogie zu einer physikalischen Größe als Entropie bezeichnet. Der Begriff Entropie wurde von Clausius und Lord Kelvin um 1850 in der Thermodynamik als abstrakte Rechengröße eingeführt, erst später erkannte Boltzmann sie als Maß für die Wahrscheinlichkeit eines physikalischen Makrozustandes.

Was ist nun ein Experiment in der Nachrichtentechnik, was sind die Häufigkeiten und Wahrscheinlichkeiten? Das Experiment entspricht dem Empfangen einer Nachricht, beispielsweise eines deutschen Textes, der aus einzelnen Zeichen besteht. Nur endlich viele verschiedene Zeichen – Buchsta-

ben, Ziffern, Satzzeichen – werden benutzt, ihre Häufigkeiten sind höchst unterschiedlich, aber charakteristisch für die Sprache. In einem langen Text können wir die Häufigkeiten der einzelnen Zeichen zählen, als Maß ihrer Wahrscheinlichkeit hernehmen und nach der oben angegebenen Formel die Entropie berechnen. Nach dem bisher Gesagten wäre das eine reine Spielerei mit Zahlen. Die Nützlichkeit solcher Berechnungen zeigt der um 1950 von dem Amerikaner Shannon gefundene Fundamentalsatz der Kodierung (*Jaglom und Jaglom*, 163f.):

Zur Übertragung einer langen Nachricht aus M Zeichen in einer Sprache mit der Entropie H sind H*M Bits erforderlich.

Dieser Satz soll als Beispiel auf die Faksimile-Übertragung (Faxgerät) angewandt werden. Dabei wird eine Vorlage in eine endliche Zahl M von Punkten zerlegt, für jeden Punkt wird gemessen, ob er schwarz oder weiß ist, das Ergebnis wird übermittelt und beim Empfänger reproduziert. Die „Sprache" enthält also nur die beiden Zeichen „Schwarz" und „Weiß" , die mit einem Bit verschlüsselt werden können. Auf dem ersten Blick benötigt man so viele Bits wie Punkte zu übertragen sind. Üblicherweise wird aber mit schwarzer Farbe auf weißem Papier geschrieben, das Zeichen „Schwarz" ist also viel seltener als „Weiß". Nimmt man für Schwarz und Weiß die Wahrscheinlichkeiten 0,1 und 0,9 an, so ergibt sich die Entropie als

$$H = -(0{,}1 * \mathrm{ld}(0.1) + 0.9 * \mathrm{ld}(0.9)) = 0{,}469$$

Nach dem Fundamentalsatz erniedrigt sich daher die Zahl der nötigen Bits auf 0,469*M, auf weniger als die Hälfte der ursprünglich angenommenen Zahl. Um eine solche Verschlüsselung wirklich zu erreichen, faßt man Gruppen von Zeichen zusammen, gibt den häufig vorkommenden Kombinationen einen kurzen, den seltenen einen längeren Code. Die bessere Verschlüsselung senkt die Übertragungszeit und spart Telefonkosten. Dieses Beispiel läßt ahnen, welche Zeiten und Kosten bei weltweiten Übertragungen von Computerdaten oder Fernsehsendungen durch geeignete Kodierung einzusparen sind. Schon Samuel Morse hatte dieses erkannt, im Morse-Alphabet sind die häufigen Buchstaben „E" und „T" mit nur einem Punkt beziehungsweise mit nur einem Strich verschlüsselt, während die seltener vorkommenden Buchstaben aus bis zu vier Punkten und Strichen bestehen.

Zählt man in einem Text einer Sprache die Häufigkeiten der Buchstaben und ermittelt die Entropie, so zeigt sich, daß diese ziemlich unabhängig ist von der Wahl des Textes, hingegen charakteristisch für die Sprache. Bei diesen Untersuchungen unterscheidet man nicht zwischen großen und

kleinen Buchstaben, alle Ziffern und Satzzeichen interpretiert man als Zwischenraum, dargestellt als „_“, es kommen also 27 verschiedene Zeichen vor. Die folgende Aufstellung zeigt für drei Sprachen die häufigsten Buchstaben und die berechnete Entropie H_1. Für das Englische sind auch H_2 und H_3 angegeben, das sind die Entropien, die aus den Häufigkeiten von Buchstabenpaaren oder -tripeln berechnet wurden (*Jaglom & Jaglom*,191)

Deutsch:	_ E N I S T R A D	$H_1 = 4{,}10$		
Französisch:	_ E S I A N T U R	$H_1 = 3{,}96$		
Englisch :	_ E T A O N R I	$H_1 = 4{,}03$	$H_2 = 3{,}32$	$H_3 = 3{,}10$

Bild 4

Wären alle 27 Zeichen gleich häufig, so benötigte man für die Übertragung eines Textes ld(27) = 4,76 Bits pro Zeichen. Die Ausnutzung der Häufigkeitsverteilung einzelner Zeichen in deutschen Texten ermöglicht eine Verringerung auf 4,10 Bits pro Zeichen. Wie das Englische zeigt, läßt sich dieser Wert noch erheblich verkleinern, wenn die Verschlüsselung auf Zweier- oder Dreiergruppen angewendet wird. Den Wert H_8 schätzt man auf 1,9 für die englische Sprache, das heißt, die zur Übertragung erforderliche Bitmenge ließe sich mehr als halbieren.

Die Tabelle unten zeigt die Buchstabenverteilungen in fünf verschiedenen Texten:

Bibelübersetzung, Martin Luther,
Simplicius Simplicissimus (Auszug), Grimmelshausen,
Geschichte des Dreißigjährigen Kriegs (Auszug), Friedrich von Schiller,
Gladius Dei (Auszug), Thomas Mann,
Hamlet, William Shakespeare.

Die Textauszüge waren ziemlich klein, darin mag die Ursache liegen für Abweichungen gegenüber der bei Jaglom angegebenen Reihenfolge. Man sieht, daß sich die Verteilung in den deutschen Texten über die Jahrhunderte nur wenig geändert hat, während sie vom Englischen stark abweicht. Nur eines ist in der deutschen Entwicklung auffällig: Das häufigste Zeichen bei Luther, der Zwischenraum (erste Spalte), wurde vom Buchstaben „E“ abgelöst; das zeigt die Tendenz zu längeren Wörtern auf. Im Schnitt enthält heute jedes Wort mehr als ein „E“. Bei Shakespeare ist der Zwischenraum viel häufiger als in den deutschen Beispielen, er benutzt also kürzere Wörter. Auffallend ist, daß E und N seltener vorkommen als im Deutschen, T, O und Y dagegen öfter.

Autor	_	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Luther	180	59	17	21	54	142	9	23	46	56	3	8	22	24	86	22	5	0	58	57	41	40	6	11	0	1	8
Grimmelshausen	157	49	13	25	33	155	8	26	43	58	2	10	36	20	104	28	6	0	59	55	45	33	7	16	1	0	12
Schiller	142	48	15	28	43	158	11	32	42	65	3	9	29	18	80	24	9	0	65	57	48	39	5	12	0	0	11
Th. Mann	142	47	13	18	53	167	11	20	33	60	3	12	37	22	108	19	10	0	56	53	48	49	7	8	0	0	6
Shakespeare	198	60	12	16	32	94	17	14	53	53	1	7	36	25	50	69	11	1	47	52	75	27	8	20	1	21	0

(Zeichenhäufigkeiten bezogen auf 1000) *Bild 5*

Um zu zeigen, daß die Buchstabenhäufigkeiten für eine Sprache charakteristisch sind, wurde folgender Versuch gemacht. Jeweils für Deutsch und Englisch wurden zufällige Folgen mit den 27 Zeichen erzeugt. Dabei wurden zuerst Einzelzeichen nach den richtigen Wahrscheinlichkeiten erzeugt, dann auch Paare und Dreiergruppen mit den Häufigkeiten, mit der sie in Texten auftraten. Das Ergebnis sieht so aus:

Bild 6

	Deutsch
Einzel:	son vohaksien m ssoendet aubersallindesue ere d hasenn wossis ichrinen
Paar:	en meer scht sons aben heiden angen gotten machstern istum meind an
Tripel:	ltete moecht auf und wand rei den dass ist meeler muedeters zeder
	Englisch
Einzel:	thatomard t msuly oner fean ize ipr homuthere hondin rworon bexu do
Paar:	me bet tooks dood sts gain me in the buse haved bey sompearectim swe
Tripel:	fell bearlots ble was in i cannot and cause is more thereignity of

(Zufällige Texte mit der richtigen Häufigkeit für Einzelzeichen, Paare und Tripel)

Man sieht, daß – zumindest bei den Paaren und Dreiergruppen – die zufälligen Texte eindeutig der richtigen Sprache zugeordnet werden können.

Verschlüsselung

In der Süddeutschen Zeitung vom 10. Oktober 1996 wurde berichtet, daß der Briefbombenattentäter in Österreich seine Texte mit einem Code chiffriert hatte, der auf einer 243stelligen Zahl beruht, die das Produkt zweier unbekannter Primzahlen ist. Aus diesem aktuellen Anlaß wird hier auf dieses Verfahren eingegangen.

Das einfachste Verfahren einen Text zu chiffrieren besteht darin, eine Schlüsseltabelle zu erstellen, die jedem Zeichen umkehrbar eindeutig ein anderes zuordnet. Der Absender setzt seine Botschaft Zeichen für Zeichen um, der Empfänger benutzt die gleichen Schlüssel in umgekehrter Richtung, um den Ursprungstext zu rekonstruieren. Der Schlüssel darf nur den Partnern bekannt sein, jeder der den Schlüssel hat und die Botschaft abfängt, könnte sie sonst entschlüsseln. Dieses primitive Verfahren wäre aber auch ohne Kenntnis des Schlüssels zu knacken, denn aus dem letzten Abschnitt wissen wir, daß die Zeichenhäufigkeit charakteristisch ist für eine Sprache. Wird ein langer deutscher Text auf diese Art verschlüsselt, so

braucht man nur die Häufigkeiten der empfangenen Zeichen zu zählen. Das häufigste Zeichen wird vermutlich dem „E" entsprechen, auf die gleiche Art sucht man die Kandidaten für die anderen Zeichen und rekonstruiert so den Schlüssel. Sicherlich muß man dazu ein wenig probieren, doch ist die Anzahl der Möglichkeiten so eingeschränkt, daß diese Art der Chiffrierung keine Sicherheit bietet. In der Novelle „The Gold-Bug" beschreibt Edgar Allan Poe diese Methode sehr anschaulich (*Poe*, 1965).

Sicherer ist folgendes Verfahren. Jedem Zeichen wird eine Ziffernfolge zugeordnet. Diese Zuordnung braucht nicht geheim zu sein, beispielsweise A = 01, B = 02, ... Z = 26. Der Text wird Zeichen für Zeichen in Ziffern umgesetzt, das Ganze als eine vielstellige Zahl interpretiert. Zu dieser Zahl addiert man eine geheime Zahl, den Schlüssel, mit mindestens genausovielen Stellen wie die Nachricht. Die Summe ist die chiffrierte Nachricht, der Empfänger subtrahiert die geheime Zahl und wandelt die Differenz in den Text zurück. Die Sicherheit ist gewährleistet, solange der Schlüssel nicht in fremde Hände gerät. Darüber hinaus darf der gleiche Schlüssel nicht mehrmals benutzt werden, weil sonst auch wieder Häufigkeitsanalysen den Code brechen könnten. Eine andere Gefährdung besteht darin, daß der Schlüssel zu berechnen ist, wenn jemand irgendeinen Originaltext und seine Chiffrierung kennt.

Die Beispiele zeigen, daß der Schlüssel eine Schwachstelle bei der Chiffrierung ist. Er muß zuverlässig an den Empfänger übermittelt und von allen Partnern sorgfältig aufbewahrt und vor fremden Blicken geschützt werden. Dieses Grundproblem lösten 1978 die drei Wissenschaftler Rivest, Shamir und Adleman vom Massachusetts Institute of Technology mit einer revolutionären Idee: Man benutze einen Schlüssel, den jeder wissen darf, der aber die Eigenschaft hat, daß sich aus der Kenntnis der Verschlüsselung nicht unmittelbar die Entschlüsselung ableiten läßt. Man spricht in einem solchen Fall von einem öffentlichen Schlüssel (public key) und von einer Falltürfunktion, denn sie ist nur in einer Richtung leicht zu passieren (*Riesel*, 1985, 223ff.).

Nach den Anfangsbuchstaben ihrer Namen wird die Methode von Rivest, Shamir und Adleman RSA-Verfahren genannt. Zur Konstruktion eines öffentlichen Schlüssels verwendet es zahlentheoretische Erkenntnisse über Primzahlen. Primzahlen sind natürliche Zahlen, die nur durch Eins und sich selbst ohne Rest teilbar sind. Die ersten Primzahlen sind 2, 3, 5, 7, 11, 13, 17, 19 . Schon Euklid hat bewiesen, daß es unendlich viele Primzahlen gibt (*Euklid*, 1973, 204ff.). Ihre unregelmäßige Verteilung hat Mathematiker stets fasziniert, die Beschäftigung mit ihnen hat zu vielen fruchtbaren Ergebnissen geführt. Die Zerlegung einer sehr großen Zahl in Primfaktoren

ist langwierig, der Nachweis aber, ob eine Zahl Primzahl ist oder zusammengesetzt, kann leicht erbracht werden aufgrund einer Erkenntnis des französischen Mathematikers Pierre de Fermat (1601-1665) (*Koblitz*, 1987, 19).

Kleiner Fermatscher Satz:

Ist P eine Primzahl und A eine ganze Zahl, die nicht durch P teilbar ist, dann gilt

$A^{P-1} = 1$(modulo P).

Die Schreibweise X=Y(modulo P) bedeutet, daß X und Y beim Teilen durch P den gleichen Rest haben.

Beispiel: P = 5 , A = 3, $A^{P-1} = 3^4 = 81$. 81 durch P=5 geteilt hat den Rest 1.

Die Potenzbildung (modulo P) läßt sich auch bei sehr großen Zahlen schnell durchführen, denn man kann nach allen Rechenschritten sofort durch P teilen und mit dem Rest weiterrechnen; auf diese Art bleiben die Zahlen durch P beschränkt. Zur Prüfung, ob P eine Primzahl ist, berechnet man A^{P-1} (modulo P) für verschiedene Werte von A. Bei zusammengesetztem P, findet man gewöhnlich schnell einen Wert A, der nicht Eins liefert.

Wir nehmen an, daß die Nachricht - wie im vorherigen Beispiel - schon als Zahl A vorliegt. Beim RSA-Verfahren erzeugt sich der Empfänger zwei große Primzahlen P und Q und berechnet das Produkt M = P*Q. P und Q hält er geheim. M und eine Zahl C veröffentlicht er und teilt mit, wer ihm eine geheime Nachricht A zukommen lassen wolle, der solle $X = A^C$ (modulo M) berechnen und ihm nur das X senden. Nach dem kleinen Fermatschen Satz gilt: $A^{P-1} = 1$ (modulo P) und $A^{Q-1} = 1$ (modulo Q). Wegen M = P*Q folgt daraus $A^{(P-1)*(Q-1)} = 1$ (modulo M) . Bezeichnet man (P-1)* (Q-1) = S , dann gilt $A^{S+1} = A$ (modulo M), man hat also mit S+1 eine Hochzahl, mit der sich jedes A modulo M reproduzieren läßt. Bestimmt man jetzt eine Zahl D, so daß C*D = k*S+1 gilt für eine geeignete ganze Zahl k , dann folgt
$X^D = (A^C)^D = A^{C*D} = A^{k*S+1} = A^{k*S} * A = A$ (modulo M).
Durch Potenzieren mit D wird also die ursprüngliche Zahl – und damit die Nachricht – aus der chiffrierten Zahl X zurückgewonnen. D läßt sich aber nur ermitteln, wenn man die Primfaktoren P und Q kennt. Wer den Code brechen will, müßte die Primfaktoren P und Q von M bestimmen. Die Sicherheit des Verfahrens beruht darauf, daß keine schnellen Verfahren bekannt sind, diese Zerlegung zu finden. Bild 7 soll einen Eindruck geben von der Größe der benutzten Zahlen. Die Faktorzerlegung solcher Zahlen erfordert Tausende von Rechenstunden auf schnellen Computern. Sollten die Computer noch schneller werden oder sollten effizientere Algorithmen zur Faktorisierung gefunden werden, dann ließe sich dieser Gewinn durch

die Wahl noch größerer Zahlen P und Q wieder zunichte machen, die Sicherheit des RSA-Verfahrens bliebe erhalten.

M = 881065616861466593904932969013305173990854392515664508767558304229165507346943876841474142022214660332631678449450507625349814801812355008867444336813145252525264038801488920850278532463638743304156278500956830090004910666093465122858208598461

P = 24968051039195866023353174745547832341665366443135466443225750099978329590968635896267478926272981726350346543280831685991

Q = 35287720914953826898740516863091818752722241957287265666773769460700119566390187573588098196314322704795780586884992634171

M = P * Q

(244stellige Zahl als Produkt zweier 122stelliger Primzahlen)

Bild 7

Literatur

EUKLID: Die Elemente. Nach Heibergs Text aus dem Griechischen übersetzt und herausgegeben von Clemens Thaer, Darmstadt 1973.

JAGLOM, A .M. und JAGLOM I. M. (1956): Wahrscheinlichkeit und Information. Frankfurt 1984.

KOBLITZ, Neal (1987): A Course in Number Theory and Cryptography. New York 1988.

POE, Edgar Allan: The Complete Tales and Poems of Edgar Allan Poe. London 1965.

RIESEL, Hans (1985): Prime Numbers and Computer Methods for Factorization. Boston.

SCHRÖDER, Rudolf Alexander: Homers Odyssee. Frankfurt.

WEIZSÄCKER, Carl Friedrich von (1971): Wissenschaft, Sprache und Methode, aus: Die Einheit der Natur, München 1995.