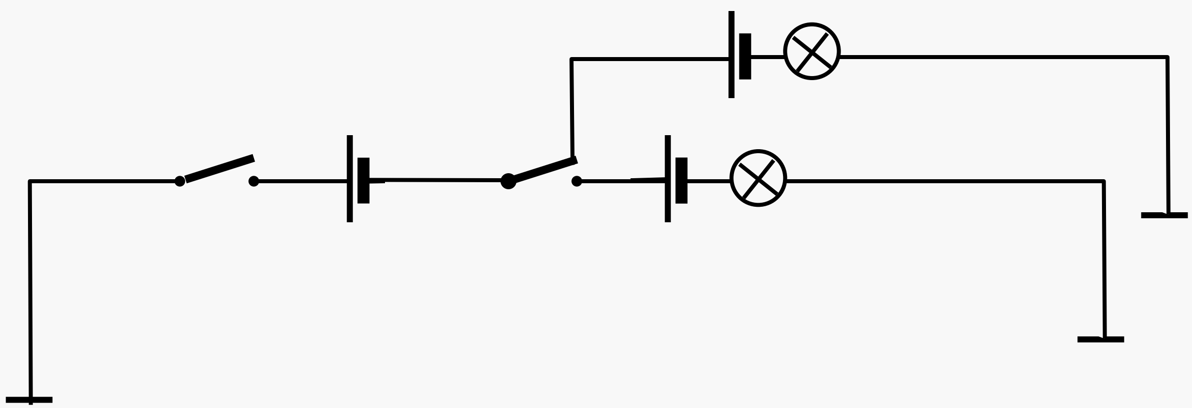
Signale und Logik (2)

Vereinfachte (Relais-/)Schalterdarstellung:

Trennung von Nutz- u. Schaltsignal macht Adressaten „anwählbar“:

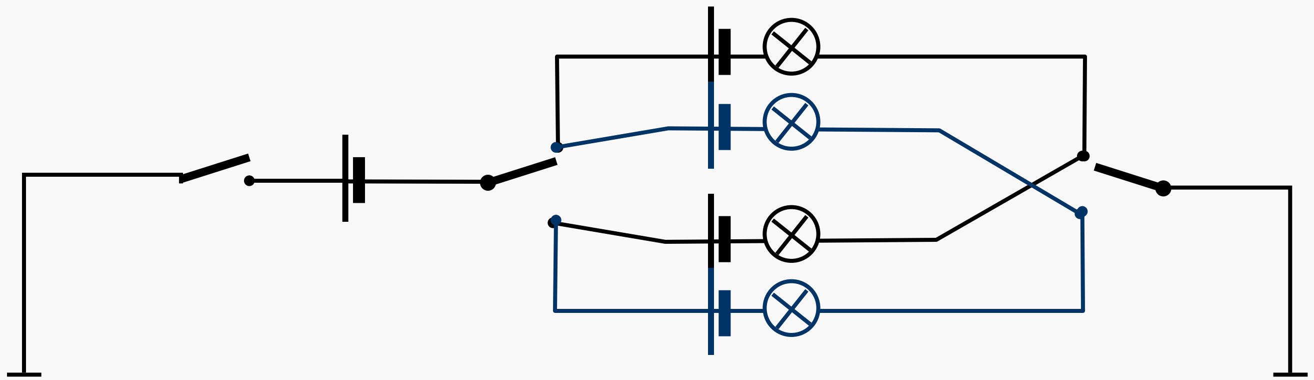
Bei zwei möglichen Empfängern genügt ein Schalter, um einen der beiden zu wählen.



(Hier: Umwidmung des Verstärker-Relais für die Adressierung - in Wirklichkeit benötigen Verstärkung u. Zuschaltung getrennte Relais!)

Signale und Logik (2)

Zur getrennten Ansteuerung von 4 Empfängern werden 2 Schalter notwendig: (... und weiter?)



Benötigt werden Aussagen darüber, wie eine gegebene Anzahl unterscheidbarer Zustände (hier: Empfängerwahl) mit möglichst wenigen binären Elementen (hier: Schaltern) eindeutig darzustellen -zu „codieren"- ist.

Das ist Gegenstand der Informationstheorie.

Die Informationstheorie untersucht **Darstellung**, **Speicherung** und **Übertragung** von Information.

Anmerkungen:

- Im Vergleich zur (später entstandenen) Informatik überwiegen hier formal-theoretische Aspekte wie Wesen, Erhaltung oder Wiedergewinnung von Information.

- Die Informationstheorie bedient sich meist der Methoden der Wahrscheinlichkeitsrechnung und der mathematischen Statistik.

- Information benötigt zu ihrer Darstellung **Symbole** (=unterscheidbare physikalische Zustände).

  Bsp.: ' * ' (auf Schreibmaschinen- oder Telefontastatur)

  auch: Gesten, Körperhaltungen, Flaggenstellungen

- Ein oder mehrere Symbole werden als **Zeichen** mit eigener Bedeutung definiert (= vereinbart).

  Bsp.: ' • —– ' (‘a‘ bei Morse) oder ' **;** ‘

  auch: Schrift-, Licht-, Hand-, Vogelzeichen

- Ein geschlossen verwendeter, geordneter, verbreiteter Vorrat von mindestens zwei Zeichen bildet ein **Alphabet**.

  Bsp.: Griechisches Alphabet, lateinisches ABC,

  American **S**tandard **C**ode for **I**nformation **I**nterchange (ASCII)

Anmerkungen:

- Es gibt Zeichen, die sich aus Symbolen zusammensetzen: ‘**?**‘ und ‘**!**‘ bestehen z.B. aus je 2 Symbolen, von denen das jeweils obere allein bedeutungslos ist.
- Flaggenstellungen können Nachrichten durch Flaggenalphabet aber auch Geisteshaltungen u.ä. übermitteln („Halbmast“).
- Die Frage, ob ein Alphabet auch geordnet sein muß, wird in der Literatur widersprüchlich behandelt: Gehören die ungeordneten Satzzeichen (, . ;) zum ABC?
- Zur Bezeichnung 'Alphabet‘ gehört eine gewisse Standardisierung, Verbreitung oder Bekanntheit; insofern kann Code auch die Entsprechungen zwischen willkürlich ausgesuchten Zeichenmengen betreffen (vgl. Verschlüsselung).
- Am Morse-Bsp. wird deutlich, wie wichtig die Code-Wahl ist: ‘E‘ ist im Englischen -wie im Dt.- der häufigste Buchstabe, deshalb im Morse-Alphabet der kürzeste.

Die eindeutige Zuordnung zwischen Alphabeten wird **Code**, ihr Einsatz **Codierung** genannt.

Anmerkungen:

- Codierungen sind nicht immer umkehrbar eindeutig!

  Bsp.: Der Wechsel zwischen dt. Klein- u. Großbuchstaben:

  'Gießen' $\Rightarrow$ 'GIESSEN'

  Bsp.: Der Morse-Code ist umkehrbar eindeutig, weil jedes Zeichen mit einem Leerzeichen abgeschlossen wird:

  'Eis?' $\Leftrightarrow$ '• •• ••• ••– –••'

- Auch die Umsetzung abstrakter Information (Gedanken) in eine Sprache erfolgt durch Codierung; Programmier-sprachen sind Gegenstand (u.a.) der Praktischen Informatik, menschliche Sprachen (u.a.) der Linguistik (vgl. Soziolinguistik: restringierter, elaborierter Code) ¬

- Häufig verwendete oder typische Zeichenfolgen (Zeichenketten) werden oft auch als neues Zeichen aufgefaßt. Das daraus hervorgehende neue Zeichen wird auch **Codewort** (auch Wort oder Superzeichen) genannt.

  Bsp.: '**u**', '**n**', '**d**', '**&**' sind Zeichen (keine Wörter); '**und**' ist ein Zeichen oder (Code-) Wort.

- Innerhalb eines Codewortes dienen die Zeichen, die es bilden, als **Elemente** (bei Zahlen auch: **Stellen**).

  Bsp.: Das (Zeichen oder) Wort '**BARBARA**' enthält 7 Elemente, aber nur 3 (Zeichen oder) Symbole. ¬

- Innerhalb eines Codes kann die Anzahl der Zeichen (Symbole) in einem Codewort variieren oder für alle Wörter gleich sein.

  Bsp.: Im Morse-Code sind Zeichen ungleich lang, im ASCII-Code gleich lang:

Schriftzeichen	Morse	ASCII
e	•	`0110 0101`
?	• • – – • •	`0011 1111`

- Codes mit einem Zeichenvorrat von nur zwei Zeichen (z.B. 0 und 1) werden **Binärcodes** genannt. Sie sind technisch interessant, weil sie auch mit einem Schalter (Transistor) realisiert werden können.

  Der Morse-Code ist als Binärcode nicht umkehrbar eindeutig!

  (Daher: Einführung des Zwischenraums als drittes Zeichen) ¬

Grundzüge der Informationstheorie

- Codes und Codierung bilden einen wichtigen Teil der Informationstheorie; denn sie sind eine grundlegende Voraussetzung für die maschinelle DV.

- **Optimal** werden Codes genannt, die je Zeichen möglichst viel Information verschlüsseln (=codieren). Dazu bedarf es einer Metrik für Information.

Beispiel:

Zwei Skatspieler sprechen sich ab: Bevor einer eine Karte spielt, kann er vom anderen ein Zeichen bekommen:

- Bierglas oben / unten anfassen:
  besonders hohe / niedrige Karte spielen!
- Aus dem Bierglas trinken / nicht trinken:
  Partner selbst kann / kann nicht zur Lage beitragen.

- Je „selbstverständlicher“ (wahrscheinlicher) ein Zeichen für die Spielsituation ist, umso weniger interessant ist es für den anderen (Information is data which is used in decision-making).
- Die Zeichen des Spielpartners „summieren sich“ zur Gesamt-Kommunikation.

- Der **Informationsgehalt** *h* einer Nachricht bzw. eines Zeichens *x* hängt von der Wahrscheinlichkeit *p(x)* seines Auftretens ab: Je unwahrscheinlicher das Auftreten eines Zeichens *x* ist, desto höher ist sein Informationsgehalt:

$$h(x) = \quad f\,(\,1/p(x)\,)$$

(f: vorerst unbekannte Funktion)

- Für den Informationsgehalt des Empfangs mehrerer voneinander unabhängiger Zeichen folgt daraus:
  - Es muß gelten (Wahrscheinl/keit gleichzeitigen Auftretens):

$$h(xy...) = f\,(\,p(x)^{-1} \cdot p(y)^{-1} \cdot \,...)$$

  - Anschauungshalber sollte außerdem gelten:

$$h(xy...) = h(x) + h(y) + \,... \qquad \text{d.h.:}$$

$$f\,(\,p(x)^{-1} \cdot p(y)^{-1} \cdot \,...) = f\,(\,p(x)^{-1}\,) + f\,(\,p(y)^{-1}\,) + \,...$$

Zusammenfassung:
Forderungen an den Informationsgehalt *h* der Nachrichten *x*, *y*...:

$$h(xy...) = f\,(p(x)^{-1} \cdot p(y)^{-1} \cdot ...) = f\,(p(x)^{-1}) + f\,(p(y)^{-1}) + ... = h(x) + h(y) + ...$$

frei assoziiert:

$$\log\,(x \cdot y \cdot ...) = \log x + \log y + ...$$

Definition (C. Shannon):

Der **Informationsgehalt** *h* eines Zeichens *x* ist definiert als der Logarithmus dualis des Reziprokwertes der Wahrscheinlichkeit, mit der das Zeichen auftritt:

$$\begin{aligned} h(x) &= \mathrm{ld}\,[\,1/p(x)\,] \\ &= -\,\mathrm{ld}\,[\,p(x)\,] \end{aligned}$$

Kurzes Logarithmen-Repetitorium

Definition: $\log_B Z = L \quad \Leftrightarrow \quad B^{\log_B Z} \equiv B^L = Z$

Speziell:

$\lg Z \equiv \log_{10} Z \qquad \ln Z \equiv \log_e Z \qquad \mathrm{ld}\, Z \equiv \log_2 Z$

$\text{(Def.)} \Rightarrow Z_1 \cdot Z_2 = B^{L1} \cdot B^{L2} = B^{L1+L2}$

Interessante Regeln:

$\log(Z_1 \cdot Z_2) = \log(Z_1) + \log(Z_2)$	$\log(Z^E) = E \cdot \log Z$
$\log(Z_1 / Z_2) = \log(Z_1) - \log(Z_2)$	$\log(\sqrt[W]{Z}) = (\log Z) / W$

Kurzes Logarithmen-Repetitorium

Wechsel der Logarithmenbasis:

$$Z = B_1^{\log_{B1} Z}$$

$$\log_{B2} Z = \log_{B2}(B_1^{\log_{B1} Z})$$

$$= \log_{B1} Z \cdot \log_{B2} B_1$$

$$\boxed{\log_{B1} Z \;=\; \log_{B2} Z \,/\, \log_{B2} B_1}$$

Oft benötigte Umrechnungen:

$$\lg 10 = \ln e = \mathrm{ld}\, 2 = 1$$

$$\lg 1 = \ln 1 = \mathrm{ld}\, 1 = 0$$

$$\mathrm{ld}\,(1/x) = -\mathrm{ld}\, x \quad ; \quad \lg\,(1/x) = -\lg x$$

$$\lg\,(10^x) = \ln\,(e^x) = \mathrm{ld}\,(2^x) = x$$

$$10^{\lg x} = e^{\ln x} = 2^{\mathrm{ld}\, x} = x$$

$$\mathrm{ld}\, x = \lg x \,/\, \lg 2 = \ln x \,/\, \ln 2$$

Aus der Definition des Informationsgehalts

$$h(x) = -\mathrm{ld}\,[\,p(x)\,]$$

folgt:

Der Informationsgehalt einer aus mehreren (voneinander unabhängigen) Zeichen bestehenden **Sequenz** ist gleich der **Summe** der Informationsgehalte der einzelnen Zeichen:

$$h(xy\ldots) = -\mathrm{ld}\,[\,p(x) \cdot p(y) \cdot \ldots]$$

$$= -\mathrm{ld}\,[\,p(x)] - \mathrm{ld}\,[\,p(y)\,] - \ldots$$

$$= h(x) + h(y) + \ldots$$

ursprüngliche Forderung ¬

Was bedeutet: „Informationsgehalt von *x* ist eins“?

$$h(x) = -\mathrm{ld}\,[\,p(x)\,] = \mathrm{ld}\,[\,1/p(x)\,] = 1,$$

woraus folgt:

$$1/p(x) = 2^1 \text{ bzw. } p(x) = ½\,,$$

d.h.:

- Der Informationsgehalt der Antwort auf eine Frage, die nur zwei (gleich wahrscheinliche) Möglichkeiten zuläßt, ist die Einheit des Informationsgehalts; sie wird **bit** genannt (**b**asic **i**ndissoluble information uni**t**).

- Ein bit ist der Informationsgehalt eines Zeichens in einem binären Alphabet mit gleicher Auftretungswahrscheinlichkeit.

  Das binäre Alphabet kann aus den Wertepaaren bestehen: ja/nein; wahr/falsch; schwarz/weiß ; hell/dunkel etc. ¬

- Haben alle *n* Zeichen eines Zeichenvorrats die gleiche Auftretungswahrscheinlichkeit $p=1/n$, so beträgt d. Informationsgehalt jedes einzelnen Zeichens: $h = \mathrm{ld}\,[\,1/p\,] = \mathrm{ld}\,n = \lg n / \lg 2$

  Beispiel:

  Informationsgehalt jedes der 10 Dezimalzeichen:

  $h = \mathrm{ld}\,10 = \lg 10 / \lg 2 = 3{,}322$ bit

- Bei Zeichenvorräten mit unterschiedlich wahrscheinlichen Zeichen bezeichnet man als **mittleren Informationsgehalt** *H* die Summe der mit den individuellen Auftretungswahrscheinlichkeiten p_i ($\Sigma p_i=1$; $i=1,\ldots,n$) gewichteten Informationsgehalte h_i der einzelnen Zeichen :

  $$H = \Sigma\,(p_i \cdot h_i) = \Sigma\,(p_i \cdot \mathrm{ld}\,[1/p_i]) = -\,\Sigma\,(\,p_i \cdot \mathrm{ld}\,p_i\,)$$

  Beispiel:

  Untersuchungen ergeben für d. dt. Sprache $H \approx 4{,}15$ bit/Zeichen; Berücksichtigung typischer Kombinationen („qu“, „ung“ etc.) reduziert das auf $H \approx 1{,}6$ bit/Zeichen

Beispiel: Häufigkeit von Buchstaben im Deutschen

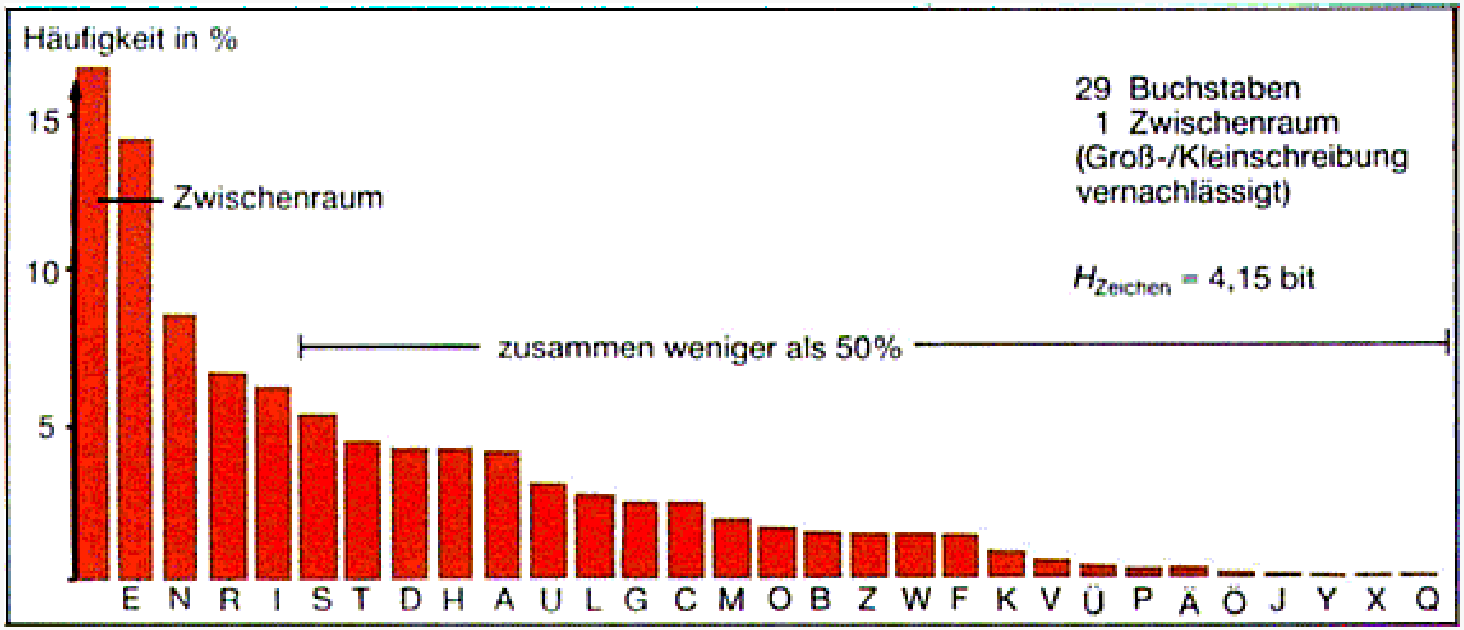


Bild: H.Breuer: „dtv-Atlas zur Informatik“, dtv 1995

Ist die Wahrscheinlichkeit für das Auftreten des Buchstabens E: p(E)=0,147 und die für H: p(H)=0,045, so ist der Informationsgehalt der Zeichenkette 'EHE' [1]:

$$-\mathrm{ld}\,(0{,}147 * 0{,}045 * 0{,}147)\ \text{bit} \approx [\,-\lg(0{,}000972405) / \lg 2\,]\ \text{bit}$$
$$\approx [\,3{,}012 / 0{,}30\,]\ \text{bit} \approx 10{,}006\ \text{bit}$$

1) (Zeichenkette gebildet mit zufällig aus einem Text herausgegriffenen Zeichen)

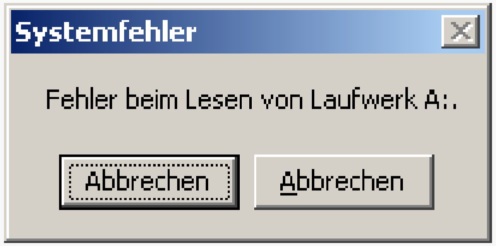
Grundzüge der Informationstheorie

- Die Einheit zur binären Darstellung von Daten heißt **Bit** (**B**inary dig**it**). Ihr Inhalt wird meist mit 0 bzw. 1 codiert.

- Mit *n* Bit lassen sich 2^n Zustände darstellen.
- Es gibt nur ganzzahlige Bit – im Gegensatz zu bit.
- Zur Darstellung von *n* bit benötigt eine elektronische Rechenanlage mindestens *n* Bit.
- Informationsgehalt wird in bit **berechnet**: Es gibt keine Instrumente zu seiner Messung.

Die technische Darstellung erfolgt u.a. mit Hilfe von:

- Ladung
  - 0 = ungeladen
  - 1 = geladen
- Spannung
  - 0 = 0 Volt
  - 1 = ca. 6 Volt
- Magnetisierung
  - 0 = gleichbleibende Magnetisierung
  - 1 = Magnetisierungswechsel
- Licht
  - 0 = kein Licht
  - 1 = Licht
- ...



Grundzüge der Informationstheorie

- Aus bestimmten technischen Gründen wie:
  - Geschwindigkeit von Lese- und Schreiboperationen
  - Darstellungsmöglichkeit „häufiger" Zeichen (z.B. Alphabet)
  - Darstellungsmöglichkeiten von Zahlen, etc.

  werden in der Informatik oft Vielfache von 8-Bit-Gruppen verwendet (8Bit, 16Bit, ...)

  Eine 8-Bit-Sequenz heißt ein **Byte.**
- Bestimmte 2er-Potenzen werden in der Informatik häufig als Maßzahlen (z.B. für Speichergrößen) verwendet:
  - 1 KByte = 2^{10} Byte = 1024 Byte (1 Kilobyte)
  - 1 Mbyte = $2^{10} \cdot 2^{10}$ Byte (1 Megabyte)
  - 1 Gbyte = $2^{10} \cdot 2^{10} \cdot 2^{10}$ Byte (1 Gigabyte)
  - 1 Tbyte = $2^{10} \cdot 2^{10} \cdot 2^{10} \cdot 2^{10}$ Byte (1 Terabyte)

- Eine Zeichenfolge zur Codierung eines bestimmten Informationsgehalts wird als **Wort** bezeichnet, die Anzahl der darin enthaltenen Zeichen als **Wortlänge**.

- Mit einer Wortlänge von **L** Zeichen, von denen jedes **B** unterscheidbare Zustände annehmen kann, lassen sich insg. **N** unterschiedliche Zustände codieren, mit

$$\mathbf{N = B^{L}} \quad (B > 1\ !)$$

Die Binärcodierung von ***n*** Zuständen benötigt eine Wortlänge von ***l*** Bit:

$$2^l \geq n \quad \Rightarrow \quad ld(2^l) = l * ld\ 2 \geq ld\ n \quad \Rightarrow \quad \mathbf{l \geq ld\ n} \qquad (ld\ 2 = 1)$$

Beispiel:

Codierung der Himmelsrichtungen (N,O,S,W – d.h.: n=4)

$I \geq ld\ 4 = ld\ (2^2) = 2 * ld2 \quad \Rightarrow \quad I \geq 2$ Bit

- Eine mögliche Codierung ist:

  00 = Norden

  01 = Osten

  10 = Süden

  11 = Westen

- Obige Codierung läßt sich in zwei Fragen umsetzen:
  - Weht der Wind aus S bzw. W (ja / nein) ?
  - Weht der Wind aus O bzw. W (ja / nein) ?

- Nimmt man noch die Zwischenrichtungen NO, SO, SW und NW hinzu, so können die 8 Zustände mit 3 Bit codiert werden

- **Redundanz** ist ein Maß für die Nicht-Nutzung der Möglichkeiten eines Codes bzw. Maß für den Anteil einer Nachricht, der keine Information enthält.

- Im Falle gleich wahrscheinlicher, gleich lang codierter Wörter (*l* binäre Zeichen [Bit]) ist die Redundanz *r* definiert als Differenz zwischen der Wortlänge und Informationsgehalt *h* [bit]:

$$r = l - h$$

Redundanz wird oft genutzt zur Erstellung prüfbarer und korrigierbarer Codierungen.

Ein Beispiel:

Die Binärcodierung eines Tripels (3 Zustände, 3 Spieler,...) benötigt eine Wortlänge von l=2 Bit; denn es ist gefordert:

$$2^l \geq 3 \;\Rightarrow\; \mathrm{ld}(2^l) \geq \mathrm{ld}\ 3 \Rightarrow l \cdot \mathrm{ld}2 \geq \mathrm{ld}\ 3$$

(ld2=1)

$$\Rightarrow \quad l \geq \mathrm{ld}\ 3 = \lg 3 / \lg 2 = 0{,}477/0{,}301 = 1{,}585 \text{ Bit}$$

d.h., die kleinste ganze Zahl Bit, die ausreicht, ist 2.

Diese Codierung hat damit eine Redundanz von:

$$r = l - h = 2 - 1{,}585 = 0{,}415 \text{ bit}$$

(i.d.R.: r < 1 !?) ¬

Codierungsstandards

- **ASCII** (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange):
  - 7 Bit / Zeichen;
  - erweiterte Version mit 8 Bit / Zeichen;
  - Breiteste Verwendung (Unix, MS-DOS, Programmiersprachen, ...)
- **ANSI**-Code (**A**merican **N**ational **S**tandards **I**nstitute):
  - 8 Bit (=1 Byte) / Zeichen;
  - Positionen 32-127 wie bei ASCII (Buchstaben, Ziffern, Satzzeichen);
  - Verwendung in Windows95 ff. *
- **Unicode:**
  - 16 Bit (=2 Byte) / Zeichen;
  - Erweiterung des ASCII-Codes: Buchstaben und Symbole aus allen bekannten geschriebenen Sprachen der Welt
  - Verwendung ab Windows NT ¬

Codierungsstandards

American Standard Code for Information Interchange

@	**NUL**	**000**	T	DC4	**020**	(	**040**	<	**060**	P	**080**	d	**100**	x	**120**
A	SOH	**001**	U	NAK	**021**	)	**041**	=	**061**	Q	**081**	e	**101**	y	**121**
B	STX	**002**	V	SYN	**022**	*	**042**	>	**062**	R	**082**	f	**102**	z	**122**
C	ETX	**003**	W	ETB	**023**	+	**043**	?	**063**	S	**083**	g	**103**	{	**123**
D	EOT	**004**	X	CAN	**024**	,	**044**	@	**064**	T	**084**	h	**104**	_|	**124**
E	ENQ	**005**	Y	EM	**025**	-	**045**	A	**065**	U	**085**	i	**105**	}	**125**
F	ACK	**006**	Z	SUB	**026**	.	**046**	B	**066**	V	**086**	j	**106**	~	**126**
G	BEL	**007**	[	ESC	**027**	/	**047**	C	**067**	W	**087**	k	**107**	DEL	**127**
H	BS	**008**	\	FS	**028**	0	**048**	D	**068**	X	**088**	l	**108**		
I	HT	**009**	]	GS	**029**	1	**049**	E	**069**	Y	**089**	m	**109**		
J	LF	**010**	^	RS	**030**	2	**050**	F	**070**	Z	**090**	n	**110**		
K	VT	**011**	_	**US**	**031**	3	**051**	G	**071**	[	**091**	o	**111**		
L	FF	**012**	SP		**032**	4	**052**	H	**072**	\	**092**	p	**112**		
M	CR	**013**	!		**033**	5	**053**	I	**073**	]	**093**	q	**113**		
N	SO	**014**	"		**034**	6	**054**	J	**074**	^	**094**	r	**114**		
O	SI	**015**	#		**035**	7	**055**	K	**075**	_	**095**	s	**115**		
P	DLE	**016**	$		**036**	8	**056**	L	**076**	`	**096**	t	**116**		
Q	DC1	**017**	%		**037**	9	**057**	M	**077**	a	**097**	u	**117**		
R	DC2	**018**	&		**038**	:	**058**	N	**078**	b	**098**	v	**118**		
S	DC3	**019**	'		**039**	;	**059**	O	**079**	c	**099**	w	**119**		

Codierungsstandards

Bedeutung der Sonderzeichen im ASCII-Code:

@	NUL	Null, or all zeros	R	DC2	DeviceControl2
A	SOH	StartHeading	S	DC3	DeviceControl3(XOFF)
B	STX	StartText	T	DC4	DeviceControl4
C	ETX	EndText	U	NAK	Neg.Acknowledge
D	EOT	EndTransmission	V	SYM	SynchronousIdle
E	ENQ	Enquiry	W	ETB	EndTrans.Block
F	ACK	Acknowledge	X	CAN	Cancel
G	BEL	Bell	Y	EM	EndofMedium
H	BS	Backspace	Z	SUB	Substitute
I	HT	HorizontalTab	[	ESC	Escape
J	LF	LineFeed	\	FS	FileSeparator
K	VT	VerticalTab	]	GS	GroupSeparator
L	FF	FormFeed	^	RS	RecordSeparator
M	CR	CarriageReturn	_	US	UnitSeparator
N	SO	ShiftOut		SP	Space
O	SI	ShiftIn			
P	DLE	DataLinkEscape	?	DEL	Delete
Q	DC1	DeviceControl1(XON)			

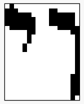
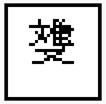
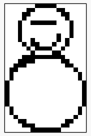
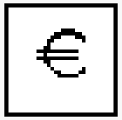
Codierungsstandards

Unicode:

- 1991: Version 1.0
- Sept. 2013: Version 6.3 (110.187 Zeichen)
- 2 Byte/Zeichen: Buchstaben und Symbole aus allen bekannten geschriebenen Sprachen der Welt
  - Amerika, Europa, Mittlerer Osten, Afrika, Indien, Asien, Pazifik
  - Symbole
  - Satzzeichen
  - Sonderzeichen
  - Währungszeichen …
- Unterstützt max. 1.114.112 Zeichen
- Genormt in ISO/IEC 10646

Codierungsstandards

Unicode:

- 05F1 
- FA0E 
- 2603 
- 20AC 
- xxD0 - xxDF