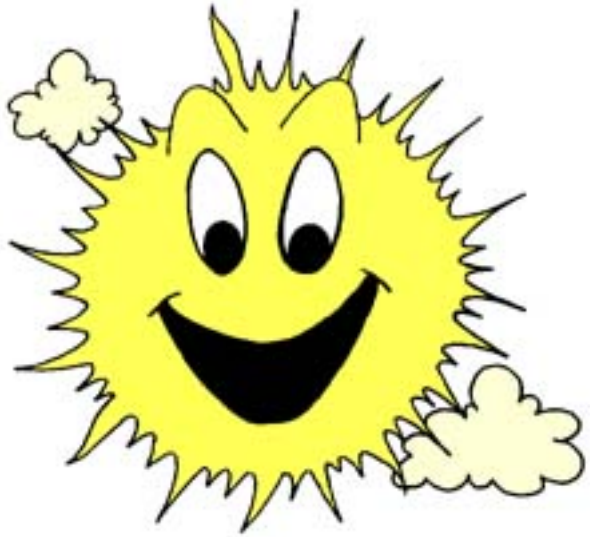
Wiederholung der 5. Vorlesung



- Nachricht und Information
  - ASCII-Code
  - UNICODE
- Information & Entscheidungsgehalt

3.5 Unicode

- ASCII unterstützt nur lateinischen Zeichensatz
- Was ist mit Arabisch, Hebräisch, Russisch, Chinesisch, Japanisch, ... ?
- Unicode enthält 16 Bit. Damit können direkt 65536 Zeichen codiert werden. (Mit Control- und Escape-Modi entsprechend mehr)
- In Version 2.0 belegt er 38885 Codewörter, deckt die wichtigsten Sprachen der Welt ab.
- 0000-007F identisch mit ASCII
- Standard im Fluss: www.unicode.org

3.5 Unicode (Auswahl 0000-00FF)

0000	NUL	0020	SP	0040	@	0060	`	0080	Ctrl	00A0	NBS	00C0	À	00E0	à
0001	SOH	0021	!	0041	A	0061	a	0081	Ctrl	00A1	¡	00C1	Á	00E1	á
0002	STX	0022	"	0042	B	0062	b	0082	Ctrl	00A2	¢	00C2	Â	00E2	â
0003	ETX	0023	#	0043	C	0063	c	0083	Ctrl	00A3	£	00C3	Ã	00E3	ã
0004	EOT	0024	$	0044	D	0064	d	0084	Ctrl	00A4	¤	00C4	Ä	00E4	ä
0005	ENQ	0025	%	0045	E	0065	e	0085	Ctrl	00A5	¥	00C5	Å	00E5	å
0006	ACK	0026	&	0046	F	0066	f	0086	Ctrl	00A6	¦	00C6	Æ	00E6	æ
0007	BEL	0027	'	0047	G	0067	g	0087	Ctrl	00A7	§	00C7	Ç	00E7	ç
0008	BS	0028	(	0048	H	0068	h	0088	Ctrl	00A8	¨	00C8	È	00E8	è
0009	HT	0029	)	0049	I	0069	i	0089	Ctrl	00A9	©	00C9	É	00E9	é
000A	LF	002A	*	004A	J	006A	j	008A	Ctrl	00AA	ª	00CA	Ê	00EA	ê
000B	VT	002B	+	004B	K	006B	k	008B	Ctrl	00AB	«	00CB	Ë	00EB	ë
000C	FF	002C	´	004C	L	006C	l	008C	Ctrl	00AC	¬	00CC	Ì	00EC	ì
000D	CR	002D	-	004D	M	006D	m	008D	Ctrl	00AD	–	00CD	Í	00ED	í
000E	SO	002E	.	004E	N	006E	n	008E	Ctrl	00AE	®	00CE	Î	00EE	î
000F	SI	002F	/	004F	O	006F	o	008F	Ctrl	00AF	–	00CF	Ï	00EF	ï
0010	DLE	0030	0	0050	P	0070	p	0090	Ctrl	00B0	°	00D0	Đ	00F0	ð
0011	DC1	0031	1	0051	Q	0071	q	0091	Ctrl	00B1	±	00D1	Ñ	00F1	ñ
0012	DC2	0032	2	0052	R	0072	r	0092	Ctrl	00B2	²	00D2	Ò	00F2	ò
0013	DC3	0033	3	0053	S	0073	s	0093	Ctrl	00B3	³	00D3	Ó	00F3	ó
0014	DC4	0034	4	0054	T	0074	t	0094	Ctrl	00B4	´	00D4	Ô	00F4	ô
0015	NAK	0035	5	0055	U	0075	u	0095	Ctrl	00B5	µ	00D5	Õ	00F5	õ
0016	SYN	0036	6	0056	V	0076	v	0096	Ctrl	00B6	¶	00D6	Ö	00F6	ö
0017	ETB	0037	7	0057	W	0077	w	0097	Ctrl	00B7	·	00D7	×	00F7	÷
0018	CAN	0038	8	0058	X	0078	x	0098	Ctrl	00B8	¸	00D8	Ø	00F8	ø
0019	EM	0039	9	0059	Y	0079	y	0099	Ctrl	00B9	¹	00D9	Ù	00F9	ù
001A	SUB	003A	:	005A	Z	007A	z	009A	Ctrl	00BA	º	00DA	Ú	00FA	ú
001B	ESC	003B	;	005B	[	007B	{	009B	Ctrl	00BB	»	00DB	Û	00FB	û
001C	FS	003C	<	005C	\	007C	|	009C	Ctrl	00BC	1/4	00DC	Ü	00FC	ü
001D	GS	003D	=	005D	]	007D	}	009D	Ctrl	00BD	1/2	00DD	Ý	00FD	ý
001E	RS	003E	>	005E	^	007E	~	009E	Ctrl	00BE	3/4	00DE	Þ	00FE	þ
001F	US	003F	?	005F	_	007F	DEL	009F	Ctrl	00BF	¿	00DF	ß	00FF	ÿ

Vor dem Empfang der Information gibt es eine Menge von Ereignissen, die eintreten könnten (jedes mit einer gewissen Wahrscheinlichkeit). Man weiß aber nicht, welches dieser Ereignisse eintreten wird. Erst **nach Empfang** der Information (=Eintreten eines Ereignisses) weiß man, welches Ereignis eingetreten ist. Damit ist die **Unsicherheit beseitigt**. Nicht beseitigt ist allerdings die Ungewissheit, welches Ereignis als nächstes eintreten wird.

Wenn man **ganz sicher** weiß, welches Ereignis eintreten wird, dann ist die Wahrscheinlichkeit **p=1**.

Beispiel:

Ein Wanderer kommt an eine Weggabelung, die nicht ausgeschildert ist. Hat er keine weiteren Informationen hat, welcher Weg der richtige ist, sind alle Möglichkeiten gleich wahrscheinlich.

Bei einer

- 2fachen Weggabelung also p=0.5
- 3fachen Weggabelung ist p=0.333
- 4fachen ist p=0.25.
- allgemein: m Möglichkeiten ➔ $p_i = 1/m$

$P(\text{gesamt}) = p_1 + p_2 + \ldots + p_n = \mathbf{1}$

4.1 Information

- Das Wesen der Information besteht u.a. darin, Unsicherheit zu beseitigen. Vor Eintreffen der Information sind viele Möglichkeiten wahrscheinlich. Nach Eintreffen der Information sind Unsicherheiten beseitigt.

Beispiel: Bei zwei Möglichkeiten: Ja oder Nein. und gleicher Wahrscheinlichkeit p = 0.5 ist keine Vorhersage möglich.

Sobald die Information, z.B. „Ja“, eingetroffen ist, $p_0=0$, $p_1=1$, ist die Unsicherheit beseitigt.

4.1 Informationswert

- Je unwahrscheinlicher die Information, also je geringer die Wahrscheinlichkeit einer richtigen Vorhersage und um so größer ist ihr Informationswert.
- Also ist der Wert (Gehalt) einer Information (z.B. einer Nachricht, die wir erhalten) um so größer, je weniger man sie erwartet, je seltener sie eintrifft.
- Beispiel:
  - „Heute ist Mittwoch“
  - „Sie haben 1 Mio € im Lotto gewonnen“

4.1 Information (wahrscheinlichkeitstheoretisch)

- Informationsübertragung durch einen Kanal
- Information **I** des Zeichens hängt von dessen Häufigkeit bzw. Wahrscheinlichkeit $\mathbf{p_i}$ ab.
- Zwei zu übertragende Zeichen seien unabhängig voneinander, die Gesamtinformation die Summe der Einzelinformation: $\mathbf{I = I(p_x) + I(p_y)}$
- Wahrscheinlichkeit für zwei Zeichen
  $\mathbf{p(x,y) = p_x \times p_y}$
- Informationsgehalt beider Zeichen ist
  $I = I(\ p(x,y)\ ) = I(p_x \times p_y)$
- aus $I(p_x) + I(p_y) = I(p_x \times p_y)$ folgt
  $\mathbf{I(p) = ld(1/p) = -ld(p)}$ ld: logarithmus dualis

4.2 Entscheidungsgehalt

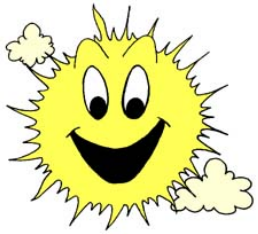
- Wie viele Entscheidungen (binäre Fragestellungen) sind notwendig, um eine bestimmte Nachricht aus einer Vielzahl von Nachrichten auszuwählen?
- Wenn Informationsquelle n Nachrichten liefern kann, dann sind mindestens ld(n) Fragen nötig um die Information einer Nachricht zu ermitteln.
- **Entscheidungsgehalt H_0** einer Nachricht: **$H_0(n) = ld(n)$** ; Maßeinheit ist das bit (binary digit)
- Beispiel:
  - n=2, $H_0(2) = ld(2) = 1$ bit
  - n=7, $H_0(7) = ld(7) = 2.81$ bit
- Entscheidungsgehalt = Speicherkapazität eines Systems

4.2 Beispiele

- **Informationsfluss** = Information / Zeiteinheit gemessen in **bit / s**
- **Mensch** kann einen Informationsfluss von **10-50 bit/s** bewusst verarbeiten.

- Ein **Fernsehbild** in PAL-TV mit 720 x 576 Bildpunkten und etwa 50 Helligkeitsstufen beträgt der Entscheidungsgehalt
  - H0 = ld(720 * 576 * 50) = 24,3 bit
- Informationsfluss (25 Hz)
  - 5,6 bit * 720/2 * 576 * 25 Hz = 29 Mbit/s

Ende der Wiederholung



4.3 Entropie

- Haben mögliche Ereignisse oder Zustände eines Systems verschiedene Wahrscheinlichkeiten, wird aus diesen der mittlere Informationsgehalt bestimmt. ➔ Entropie (aus Thermodynamik)
- Entropie (gr., „sich zu etwas hinwenden“)
- natürliche Systeme streben Gleichgewicht an.
- je wahrscheinlicher ein Zustand desto größer dessen Entropie (= Maß für Unordnung)
- Entropie ist Maß für Wahrscheinlichkeit, mit der ein System vom Zustand i nach j übergeht.

4.3 Entropie

- Haben mögliche Ereignisse oder Zustände eines Systems verschiedene Wahrscheinlichkeiten, wird aus diesen der mittlere Informationsgehalt bestimmt. ➔ Entropie (aus Thermodynamik)
- Entropie (gr., „sich zu etwas hinwenden“)
- natürliche Systeme streben Gleichgewicht an.
- je wahrscheinlicher ein Zustand desto größer dessen Entropie (= Maß für Unordnung)

Informatik

Der **mittlere Informationsgehalt** einer Informationsquelle nennt man **Entropie H**

$$H = \sum_i I(N_i)\, p_i = \sum_i p_i\, ld(1/p_i)$$

$$= -\sum_i p_i\, ld(p_i)\ ;\ i = 1, 2, ..., n$$

wenn ferner gilt $\sum_i p_i = p_1 + p_2 + ... + p_n = 1$

- **Entropie** ist **maximal**, wenn alle Zustände (Nachrichten) gleich wahrscheinlich sind,
- wenn gilt: $\mathbf{p_i = 1/n}$ für alle i = 1, ..., n
- dann gilt $H_{max}(n) = H_0(n)$

4.4 Redundanz

- Redundanz ist die Differenz aus maximalem und tatsächlichem Informationsgehalt.
- Redundanz bedeutet „Weitschweifigkeit“, „Überfluss“ (lat. redundare)
- Redundanz kann nützlich sein. Störungen können ggf. kompensiert werden:
  - Vorlesung am Mittwoch, den 24.11.04 im 2. Block von 10:15 – 11:45 ...
  - %orle{%ung am M[@~woch, den 2?.11.04 im 2. Bl!“k von 1-:ß5 – |1:45 ...

4.4 Redundanz (2)

- Die Maße **Redundanz** und **Wirkungsgrad** charakterisieren eine Informationsquelle
- Redundanz: $\mathbf{R = |H(n) - H_0(n)|}$
- Wirkungsgrad: $\eta = \mathbf{H(n) / H_0(n)}$

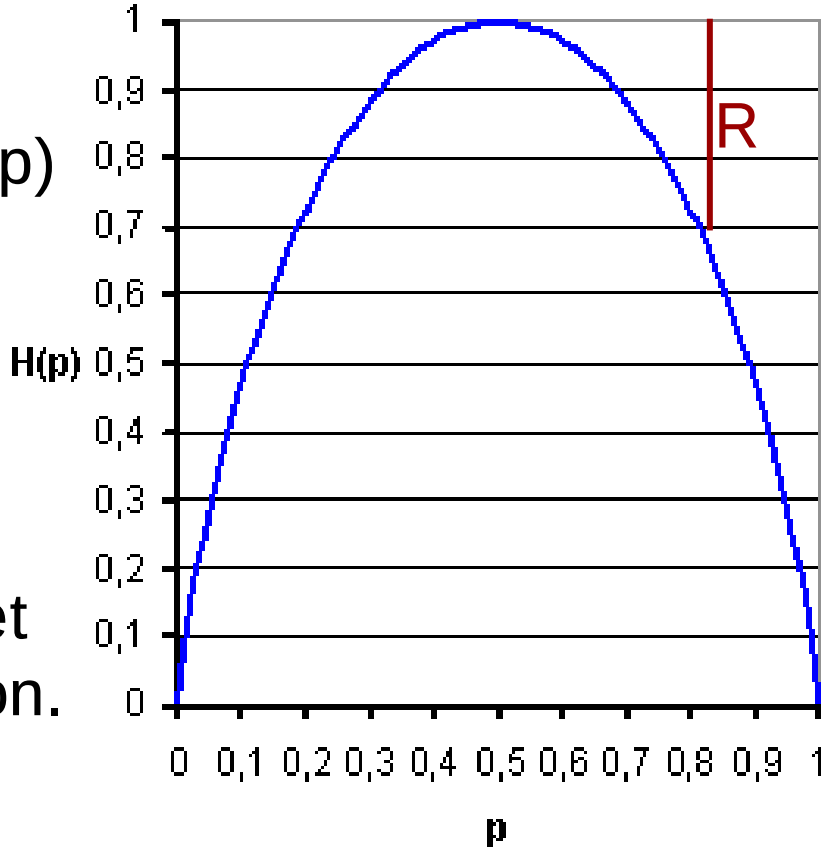
Z	p_i	$-\mathrm{ld}\,p_i$	$\lceil -\mathrm{ld}\,p_i \rceil$
a	0,4	1,322	2
b	0,2	2,322	3
c	0,2	2,322	3
d	0,1	3,322	4
e	0,1	3,322	4

$H_0(5) = 2{,}322$
$H(5) = 2{,}122$
Redundanz:
$R = 0{,}200$
Wirkungsgrad:
$h = 0{,}914$

4.4 Shannon-Funktion

Sendet eine Nachrichtenquelle zwei Signale (0 und 1) aus, so tritt 0 mit der Wahrscheinlichkeit p und 1 mit der Wahrscheinlichkeit $q = 1-p$ auf. H hängt also von p ab.

$$H(p) = -(p \text{ ld } p + q \text{ ld } q)$$
$$= -\text{ p ld p} + (1-\text{p}) \text{ ld } (1-\text{p})$$



Die Funktion H(p) bezeichnet man als die Shannon-Funktion.

5. Datenkomprimierung

1. Wie kann man Daten komprimieren?
2. Komprimierung nach Huffman

5.1 Wie kann man Daten komprimieren?

Informatik

Beispiel:

Stellen wir uns eine Schwarz-Weiss-Grafik vor.

Diese bestehe in B x H aus 300 x 200 Punkten.

Die erste, weisse Zeile kann nun auf unterschiedliche Art angegeben werden:

- *weiss, weiss, weiss, weiss, ... (300 mal)*
  Jeder Punkt wird unabhängig vom vorhergehenden angegeben.
- *300 x weiss* (wesentlich weniger Daten als 1.)
- *1. Zeile weiss* (noch weniger Daten als 2.)

5.1 Wie kann man Daten komprimieren?

Informatik

Beispiel: Textdatei soll komprimiert werden.

Ersten Buchstaben in ASCII, weitere relativ dazu...

Zeichen	**H**	**U**	**B**	**E**	**R**	**T**
Dezimal	72	85	66	69	82	84
Hexadez	48	55	42	45	52	54
Binär	01001000	01010101	01000010	01000101	01010010	01010100
Komprimierung	Umwandlung in andere Sequenzen					
Dezimal	72	+13	-19	+3	+13	+2
Binär	01001000	+1101	-10011	+11	+1101	+10
Vereinbarung	Binäre Zahlen müssen einheitliche Länge haben (Nullen voranstellen)					
Binär	01001000	+01101	-10011	+00011	+01101	+00010
Vereinbarung	zusätzliches erstes Bit=1 bedeutet dazuzählen, ersten Bit=0 bedeutet abziehen					
Ergebnis	01001000	101101	010011	100011	101101	100010

5.2 Datenkomprimierung nach Huffman

- höhere Effizienz bei Komprimierung
- Häufig vorkommende Zeichen, z.B. „E“ und „I“, bekommen möglichst kurzen Code, d.h. weniger Binärstellen.
- Verfahren nach Huffmann werden in Programmen wie PKZIP und ARJ angewendet.

Wir beginnen anhand eines Beispiels, denn „EINER IST BESSER ALS KEINER“

5.2 „EINER IST BESSER ALS KEINER“

- Der zu komprimierende Text wird komplett durchsucht
- Jeder vorkommende Buchstabe wird gezählt und bekommt ein Gewicht entsprechend seiner Häufigkeit. (Unterstrich _ steht für Leerzeichen)

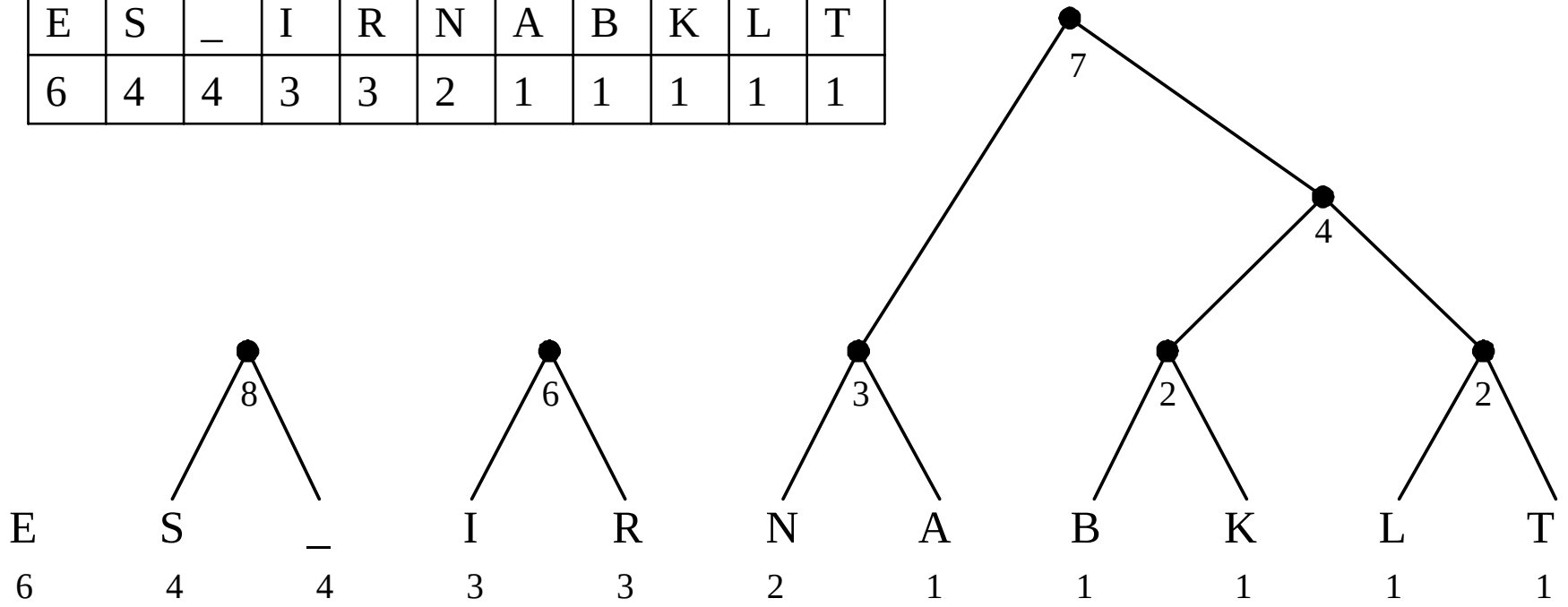
E	S	_	I	R	N	A	B	K	L	T
6	4	4	3	3	2	1	1	1	1	1

Es folgen jetzt noch weitere Schritte ...

5.2 Huffmann – 1. Schritt

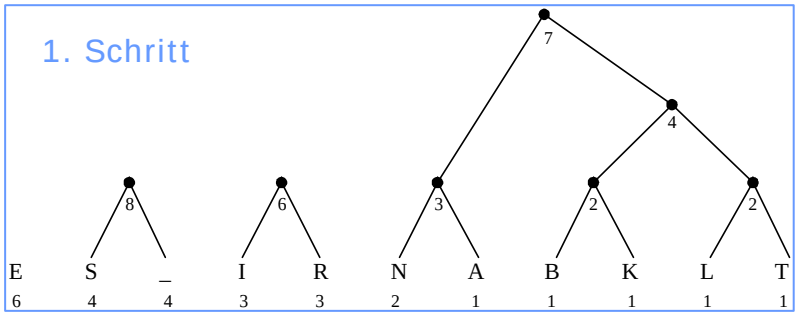
Es wird ein Baum erstellt. Jeweils **2 Knoten mit der geringsten Gewichtung** werden zu einem neuen Knoten zusammengefasst und erhält die **Summe der einzelnen Gewichtungen** der Knoten, aus dem er hervorgeht. Zusammengefasste Knoten werden **als verwendet markiert** und nicht weiter benutzt.

E	S	_	I	R	N	A	B	K	L	T
6	4	4	3	3	2	1	1	1	1	1



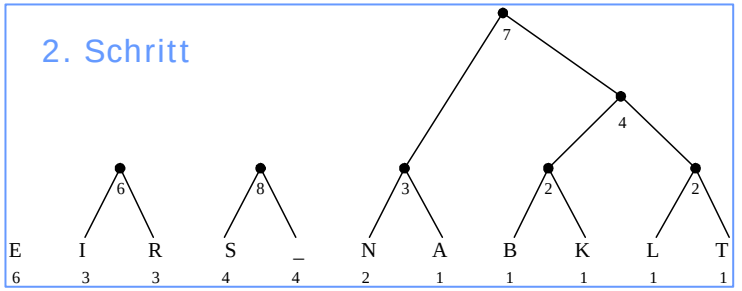
5.2 Huffmann – 2. Schritt

Als nächsten müssen die beiden Knoten mit der Gewichtung 6 zusammengefasst werden. Damit sich die Ästen in den darauf folgenden Schritten nicht überschneiden, wird (nur der Übersicht halber) umsortiert. Das „E“ wird mit dem Zweig rechts daneben vertauscht



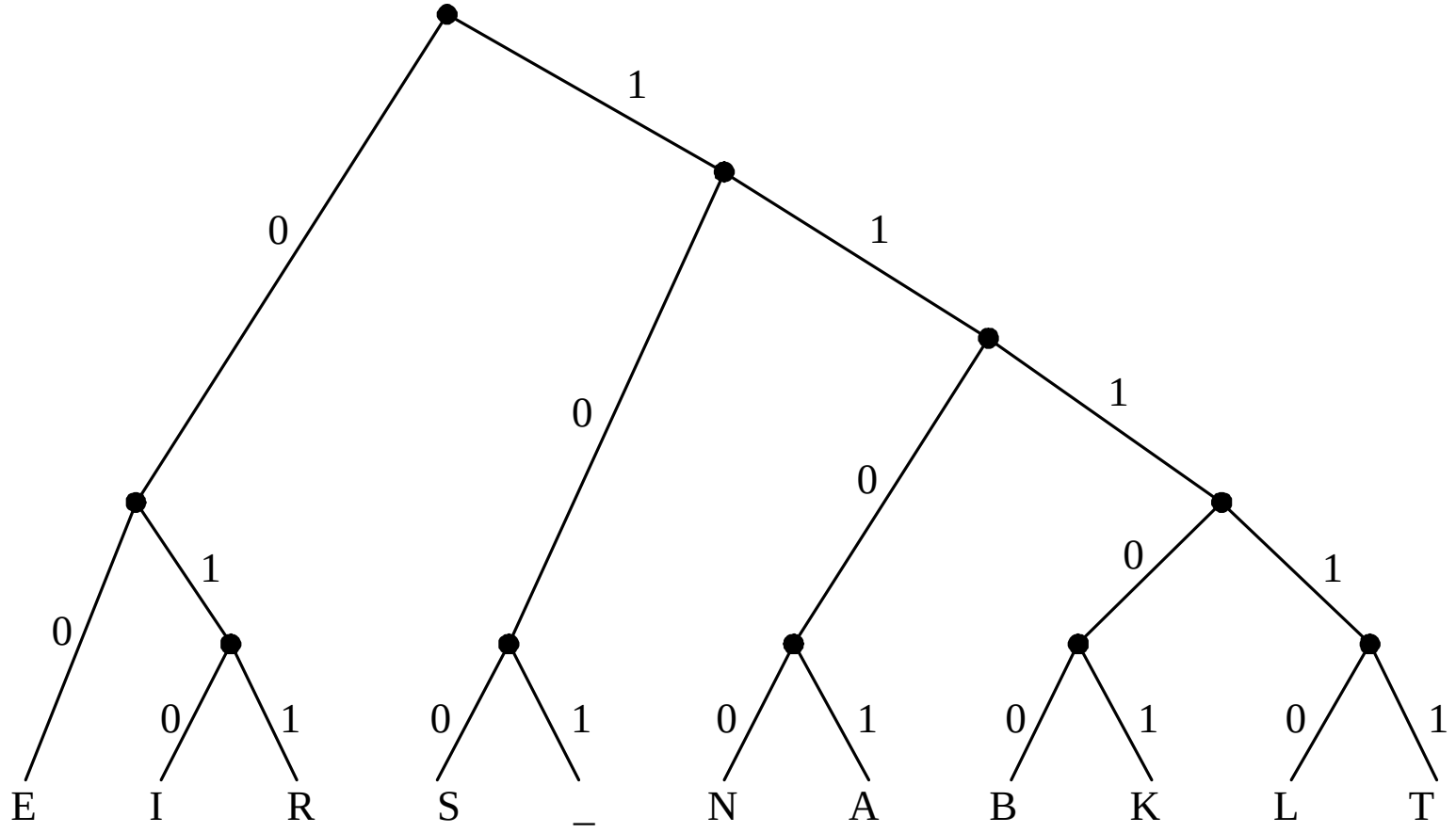
5.2 Huffmann – 3. Schritt

Fertigstellung des Baumes



5.2 Huffmann – 4. Schritt

Nun wählt man frei für je eine Richtung (links oder rechts) eines Astes des Baumes eine Binärzahl. Hier ist die Null für links und die Eins für rechts gewählt. Der fertige Baum sieht dann so aus:



5.2 Huffmann – Dekomprimierung

Beim Dekomprimieren muss der Baum bekannt sein. (Dieser wird in der Archivdatei mit abgespeichert.) Zur Rückgewinnung der ursprüngliche Daten wird für jedes Zeichen oben am Baum angefangen. Man durchläuft dann je nach Binärzahl solange die Verzweigungen, bis man unten angekommen ist. :

E	⇨	00	S	⇨	100
I	⇨	010	N	⇨	1100
R	⇨	011	K	⇨	11101

Komprimierung	**E**	**I**	**N**	**E**	**R**	**Sum**
ohne	01000101	01001001	01001110	01000101	01010010	40 Bit
mit	00	010	1100	00	011	14 Bit

Buchstaben, die häufig vorkommen, haben eine kürzere Sequenz. Binärzahlen müssen keine einheitliche Länge haben

5.1 Kompressionsverfahren Übersicht

Klassifikation:

- universelle ↔ spezielle Verfahren
- nicht verlustbehaftet ↔ verlustbehaftet

Zeichenorientierte Verfahren:

- Lauflängencodierung (Running Length Encoding)
- ...

Statistische Verfahren:

- Huffman-Codierung
- ...