

Symbol	Code	Codewortlänge
$x_0$	$\underline{c}_0 = (10)$	$\ell_0 = 2 \text{ Bit}$
$x_1$	$\underline{c}_1 = (110)$	$\ell_1 = 3 \text{ Bit}$
$x_2$	$\underline{c}_2 = (1110)$	$\ell_2 = 4 \text{ Bit}$

Der **durchschnittliche Codewortlänge**  $L$  eines Codes wird durch die Formel

$$L = \sum_{i=1}^N p(s_i) \cdot \ell_i \left[ \frac{\text{Bit}}{\text{Symbol}} \right]$$

Die **Redundanz**  $R$  wird definiert durch die Formel

$$R = L - H(X) \left[ \frac{\text{Bit}}{\text{Symbol}} \right]$$

- $R > 0$  Code kann noch verlustfrei komprimiert werden
- $R = 0$  Code kann nicht mehr verlustfrei komprimiert werden
- $R < 0$  Code wird verlustbehaftet komprimiert

Die Kompressionsrate  $CR$  eines Codierungsverfahrens wird definiert als das Verhältnis der Grösse der ursprünglichen Daten  $D_{orig}$  zur Grösse der komprimierten Daten  $D_{comp}$ :

$$CR = \frac{D_{orig}}{D_{comp}}$$

## Run Length Encoding (RLE)

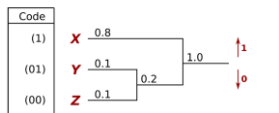
Bsp: ...TERRRRRRRRMAUGGWQCSSSSSSSSSL...

Komprimiert: ...TERA08RMA01AUA02GWQCA10SL...

Die Huffman-Codierung ist ein Verfahren zur optimalen Quellencodierung, das darauf abzielt, die durchschnittliche Codewortlänge zu minimieren. Der Algorithmus funktioniert wie folgt:

1. Liste aller Symbole mit ihren Wahrscheinlichkeiten erstellen.
2. Die beiden Symbole mit den kleinsten Wahrscheinlichkeiten auswählen und zu einem neuen Symbol zusammenfassen, dessen Wahrscheinlichkeit die Summe der beiden ist.
3. Diesen Vorgang wiederholen, bis nur noch ein Symbol übrig ist.
4. Den Baum von unten nach oben durchgehen und jedem Symbol eine Code Zeile zuweisen, wobei links eine 0 und rechts eine 1 hinzugefügt wird

$$P(X) = 0.80 \quad P(Y) = 0.10 \quad P(Z) = 0.10$$

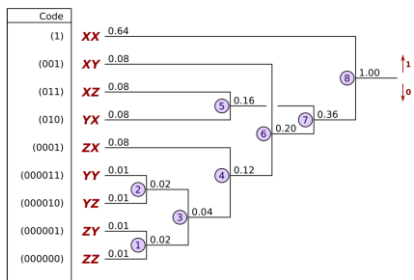


Um die Redundanz weiter zu reduzieren, kann die Huffman-Codierung auf Symbolgruppen (z.B. Paare oder Tripel von Symbolen) angewendet werden, anstatt auf einzelne Symbole. Dafür gilt:

$$p(x_1, x_2, \dots, x_n) = p(x_1) \cdot p(x_2) \cdot \dots \cdot p(x_n)$$

$$H_n(X) = n \cdot H(X) \quad \left[ \frac{\text{Bit}}{n \text{ Symbole}} \right]$$

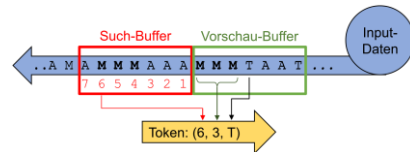
Hierbei ist  $H_n(X)$  die Entropie der Symbolgruppe der Länge  $n$ .



$H, L, R$  sind in diesem Bsp. nun Bit/2 Symbole.

Die LZ77-Codierung ist ein verlustfreies Kompressionsverfahren, das auf der Ersetzung von wiederkehrenden Mustern durch Verweise basiert:

1. Suche im Suchbuffer nach der längsten Übereinstimmung mit dem Anfang des Lookahead-Buffers.
2. Erstelle ein Tripel (Offset, Länge, Nächstes Symbol), wobei:
  - a. **Offset:** die Position der Übereinstimmung im Suchbuffer ist,
  - b. **Länge:** die Länge der Übereinstimmung ist,
  - c. **Nächstes Symbol:** das erste Symbol im Lookahead-Buffer nach der Übereinstimmung ist.
  - d. Bei keiner Übereinstimmung: (0, 0, Nächstes Symbol)
3. Verschiebe den Suchbuffer und den Lookahead-Buffer entsprechend der Länge der Übereinstimmung und des nächsten Symbols.
4. Wiederhole den Vorgang, bis alle Daten codiert sind.



Such-Buffer													Vorschau-Buffer			Offset	Länge	Symbol	
10	9	8	7	6	5	4	3	2	1		A	M	A	M	M	M	0	0	A
Such-Buffer													Vorschau-Buffer				0	0	M
10	9	8	7	6	5	4	3	2	1		A	M	A	M	M	M	2	2	M
10	9	8	7	6	5	4	3	2	1		A	M	A	M	M	A	4	2	A
Such-Buffer													Vorschau-Buffer				6	4	T
10	9	8	7	6	5	4	3	2	1		A	M	A	M	M	A			

Bit ganz rechts im Vorschau-Buffer dürfen nicht codiert werden.

Für den Abschluss des Codes muss eine Lösung definiert werden (Token / Grösse im Header)

Die LZW-Codierung (Lempel-Ziv-Welch) ist ein weiteres verlustfreies Kompressionsverfahren, das auf der Erstellung eines Wörterbuchs basiert:

1. Initialisiere das Wörterbuch mit allen möglichen Symbolen der Eingabedaten oder andere Charsets (ASCII).
2. Lese die Eingabedaten und finde die längste Zeichenkette, die im Wörterbuch vorhanden ist.
3. Gib den Index dieser Zeichenkette im Wörterbuch aus.
4. Füge die längste Zeichenkette plus das nächste Symbol zur Eingabe zum Wörterbuch hinzu.
5. Wiederhole den Vorgang, bis alle Daten codiert sind.

Bsp: AMAMMMAAAMMMTAAT

Index	Eintrag	Token
0...255	ASCII	
256	AM	65 (A)
257	MA	77 (M)
258	AMM	256 (AM)
259	MM	77 (M)
260	MAA	257 (MA)
261	AA	65 (A)
262	AMMM	258 (AMM)
...	...	...