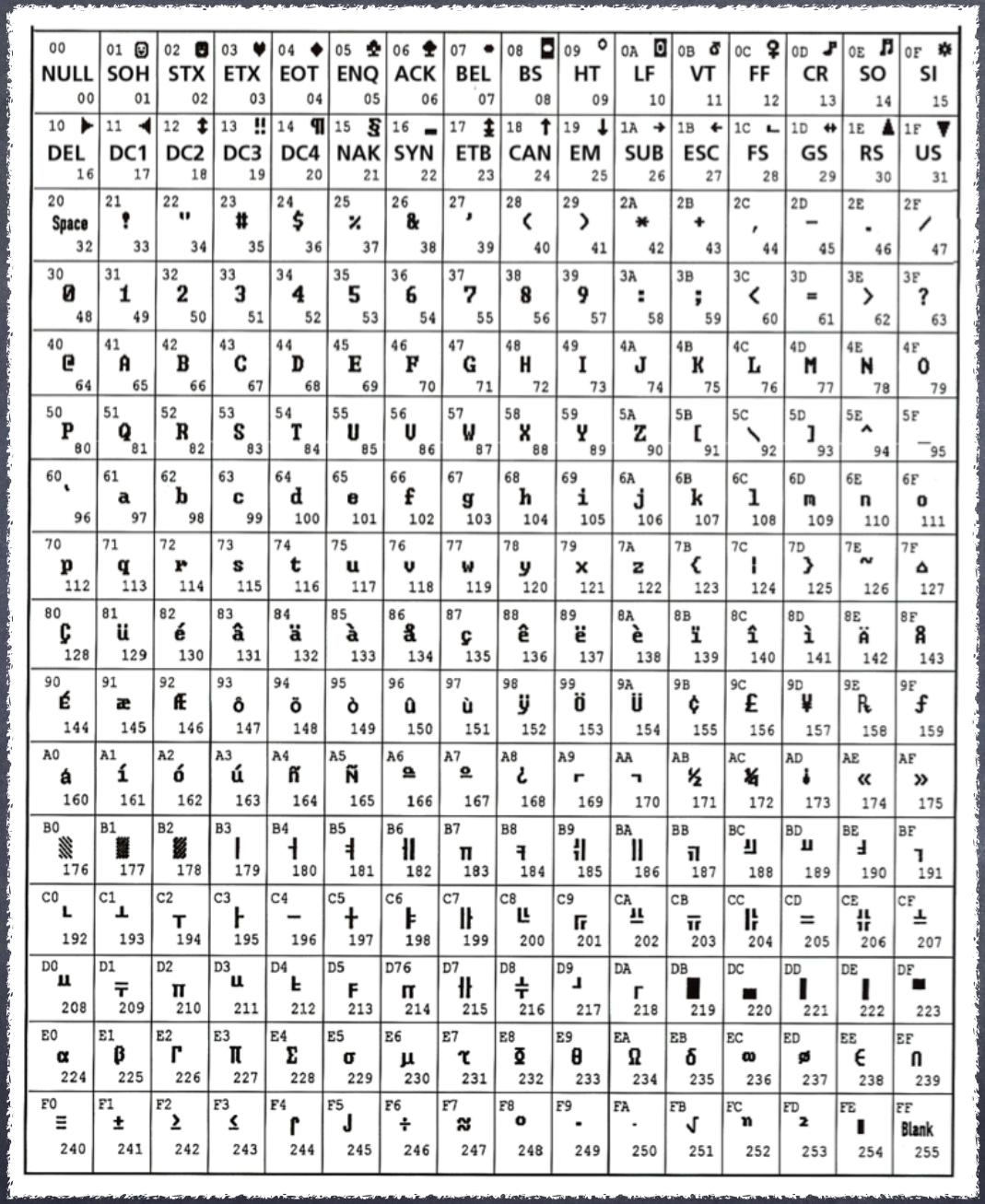
Der Turm zu Babel

Das "Durcheinander der Sprachen" oder "Wie ich lernte, Unicode zu schätzen"

ARJ/August 2015



00 NULL 00	01 SOH 01	02 STX 02	03 ETX 03	04 EOT 04	05 ENQ 05	06 ACK 06	07 BEL 07	08 BS 08	09 HT 09	0A LF 10	0B VT 11	0C FF 12	0D CR 13	0E SO 14	0F SI 15
10 DEL 16	11 DC1 17	12 DC2 18	13 DC3 19	14 DC4 20	15 NAK 21	16 SYN 22	17 ETB 23	18 CAN 24	19 EM 25	1A SUB 26	1B ESC 27	1C FS 28	1D GS 29	1E RS 30	1F US 31
20 Space 32	21 ! 33	22 " 34	23 # 35	24 $ 36	25 % 37	26 & 38	27 ' 39	28 (40	29) 41	2A * 42	2B + 43	2C , 44	2D - 45	2E . 46	2F / 47
30 0 48	31 1 49	32 2 50	33 3 51	34 4 52	35 5 53	36 6 54	37 7 55	38 8 56	39 9 57	3A : 58	3B ; 59	3C < 60	3D = 61	3E > 62	3F ? 63
40 @ 64	41 A 65	42 B 66	43 C 67	44 D 68	45 E 69	46 F 70	47 G 71	48 H 72	49 I 73	4A J 74	4B K 75	4C L 76	4D M 77	4E N 78	4F O 79
50 P 80	51 Q 81	52 R 82	53 S 83	54 T 84	55 U 85	56 V 86	57 W 87	58 X 88	59 Y 89	5A Z 90	5B [91	5C \ 92	5D] 93	5E ^ 94	5F _ 95
60 ` 96	61 a 97	62 b 98	63 c 99	64 d 100	65 e 101	66 f 102	67 g 103	68 h 104	69 i 105	6A j 106	6B k 107	6C l 108	6D m 109	6E n 110	6F o 111
70 p 112	71 q 113	72 r 114	73 s 115	74 t 116	75 u 117	76 v 118	77 w 119	78 y 120	79 x 121	7A z 122	7B { 123	7C | 124	7D } 125	7E ~ 126	7F ⌂ 127
80 Ç 128	81 ü 129	82 é 130	83 â 131	84 ä 132	85 à 133	86 å 134	87 ç 135	88 ê 136	89 ë 137	8A è 138	8B ï 139	8C î 140	8D ì 141	8E Ä 142	8F Å 143
90 É 144	91 æ 145	92 Æ 146	93 ô 147	94 ö 148	95 ò 149	96 û 150	97 ù 151	98 ÿ 152	99 Ö 153	9A Ü 154	9B ¢ 155	9C £ 156	9D ¥ 157	9E ₧ 158	9F ƒ 159
A0 á 160	A1 í 161	A2 ó 162	A3 ú 163	A4 ñ 164	A5 Ñ 165	A6 ª 166	A7 º 167	A8 ¿ 168	A9 ⌐ 169	AA ¬ 170	AB ½ 171	AC ¼ 172	AD ¡ 173	AE « 174	AF » 175
B0 ░ 176	B1 ▒ 177	B2 ▓ 178	B3 | 179	B4 ┤ 180	B5 ╡ 181	B6 ╢ 182	B7 ╖ 183	B8 ╕ 184	B9 ╣ 185	BA ║ 186	BB ╗ 187	BC ╝ 188	BD ╜ 189	BE ╛ 190	BF ┐ 191
C0 └ 192	C1 ┴ 193	C2 ┬ 194	C3 ├ 195	C4 ─ 196	C5 ┼ 197	C6 ╞ 198	C7 ╟ 199	C8 ╚ 200	C9 ╔ 201	CA ╩ 202	CB ╦ 203	CC ╠ 204	CD ═ 205	CE ╬ 206	CF ╧ 207
D0 ╨ 208	D1 ╤ 209	D2 ╥ 210	D3 ╙ 211	D4 ╘ 212	D5 ╒ 213	D76 ╓ 214	D7 ╫ 215	D8 ╪ 216	D9 ┘ 217	DA ┌ 218	DB █ 219	DC ▄ 220	DD ▌ 221	DE ▐ 222	DF ▀ 223
E0 α 224	E1 ß 225	E2 Γ 226	E3 π 227	E4 Σ 228	E5 σ 229	E6 µ 230	E7 τ 231	E8 Φ 232	E9 Θ 233	EA Ω 234	EB δ 235	EC ∞ 236	ED ø 237	EE ∈ 238	EF ∩ 239
F0 ≡ 240	F1 ± 241	F2 ≥ 242	F3 ≤ 243	F4 ⌠ 244	F5 ⌡ 245	F6 ÷ 246	F7 ≈ 247	F8 ° 248	F9 ∙ 249	FA · 250	FB √ 251	FC ⁿ 252	FD ² 253	FE ■ 254	FF Blank 255

Die 8-Bit ANSI-ASCII-Code Tabelle

- Der Unicode kann max. 8 Byte lang sein (64 Bit): U+XXXX'XXXX
- Unicode V2.0 nützt bisher erst 1'114'112 verschiedene Zeichen U+0000'0000 bis U+0010'FFFF
- Die verbreitetste Kodierungsform ist UTF-8 und belegt pro Zeichen max. 4 Byte

Unicode-Bereich (Hexadez.)	**UTF-8 Kodierung (Binär)**	**Bemerkungen**
0000 0000 – 0000 007F	0xxx xxxx	In diesem Bereich (128 Zeichen) entspricht UTF-8 genau dem ASCII-Code: Das höchste Bit ist *0*, die restliche 7-Bit-Kombination ist das ASCII-Zeichen
0000 0080 – 0000 07FF	110xxxxx 10xxxxxx	Das erste Byte enthält binär 11xxxxxx, die folgenden Bytes 10xxxxxx; die *x* stehen für die fortlaufende Bitkombination des Unicode-Zeichens. Die Anzahl der Einsen vor der höchsten 0 im ersten Byte ist die Anzahl der Bytes für das Zeichen. (In Klammern jeweils die theoretisch maximal möglichen.)
0000 0800 – 0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx	
0001 0000 – 0010 FFFF [0001 0000 – 001F FFFF]	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx	

- Das erste Byte eines UTF-8-kodierten Zeichens nennt man dabei Start-Byte, weitere Bytes nennt man Folgebytes. Startbytes enthalten also die Bitfolge 11xxxxxx oder 0xxxxxxx, Folgebytes immer die Bitfolge 10xxxxxx.

- Ist das höchste Bit des ersten Bytes 0, handelt es sich um ein gewöhnliches ASCII-Zeichen, da ASCII eine 7-Bit-Kodierung ist und die ersten

  128 Unicode-Zeichen den ASCII-Zeichen entsprechen. Damit sind alle ASCII-Dokumente automatisch aufwärtskompatibel zu UTF-8.

- Ist das höchste Bit des ersten Bytes 1, handelt es sich um ein Mehrbytezeichen, also ein Unicode-Zeichen mit einer Zeichennummer größer als 127.

- Sind die höchsten beiden Bits des ersten Bytes 11, handelt es sich um das Start-Byte eines Mehrbytezeichens, sind sie 10, um ein Folge-Byte.

- Die lexikalische Ordnung nach Byte-Werten entspricht der lexikalischen Ordnung nach Buchstaben-Nummern, da höhere Zeichennummern mit entsprechend mehr 1-Bits im Start-Byte kodiert werden.

- Bei den Start-Bytes von Mehrbyte-Zeichen gibt die Anzahl der höchsten 1-Bits die gesamte Bytezahl des als Mehrbyte-Zeichen kodierten Unicode-Zeichens an. Anders interpretiert, die Anzahl der 1-Bits links des höchsten 0-Bits entspricht der Anzahl an Folgebytes plus eins, z. B. 1110xxxx 10xxxxxx 10xxxxxx = drei Bits vor dem höchsten 0-Bit = drei Bytes insgesamt, zwei Bits nach dem höchsten 1-Bit vor dem höchsten 0-Bit = zwei Folgebytes.

- Start-Bytes (0xxx xxxx oder 11xx xxxx) und Folge-Bytes (10xx xxxx) lassen sich eindeutig voneinander unterscheiden. Somit kann ein Byte-Strom auch in der Mitte gelesen werden, ohne dass es Probleme mit der Dekodierung gibt, was insbesondere bei der Wiederherstellung defekter Daten wichtig ist. 10xxxxxx Bytes werden einfach übersprungen, bis ein 0xxxxxxx oder 11xxxxxx Byte gefunden wird. Könnten Start-Bytes und Folge-Bytes nicht eindeutig voneinander unterschieden werden, wäre das Lesen eines UTF-8-Datenstroms, dessen Beginn unbekannt ist, unter Umständen nicht möglich.

- Das gleiche Zeichen kann theoretisch auf verschiedene Weise kodiert werden (Zum Beispiel „a“ als 0110 0001 oder fälschlich als 11000001 10100001). Jedoch ist nur die jeweils kürzest mögliche Kodierung erlaubt.

Zeichen	Unicode	Unicode (Binär)	UTF-8 (Binär)	UTF-8 (Hexadez.)
Buchstabe *y*	U+0079	00000000 01111001	01111001	0x79
Buchstabe *ä*	U+00E4	00000000 1110 0100	11000011 1010 0100	0xC3 0xA4
Zeichen für eingetragene Marke ®	U+00AE	0000 0000 1010 1110	11000010 1010 1110	0xC2 0xAE
Eurozeichen €	U+20AC	0010 0000 1010 1100	11100010 10000010 10101100	0xE2 0x82 0xAC

UTF-8 Beispiele

UTF-8 Eingabemethoden

- Notation bei HTML und XML: &#0000; für dezimale Notation bzw. &#x0000; für hexadezimale Notation, wobei das 0000 die Unicode-Nummer des Zeichens darstellt.

- Ab Windows 2000 kann in einigen Programmen (genauer in RichEdit-Feldern) der Code dezimal als Alt+<dezimales Unicode> auf dem numerischen Tastaturfeld eingegeben werden.

- Ab Microsoft Word 2002 kann Unicode auch hexadezimal eingegeben werden, indem im Dokument <Unicode> oder U+<Unicode> eingetippt wird und anschließend die Tastenkombination Alt+C im Dokument bzw. Alt+X in Dialogfeldern gedrückt wird.

- In Powerpoint: Alt-Taste gedrückt halten und auf dem Zahlenblock den Unicode in Dezimal eingeben. Z.B. Für den Pipe: Unicode in Hexadezimal 007c in Dezimal 124. Somit ALT + 124 (Auf dem Zahlenblock eingeben!)

- Ob das entsprechende Unicode-Zeichen auch tatsächlich am Bildschirm erscheint, hängt davon ab, ob die verwendete Schriftart eine Glyphe für das gewünschte Zeichen (also eine Grafik für die gewünschte Zeichen-nummer) enthält. WIN: Siehe charmap.exe

Aufgaben

- Untersuchen sie die Dateien:
  - Sample1_ascii (Enthält 7-bit ASCII-Zeichen)
  - Sample2_utf8 (Enthält 7-Bit ASCII-Zeichen)
  - Sample3_ascii (Enthält 8-Bit ASCII-Zeichen)
  - Sample4_utf8 (Enthält 8-Bit ASCII-Zeichen)
  - Sample5_utf8 (Enthält 4 chinesische Zeichen)

  je mit einem Hex-Editor und einem Texteditor!
  (Sie finden diese Files auf dem BSCW!)
- Was stellen sie fest bezüglich der Länge und Codierung der UTF-8 Zeichen gegenüber den ASCII-Zeichen?
- Was will uns Sample5_utf8 sagen?
- Erstellen sie ein Word-File mit Schriftzeichen, die ihre PC-Tastatur nicht zur Verfügung stellt. Z.B. das Euro-Zeichen. Benutzen sie die besprochenen UTF-8 Eingabemethoden!