64-040 Modul IP7: Rechnerstrukturen

4. Textkodierung

Norman Hendrich & Jianwei Zhang

Universität Hamburg
MIN Fakultät, Department Informatik
Vogt-Kölln-Str. 30, D-22527 Hamburg
{hendrich,zhang}@informatik.uni-hamburg.de

WS 2010/2011

Inhalt

Darstellung von Texten

- Ad-Hoc Kodierungen
  - Flaggen-Alphabet
  - Braille-Code
  - Morse-Code
- ASCII und ISO-8859-1
- Unicode

Wiederholung: Zeichenkette

Zeichenkette (engl. *string*): Eine Folge von Zeichen

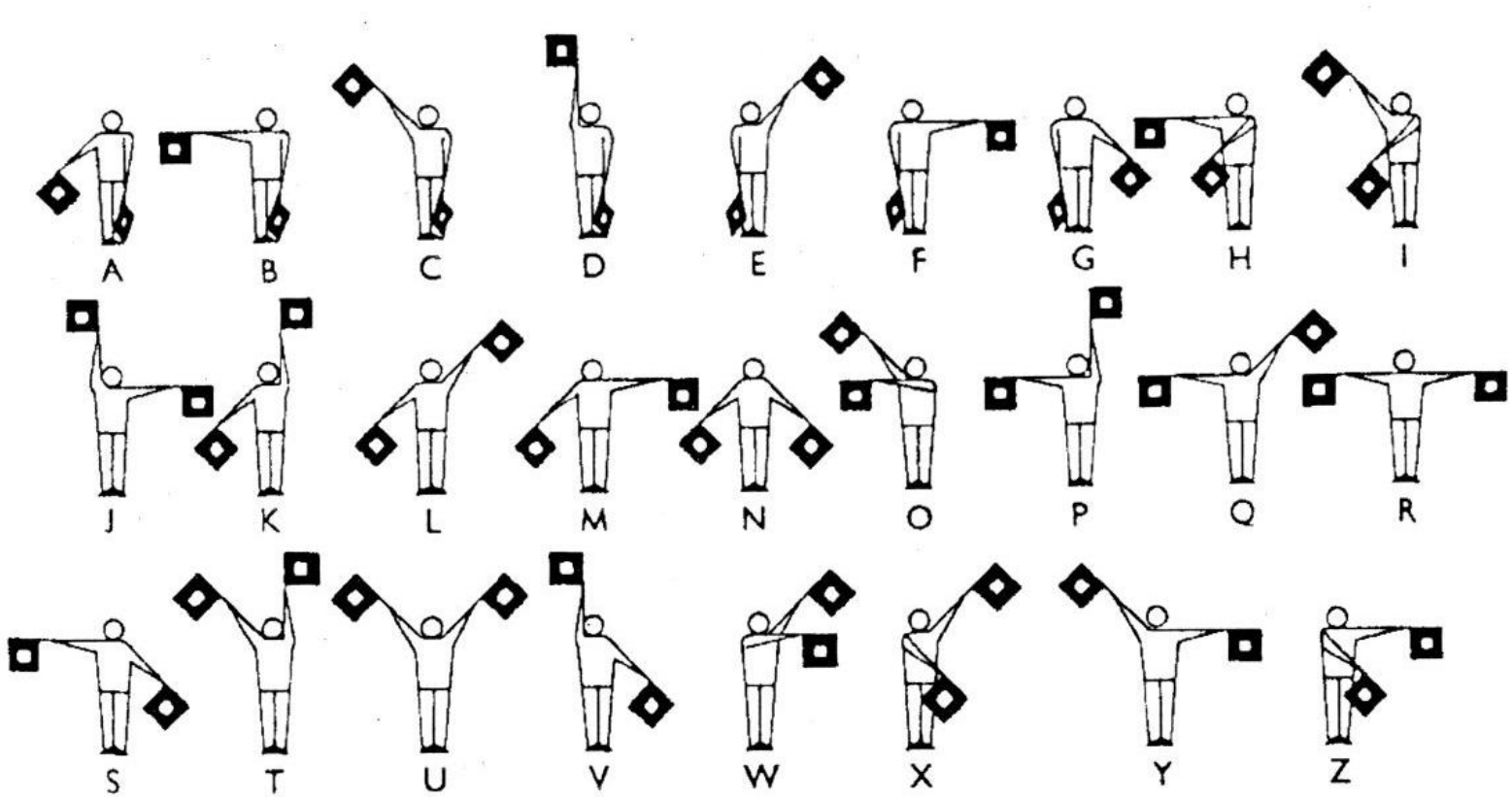
Wort (engl. *word*): Eine Folge von Zeichen, die in einem gegebenen Zusammenhang als Einheit bezeichnet wird. Worte mit 8 Bit werden als **Byte** bezeichnet.

Stelle (engl. *position*): Die Lage/Position eines Zeichens innerhalb einer Zeichenkette.

Beispiel

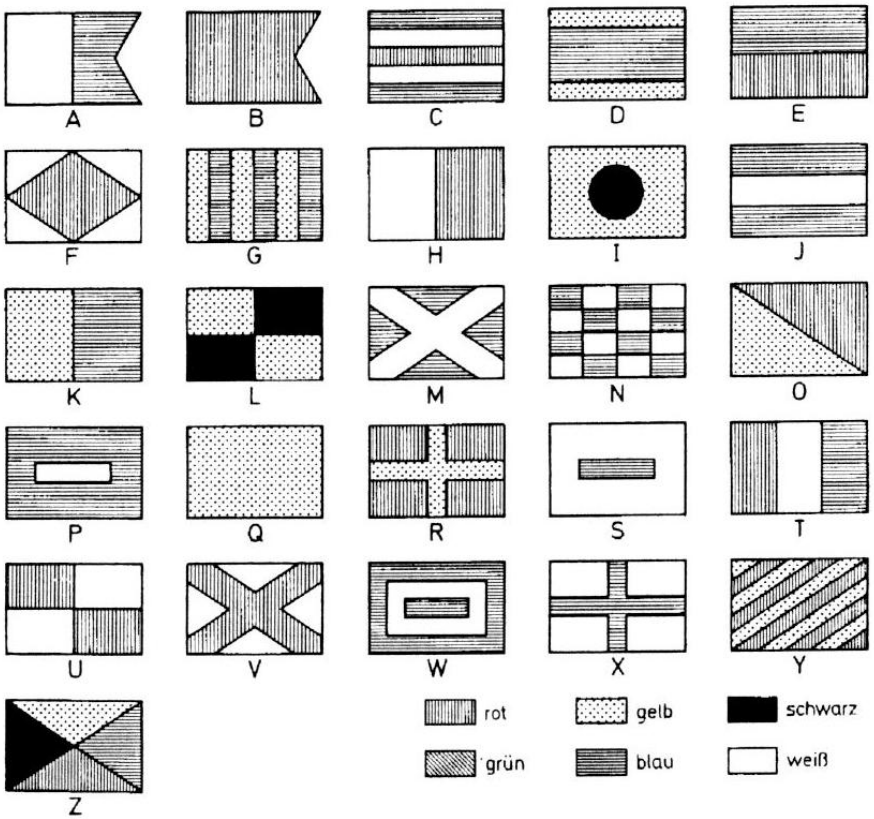
- `s = H e l l o , w o r l d !`

Flaggen-Signale



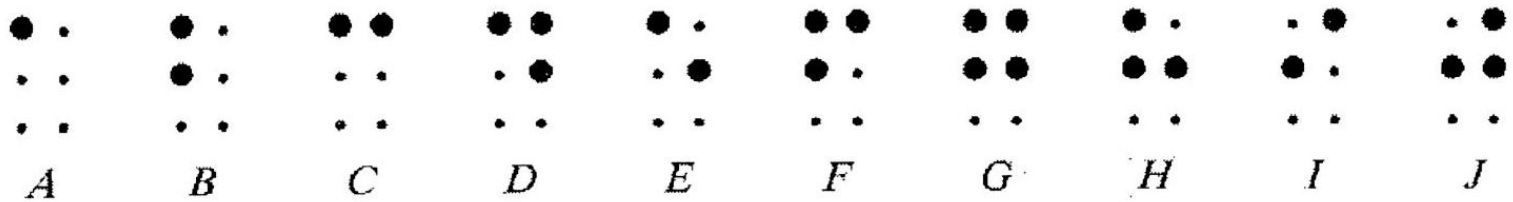
(Bauer & Goos, 1981)

Flaggen-Alphabet

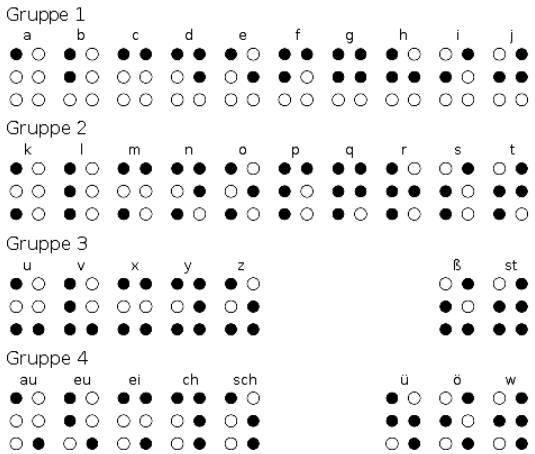


(Bauer & Goos, 1981)

Braille: Blindenschrift



- Symbole als 2x3 Matrix (geprägte Punkte)
- Erweiterung auf 2x4 Matrix (für Computer)
- bis zu 64 (256) mögliche Symbole
- diverse Varianten
  - ein Symbol pro Buchstabe
  - ein Symbol pro Silbe
  - Kurzschrift/Steno



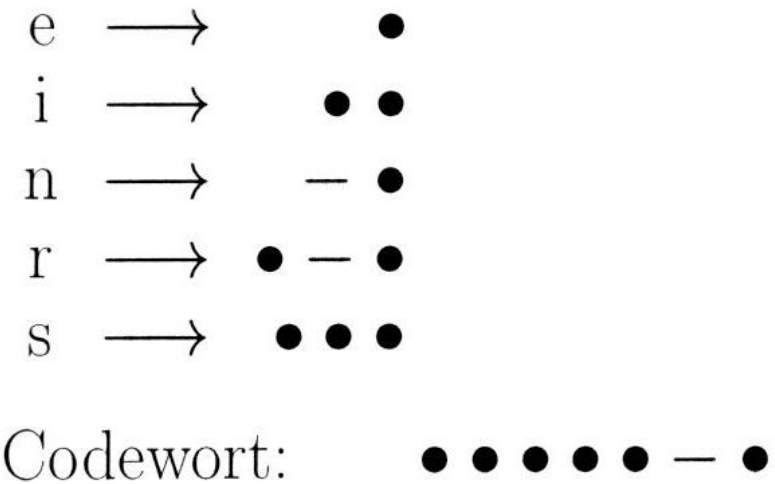
Morse-Code

Codetabelle des Morse-Codes

a	● –	o	– – –	4	● ● ●● –
ä	● – ● –	ö	– – – ●	5	● ● ● ● ●
å	● – – ● –	p	● – – ●	6	– ● ● ●●
b	– ● ● ●	q	– – ● –	7	– – ● ● ●
c	– ● – ●	r	● – ●	8	– – – ●●
ch	– – – –	s	● ● ●	9	– – – – ●
d	– ●●	t	–		
e	●	u	●● –	.	● – ● – ● –
é	●● – ●●	ü	●● – –	,	– – ●● – –
f	●● – ●	v	● ● ● –	:	– – – ● ● ●
g	– – ●	w	● – –	-	– ● ● ●● –
h	● ● ●●	x	– ●● –	'	● – – – – ●
i	●●	y	– ● – –	(	– ● – – ● –
j	● – – –	z	– – ●●	?	●● – – ●●
k	– ● –			“	● – ●● – ●
l	● – ●●	0	– – – – –	Notruf	● ● ● – – – ● ● ●
m	– –	1	● – – – –	SP	●●
n	– ●	2	●● – – –	Anfang	– ● – ● –
ñ	– – ● – –	3	● ● ● – –	Ende	● ● ● – ● –

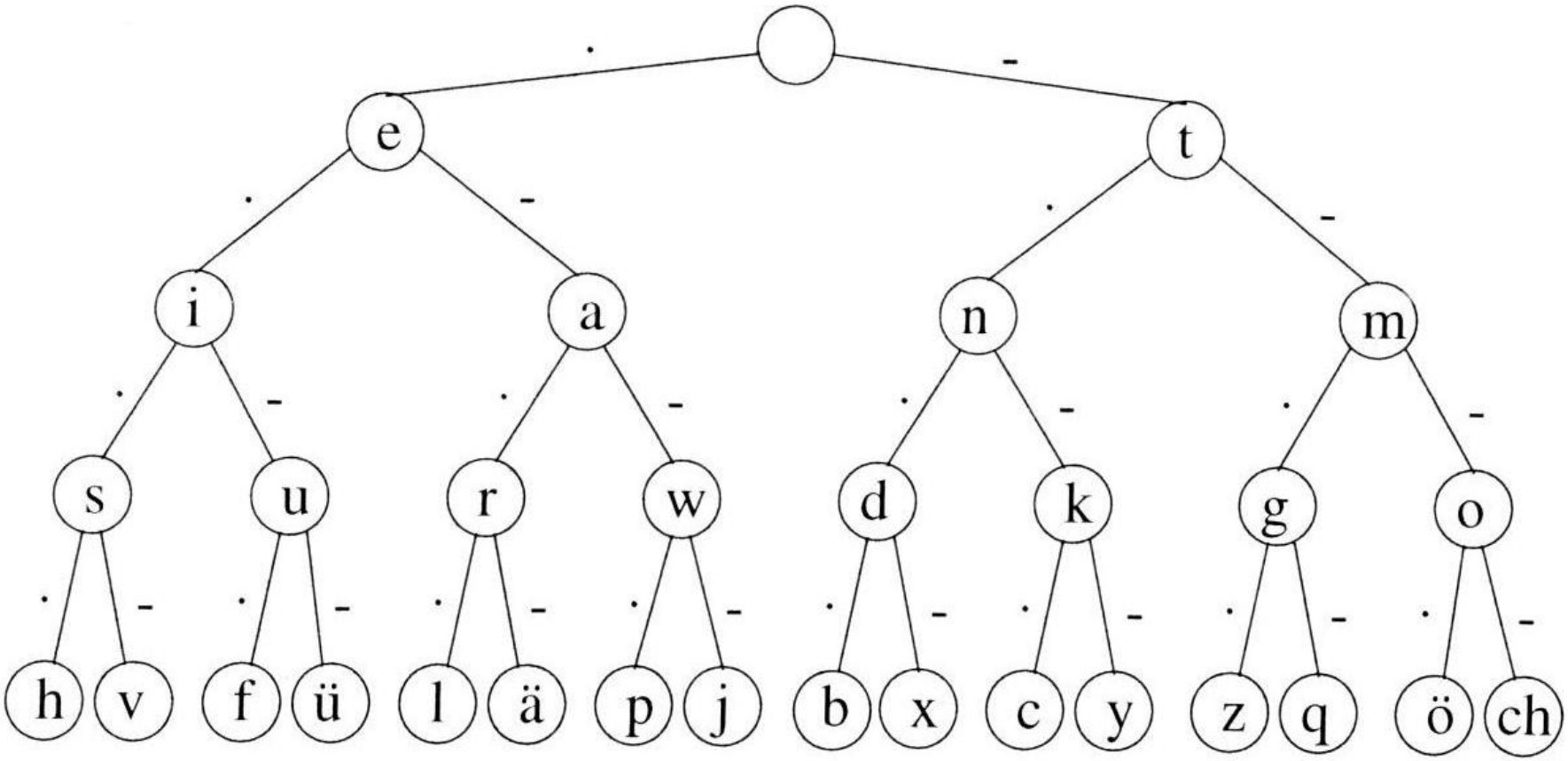
Morse-Code: Eindeutigkeit?

- Punkt steht für kurzen Ton, Strich für langen Ton



- bestimmte Morse-Sequenzen sind mehrdeutig
- Pause zwischen den Symbolen notwendig

Morse-Code: Baumdarstellung (Ausschnitt)



- Anordnung der Symbole entsprechend ihrer Kodierung

ASCII

American Standard Code for Information Interchange

- eingeführt 1967, aktualisiert 1986: ANSI X3.4-1986
- viele Jahre der dominierende Code für Textdateien
- alle Zeichen einer typischen Schreibmaschine
- Erweiterung des früheren 5-bit Fernschreiber-Codes

- 7-bit pro Zeichen, 128 Zeichen insgesamt
- 95 druckbare Zeichen: Buchstaben, Ziffern, Sonderzeichen (Kodierung im Bereich 21..7E)
- 33 Steuerzeichen (engl: *control characters*) (0..1F,7F)

ASCII: Codetabelle

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1...	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2...	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

- SP = Leerzeichen, CR = carriage-return, LF = line-feed
- ESC = escape, DEL = delete, BEL = bell, usw.

de.wikipedia.org/wiki/ASCII

ISO-8859 Familie

- Erweiterung von ASCII um Sonderzeichen und Umlaute
- 8-bit Kodierung: bis max. 256 Zeichen darstellbar

- Latin-1: Westeuropäisch
- Latin-2: Mitteleuropäisch
- Latin-3: Südeuropäisch
- Latin-4: Baltisch
- Latin-5: Kyrillisch
- Latin-6: Arabisch
- Latin-7: Griechisch
- usw.

- immer noch nicht für mehrsprachige Dokumente geeignet

ISO-8859-1: Codetabelle (1)

Erweiterung von ASCII für westeuropäische Sprachen

Code	…0	…1	…2	…3	…4	…5	…6	…7	…8	…9	…A	…B	…C	…D	…E	…F
0…	Sonderzeichen															
1…																
2…	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3…	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4…	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5…	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6…	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7…	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
8…	Sonderzeichen															
9…																
A…	NBSP	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	SHY	®	¯
B…	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C…	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D…	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E…	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F…	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

ISO-8859-1: Codetabelle (2)

Sonderzeichen gemeinsam für alle 8859 Varianten

Code	…0	…1	…2	…3	…4	…5	…6	…7	…8	…9	…A	…B	…C	…D	…E	…F
0…	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1…	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2…	wie ISO/IEC 8859, Windows-125X und US-ASCII															
3…																
4…																
5…																
6…																
7…																DEL
8…	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9…	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
A…	wie ISO/IEC 8859-1 und Windows-1252															
B…																
C…																
D…																
E…																
F…																

ISO-8859-2

Erweiterung von ASCII für slawische Sprachen

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1...	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2...	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL
8...	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9...	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
A...	NBSP	Ą	˘	Ł	¤	Ľ	Ś	§	¨	Š	Ş	Ť	Ź	SHY	Ž	Ż
B...	°	ą	˛	ł	´	ľ	ś	ˇ	¸	š	ş	ť	ź	˝	ž	ż
C...	Ŕ	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
D...	Đ	Ń	Ň	Ó	Ô	Ő	Ö	×	Ř	Ů	Ú	Ű	Ü	Ý	Ţ	ß
E...	ŕ	á	â	ă	ä	ĺ	ć	ç	č	é	ę	ë	ě	í	î	ď
F...	đ	ń	ň	ó	ô	ő	ö	÷	ř	ů	ú	ű	ü	ý	ţ	˙

Microsoft: Codepage 437, 850, 1252

- Zeichensatz des IBM-PC ab 1981
- Erweiterung von ASCII auf einen 8-bit Code
- einige Umlaute (westeuropäisch)
- Graphiksymbole

- Details: http://de.wikipedia.org/wiki/Codepage_437
- verbesserte Version: Codepage 850, 858 (€-Symbol an `0xD5`)
- Codepage 1252 entspricht (weitgehend) ISO-8859-1
- Sonderzeichen liegen an anderen Positionen als bei ISO-8859

Windows: Codepage 850

	*0	*1	*2	*3	*4	*5	*6	*7	*8	*9	*A	*B	*C	*D	*E	*F
0*		☺	☻	♥	♦	♣	♠	•	◘	○	◙	♂	♀	♪	♫	☼
1*	►	◄	↕	‼	¶	§	▬	↨	↑	↓	→	←	∟	↔	▲	▼
2*		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3*	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4*	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5*	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6*	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7*	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	⌂
8*	Ç	ü	é	â	ä	à	å	ç	ê	ë	è	ï	î	ì	Ä	Å
9*	É	æ	Æ	ô	ö	ò	û	ù	ÿ	Ö	Ü	ø	£	Ø	×	ƒ
A*	á	í	ó	ú	ñ	Ñ	ª	º	¿	®	¬	½	¼	¡	«	»
B*	░	▒	▓	│	┤	Á	Â	À	©	╣	║	╗	╝	¢	¥	┐
C*	└	┴	┬	├	─	┼	ã	Ã	╚	╔	╩	╦	╠	═	╬	¤
D*	ð	Ð	Ê	Ë	È	ı	Í	Î	Ï	┘	┌	█	▄	¦	Ì	▀
E*	Ó	ß	Ô	Ò	õ	Õ	µ	þ	Þ	Ú	Û	Ù	ý	Ý	¯	´
F*		±	‗	¾	¶	§	÷	¸	°	¨	·	¹	³	²	■	

Austausch von Texten?

- die meisten gängigen Codes (abwärts-) kompatibel mit ASCII
- unterschiedliche Kodierung für Umlaute (soweit vorhanden)
- unterschiedliche Kodierung der Sonderzeichen

- Unterschiedliche Konvention für Zeilenende
  - DOS/Windows: CR/LF (0D 0A)
  - Unix/Linux: LF
  - Mac OS 9: CR

  - Konverter-Tools: dos2unix, unix2dos, iconv

Unicode: Motivation

- zunehmende Vernetzung und Globalisierung
- internationaler Datenaustausch?
- Erstellung mehrsprachiger Dokumente?
- Unterstützung orientalischer oder asiatischer Sprachen?

- ASCII oder ISO-8859-1 reicht nicht aus
- temporäre Lösungen konnten sich nicht durchsetzen, z.B:
- ISO-2022: Umschaltung zwischen mehreren Zeichensätzen durch Spezialbefehle (*Escapesequenzen*).

- **Unicode** als System zur Kodierung aller Zeichen aller bekannten (lebenden oder toten) Schriftsysteme

Unicode: Versionen und History

- auch abgekürzt als UCS: **Universal Character Set**
- zunehmende Verbreitung (Betriebssysteme, Applikationen)
- Darstellung erfordert auch entsprechende Schriftarten
- www.unicode.org, www.unicode.org/charts/

- 1991 1.0.0: europäisch, nahöstlich, indisch
- 1992 1.0.1: ostasiatisch (Han)
- 1993 akzeptiert als ISO-10646 Standard
- ...
- 2008 5.1.0: über 100.000 Zeichen enthalten

(www.unicode.org, de.wikipedia.org/wiki/Unicode)

Unicode: Schreibweise

- ursprüngliche Version nutzt 16-bit pro Zeichen
- die sogenannte „*Basic Multilingual Plane*“
- Schreibweise hexadezimal als `U+xxxx`
- Bereich von `U+0000` .. `U+FFFF`
- Schreibweise in Java-Strings: `\uxxxx`
  z.B. `\u03A9` für Ω, `\u20A0` für das €-Symbol

- mittlerweile mehr als 2^{16} Zeichen
- Erweiterung um „*Extended Planes*“
- `U+10000` .. `U+10FFFF`

Unicode: in Webseiten (HTML)

- HTML-Header informiert über verwendeten Zeichensatz
- Unterstützung und Darstellung abhängig vom Browser
- Demo: http://www.columbia.edu/kermit/utf8.html

```
<html>
<head>
<META http-equiv="Content-Type"
      content="text/html;
      charset=utf-8">
<title>UTF-8 Sampler</title>
</head>
...
```

Unicode: Demo (Quick brown fox)

1. **English:** The quick brown fox jumps over the lazy dog.
2. **Jamaican:** Chruu, a kwik di kwik brong fox a jomp huova di liezi daag de, yu no siit?
3. **Irish:** "An bfuil do chroí ag bualadh ó fhaitíos an ghrá a mheall lena phóg éada ó shlí do leasa thú?" "D'fhuascail Íosa Úrmhac na hÓige Beannaithe pór Éava agus Ádaimh."
4. **Dutch:** Pa's wijze lynx bezag vroom het fikse aquaduct.
5. **German:** Falsches Üben von Xylophonmusik quält jeden größeren Zwerg. (1)
6. **German:** Im finſteren Jagdſchloß am offenen Felsquellwaſſer patzte der affig-flatterhafte kauzig-höfliche Bäcker über ſeinem verſifften kniffligen C-Xylophon. (2)
7. **Norwegian:** Blåbærsyltetøy ("blueberry jam", includes every extra letter used in Norwegian).
8. **Swedish:** Flygande bäckasiner söka strax hwila på mjuka tuvor.
9. **Icelandic:** Sævör grét áðan því úlpan var ónýt.
10. **Finnish:** (5) Törkylempijävongahdus (This is a perfect pangram, every letter appears only once. Translating it is an art on its own, but I'll say "rude lover's yelp". :-D)
11. **Finnish:** (5) Albert osti fagotin ja töräytti puhkuvan melodian. (Albert bought a bassoon and hooted an impressive melody.)
12. **Finnish:** (5) On sangen hauskaa, että polkupyörä on maanteiden jokapäiväinen ilmiö. (It's pleasantly amusing, that the bicycle is an everyday sight on the roads.)
13. **Polish:** Pchnąć w tę łódź jeża lub ośm skrzyń fig.
14. **Czech:** Příliš žluťoučký kůň úpěl ďábelské kódy.
15. **Slovak:** Starý kôň na hŕbe kníh žuje tíško povädnuté ruže, na stĺpe sa ďateľ učí kvákať novú ódu o živote.
16. **Greek** (monotonic): ξεσκεπάζω την ψυχοφθόρα βδελυγμία
17. **Greek** (polytonic): ξεσκεπάζω τὴν ψυχοφθόρα βδελυγμία
18. **Russian:** Съешь же ещё этих мягких французских булок да выпей чаю.
19. **Russian:** В чащах юга жил-был цитрус? Да, но фальшивый экземпляр! ёъ.
20. **Bulgarian:** Жълтата дюля беше щастлива, че пухът, който цъфна, замръзна като гьон.
21. **Sami (Northern):** Vuol Ruota geđggiid leat máŋga luosa ja čuovžža.
22. **Hungarian:** Árvíztűrő tükörfúrógép.
23. **Spanish:** El pingüino Wenceslao hizo kilómetros bajo exhaustiva lluvia y frío, añoraba a su querido cachorro.
24. **Portuguese:** O próximo vôo à noite sobre o Atlântico, põe freqüentemente o único médico. (3)
25. **French:** Les naïfs ægithales hâtifs pondant à Noël où il gèle sont sûrs d'être déçus en voyant leurs drôles d'œufs abîmés.
26. **Esperanto:** Eĥoŝanĝo ĉiuĵaŭde.
27. **Hebrew:** זה כיף סתם לשמוע איך תנצח קרפד עץ טוב בגן.
28. **Japanese** (Hiragana):

    いろはにほへど　ちりぬるを
    わがよたれぞ　つねならむ
    うゐのおくやま　けふこえて
    あさきゆめみじ　ゑひもせず (4)

(http://www.columbia.edu/kermit/utf8.html)

Unicode: Demo

Šota Rustaveli's Veṗxis Ṭ̇aosani, T̈h, *The Knight in the Tiger's Skin* (Georgian):

ვეპხის ტყაოსანი შოთა რუსთაველი

ღმერთსი შემვედრე, ნუთუ კვლა დამხსნას სოფლისა შრომასა, ცეცხლს,
წყალსა და მიწასა, ჰაერთა თანა მრომასა; მომცნეს ფრთენი და
აღვფრინდე, მივჰხვდე მას ჩემსა ნდომასა, დღისით და ღამით
ვჰხედვიდე მზისა ელვათა კრთომაასა.

Tamil poetry of Subramaniya Bharathiyar: சுப்ரமணிய பாரதியார் (1882-1921):

யாமறிந்த மொழிகளிலே தமிழ்மொழி போல் இனிதாவது எங்கும் காணோம்,
பாமரராய் விலங்குகளாய், உலகனைத்தும் இகழ்ச்சிசொலப் பான்மை கெட்டு,
நாமமது தமிழரெனக் கொண்டு இங்கு வாழ்ந்திடுதல் நன்றோ? சொல்லீர்!
தேமதுரத் தமிழோசை உலகமெலாம் பரவும்வகை செய்தல் வேண்டும்.

(http://www.columbia.edu/kermit/utf8.html)

Unicode: Latin-Zeichen

- Zeichen im Bereich `U+0000` bis `U+007F` wie ASCII
  www.unicode.org/charts/PDF/U0000.pdf

- Bereich von `U+0100` bis `U+017F` für Latin-A
  (europäische Umlaute und Sonderzeichen)
  www.unicode.org/charts/PDF/U0100.pdf

- viele weitere Sonderzeichen ab `U+0180`
  (Latin-B, Latin-C, usw.)

Unicode: Mathematische Symbole und Operatoren

- Vielfältige Auswahl von Symbolen und Operatoren

- griechisch: www.unicode.org/charts/PDF/U0370.pdf
- Letterlike Symbols: www.unicode.org/charts/PDF/U2100.pdf

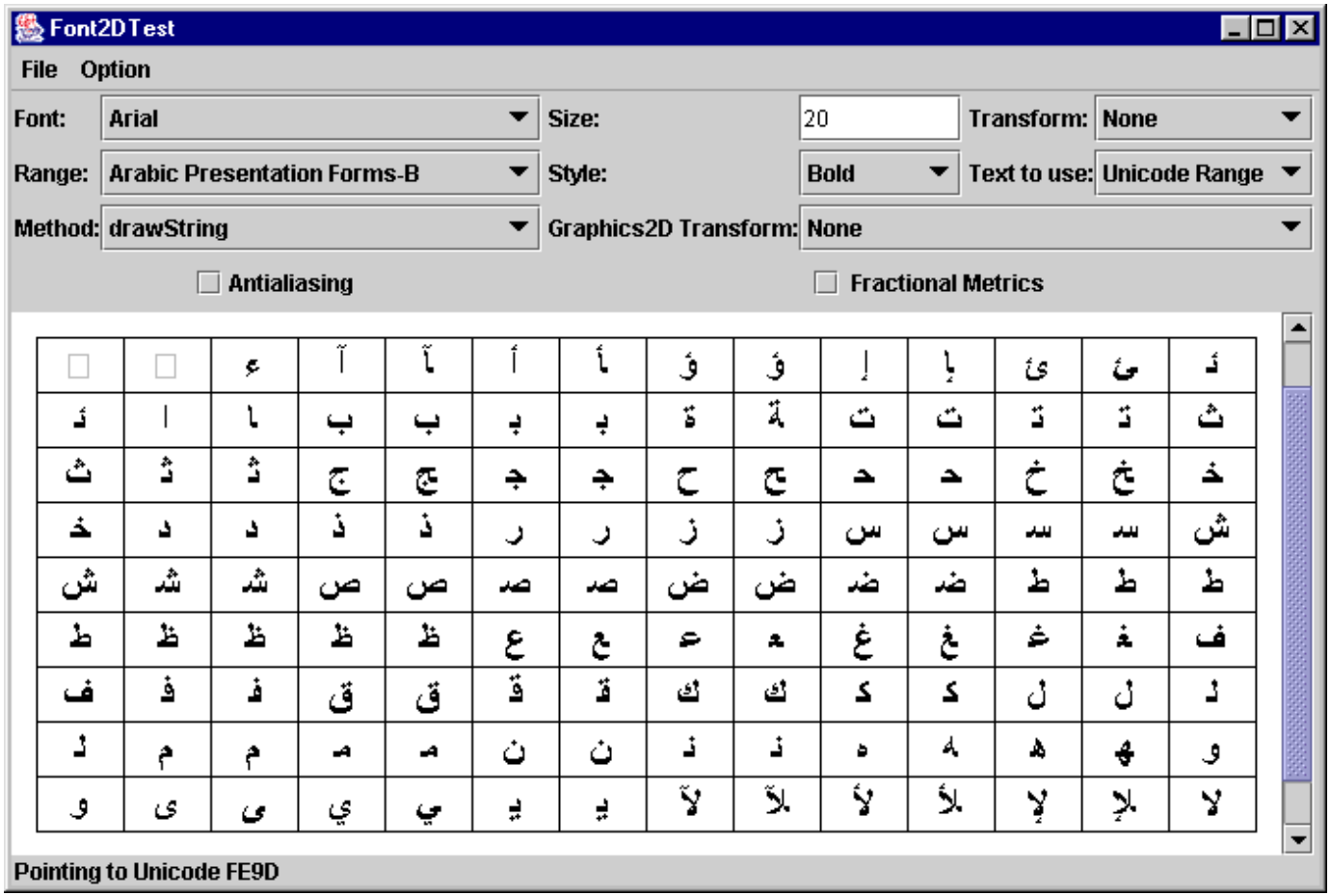
- Pfeile: www.unicode.org/charts/PDF/U2190.pdf
- Operatoren: www.unicode.org/charts/PDF/U2A00.pdf
- ...

- Dingbats: www.unicode.org/charts/PDF/U2700.pdf

Unicode: Asiatische Sprachen

- www.unicode.org/charts/PDF/U3400.pdf
- www.unicode.org/charts/PDF/U4E00.pdf
- chinesisch (traditional/simplified), japanisch, koreanisch
- U+3400 bis U+4BDF
- U+4E00 bis U+9FCF

Unicode: Java2D Fontviewer



(Sun Microsystems, Java JDK examples: Java2D/Font2DTest)

Unicode: Repräsentation?

- 16-bit für jedes Zeichen, bis zu 65536 Zeichen
- schneller Zugriff auf einzelne Zeichen über Arrayzugriffe (Index)
- aber: doppelter Speicherbedarf gegenüber ASCII/ISO-8859-1
- Verwendung u.a. in Java: Datentyp `char`

- ab Unicode-3: mehrere *Planes* zu je 65536 Zeichen
- direkte Repräsentation aller Zeichen erfordert 32-bit/Zeichen
- vierfacher Speicherbedarf gegenüber ISO-8859-1

- bei Dateien ist möglichst kleine Dateigröße wichtig
- effizientere Kodierung üblich: UTF-16 und UTF-8

UTF-8

Zeichen	Unicode	Unicode binär	UTF-8 binär	UTF-8 hexadezimal
Buchstabe *y*	U+0079	00000000 01111001	**0**1111001	0x79
Buchstabe *ä*	U+00E4	00000000 11100100	**110**00011 **10**100100	0xC3 0xA4
Zeichen für eingetragene Marke ®	U+00AE	00000000 10101110	**110**00010 **10**101110	0xC2 0xAE
Eurozeichen €	U+20AC	00100000 10101100	**1110**0010 **10**000010 **10**101100	0xE2 0x82 0xAC
Violinschlüssel	U+1D11E	00000001 11010001 00011110	**11110**000 **10**011101 **10**000100 **10**011110	0xF0 0x9D 0x84 0x9E

- effiziente Kodierung von „westlichen“ Unicode-Texten
- Zeichen werden mit variabler Länge kodiert, 1..4-Bytes
- volle Kompatibilität mit ASCII

de.wikipedia.org/wiki/UTF-8

UTF-8: Algorithmus

Unicode-Bereich (hexadezimal)	UTF-Kodierung (binär)	Anzahl (benutzt)
0000 0000 - 0000 007F	0xxx xxxx	128
0000 0080 - 0000 07FF	110x xxxx 10xx xxxx	1920
0000 0800 - 0000 FFFF	1110 xxxx 10xx xxxx 10xx xxxx	63488
0001 0000 - 0010 FFFF	1111 0xxx 10xx xxxx 10xx xxxx 10xx xxxx	bis 2^{21}

- untere 128 Zeichen kompatibel mit ASCII
- Sonderzeichen westlicher Sprachen je zwei Bytes
- führende Eins markiert Multi-Byte Zeichen
- Anzahl der führenden Einsen gibt Anzahl der Bytes an
- Zeichen ergibt sich als Bitstring aus den `xxx...x`
- theoretisch bis zu sieben Folgebytes a 6-bit: max. 2^{42} Zeichen

Sprach-Einstellungen: Locale

Locale: die Sprach-Einstellungen und Parameter

- auch: `i18n` („internationalization“)
- Sprache der Benutzeroberfläche
- Tastaturlayout/-belegung
- Zahlen-, Währungs, Datums-, Zeitformate

- Linux/POSIX: Einstellung über die Locale-Funktionen der Standard C-Library
- Java: java.util.Locale
- Windows: Einstellung über System/Registry-Schlüssel

dos2unix, unix2dos

- Umwandeln von ASCII-Texten (z.B. Programm-Quelltexte) zwischen DOS/Windows und Unix/Linux Maschinen

- Umwandeln von `a.txt` in Ausgabedatei `b.txt`:

```
dos2unix a.txt -c iso b.txt
dos2unix -c ascii a.txt -c iso b.txt
dos2unix -c mac a.txt b.txt
```

- Umwandeln von Unix nach DOS/Windows, Codepage 850:

```
unix2dos -850 a.txt b.txt
```

iconv

Das Schweizer-Messer zur Umwandlung von Textkodierungen.

Optionen:

- `-f, --from-code=<encoding>` Kodierung der Eingabedatei
- `-t, --to-code=<encoding>` Kodierung der Ausgabedatei
- `-l, --list` Liste der unterstützten Kodierungen ausgaben
- `-o, --output=<filename>` Name der Ausgabedatei

Beispiel:

```
iconv -f=iso-8859-1 -t=utf-8 -o foo.utf8.txt foo.txt
```

base64-Kodierung

Übertragung von (Binär-) Dateien zwischen verschiedenen Rechnern?

- SMTP (Internet Mail-Protokoll) verwendet 7-bit ASCII
- bei Netzwerk-Übertragung müssen alle Rechner/Router den verwendeten Zeichensatz unterstützen

- Verfahren zur Umkodierung der Datei in 7-bit ASCII notwendig
- etabliert ist das **base-64** Verfahren (RFC 2045)

- alle e-mail Dateianhänge und 8-bit Textdateien
- Umkodierung benutzt nur Buchstaben, Ziffern, drei Sonderzeichen

base64-Kodierung: Prinzip

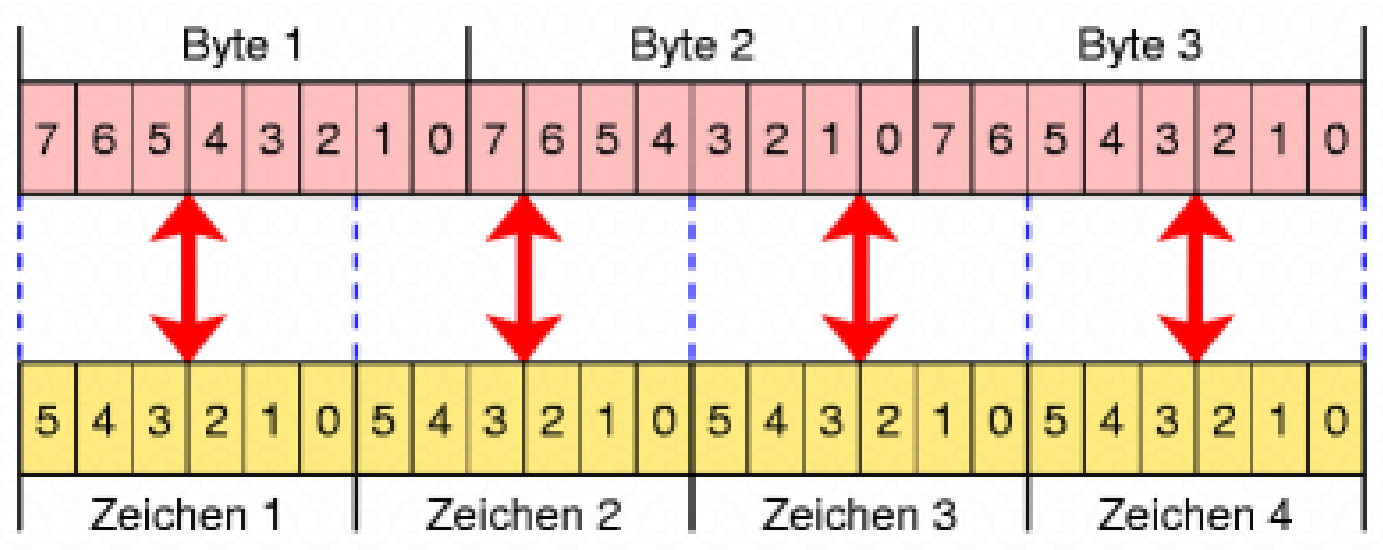
- Kodierung von drei 8-bit Bytes als vier 6-bit Zeichen
- erfordert 64 der verfügbaren 128 7-bit ASCII Symbole

`A...Z`	(Codes 0...25)
`a...z`	(Codes 26...51)
`0...9`	(Codes 52...61)
`+`	(Code 62)
`/`	(Code 63)

`=` Füllzeichen, falls Anzahl der Bytes nicht durch 3 teilbar

`CR` Zeilenumbruch (optional), meistens nach 76 Zeichen

base64-Kodierung: Gruppieren der Bits



- drei 8-bit Zeichen, re-gruppiert als vier 6-bit Blöcke
- Zuordnung des jeweiligen Buchstabens/Ziffer
- ggf. = == am Ende zum Auffüllen
- Übertragung dieser Zeichenfolge ist 7-bit kompatibel
- resultierende Datei ca. 33% größer als das Original

base64-Kodierung: Tools

- im Java JDK enthalten
- aber im inoffiziellen internen Teil

  sun.misc.BASE64Encoder sun.misc.BASE64Decoder

- aber diverse (open-source) Implementierungen verfügbar
- Beispiel: Apache Commons

  org.apache.commons.codec.binary.Base64

  org.apache.commons.codec.binary.Base64InputStream

  org.apache.commons.codec.binary.Base64OutputStream

http://commons.apache.org/codec/

base64-Kodierung: Beispiel

```
public class Base64Demo {
  public static void main( String[] args ) throws IOException {
    byte[] bytes1 = new byte[ 112 ];
    new Random().nextBytes( bytes1 );

    // bytes in String konvertieren und ausgeben, z.B.
    // QFgwDyiQ28/4G...
    // LA3YUbf96Ym2z... rlcQg==
    //
    String s = new BASE64Encoder().encode( bytes1 );
    System.out.println( s );

    // String dekodieren in byte[]
    byte[] bytes2 = new BASE64Decoder().decodeBuffer( s );
    System.out.println( Arrays.equals(bytes1, bytes2) );    // true
  }
}
```

Literatur: Vertiefung

- www.unicode.org/
- Java Tutorial: Internationalization Trail
  java.sun.com/docs/books/tutorial/i18n/index.html