64-040 Modul InfB-RSB
Rechnerstrukturen und Betriebssysteme

https://tams.informatik.uni-hamburg.de/lectures/2024ws/vorlesung/rsb

– Kapitel 5 –

Andreas Mäder

Universität Hamburg
Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Technische Aspekte Multimodaler Systeme

Wintersemester 2024/2025

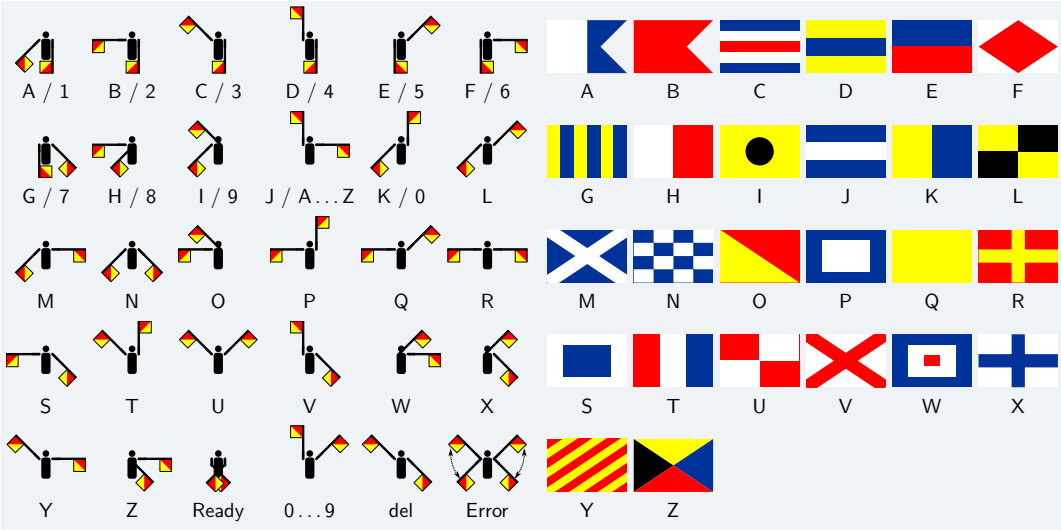
Wiederholung: Zeichen

- **Zeichen**: engl. *character*
  Element z aus einer zur Darstellung von Information vereinbarten, einer Abmachung unterliegenden, endlichen Menge Z von Elementen
- Die Menge Z heißt **Zeichensatz** oder **Zeichenvorrat** engl. *character set*
- **Binärzeichen**: engl. *binary element, binary digit, bit*
  Jedes der Zeichen aus einem Vorrat / aus einer Menge von zwei Symbolen
- **Numerischer Zeichensatz**: Zeichenvorrat aus Ziffern und/oder Sonderzeichen zur Darstellung von Zahlen
- **Alphanumerischer Zeichensatz**: Zeichensatz aus (mindestens) den Dezimalziffern und den Buchstaben des Alphabets, meistens auch mit Sonderzeichen (Leerzeichen, Punkt, Komma usw.)

Wiederholung: Zeichen (cont.)

- **Alphabet**: engl. *alphabet*
  Ein in vereinbarter Reihenfolge geordneter Zeichenvorrat
- **Zeichenkette**: Eine Folge von Zeichen engl. *string*
- **Wort**: engl. *word*
  Zeichenkette, die in einem gegebenen Zusammenhang als Einheit bezeichnet wird
- Worte aus 8 Binärzeichen (8 bit) werden als **Byte** bezeichnet
- **Stelle**: Die Position eines Zeichens in einer Zeichenkette engl. *position*

Flaggen-Signale

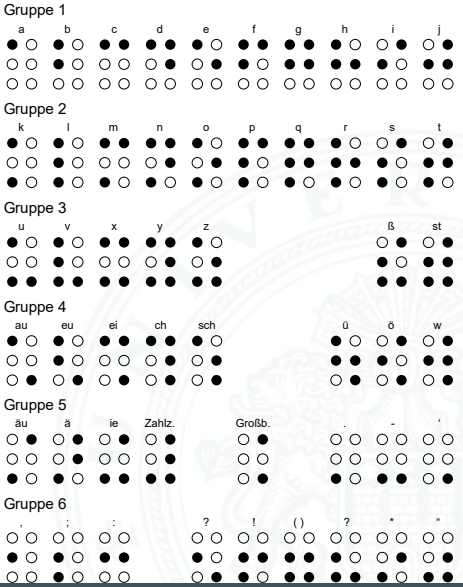


de.wikipedia.org/wiki/Winkeralphabet

de.wikipedia.org/wiki/Flaggenalphabet

Braille: Blindenschrift

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



Codetabelle

● kurzer Ton – langer Ton

A	●–	S	●●●	.	●–●–●–	S-Start	–●–●–
B	–●●●	T	–	,	––●●––	Verst.	●●●–●
C	–●–●	U	●●–	?	●●––●●	S-Ende	●–●–●
D	–●●	V	●●●–	´	●––––●	V-Ende	●●●–●–
E	●	W	●––	!	–●–●––	Error	●●●●●●●●
F	●●–●	X	–●●–	/	–●●–●		
G	––●	Y	–●––	(	–●––●	Ä	●–●–
H	●●●●	Z	––●●	)	–●––●–	À	●––●–
I	●●	0	–––––	&	●–●●●	É	●●–●●
J	●–––	1	●––––	:	–––●●●	È	●–●●–
K	–●–	2	●●–––	;	–●–●–●	Ö	–––●
L	●–●●	3	●●●––	=	–●●●–	Ü	●●––
M	––	4	●●●●–	+	●–●–●	ß	●●●––●●
N	–●	5	●●●●●	-	–●●●●–	CH	––––
O	–––	6	–●●●●	_	●●––●–	Ñ	––●––
P	●––●	7	––●●●	"	●–●●–●	...	
Q	––●–	8	–––●●	$	●●●–●●–		
R	●–●	9	––––●	@	●––●–●	SOS	●●● ––– ●●●

Morse-Code (cont.)

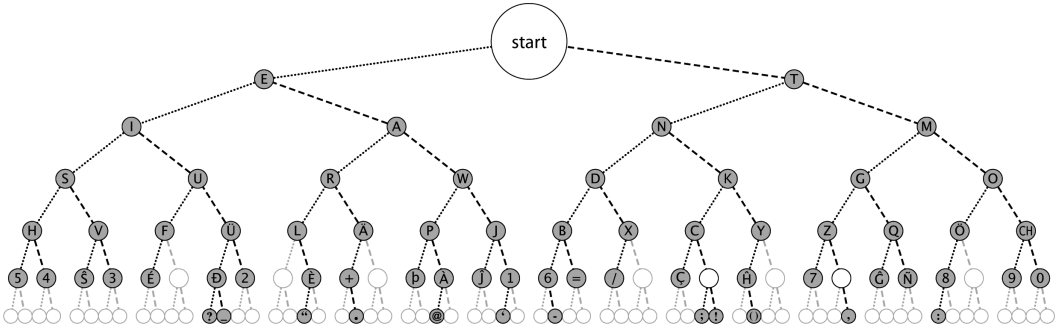
- Eindeutigkeit

Codewort:	●●●●●−●
E	●
I	●●
N	−●
R	●−●
S	●●●

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

- Codierung
  - Häufigkeit der Buchstaben $= 1$ / Länge des Codewortes
  - Effizienz: kürzere Codeworte
  - Darstellung als Codebaum

Morse-Code: Baumdarstellung (Ausschnitt)



- Anordnung der Symbole entsprechend ihrer Codierung

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 (Murray-Code)

- 7-bit pro Zeichen, 128 Zeichen insgesamt
- 95 druckbare Zeichen: Buchstaben, Ziffern, Sonderzeichen 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

de.wikipedia.org/wiki/ASCII

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

ISO-8859 Familie

- Erweiterung von ASCII um Sonderzeichen und Umlaute
- 8-bit Codierung: 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…	*nicht belegt*															
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…	*nicht belegt*															
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	˙

ISO-8859-15

Modifizierte ISO-8859-1 mit € (0xA4)

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*	¡	¢	£	€	¥	Š	§	š	©	ª	«	¬	*SHY*	®	¯
B…	°	±	²	³	Ž	µ	¶	·	ž	¹	º	»	Œ	œ	Ÿ	¿
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

Microsoft: Codepages 437, 850, 1252

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

- `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

Microsoft: Codepage 850

Code	…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 Codierung für Umlaute (soweit vorhanden)
- unterschiedliche Codierung der Sonderzeichen

- Systemspezifische Konventionen für Zeilenende
  - abhängig von Rechner- und Betriebssystem
  - Konverter-Tools: `dos2unix`, `unix2dos`, `iconv`

Betriebssystem	Zeichensatz	Abkürzung	Hex-Code	Escape
Unix, Linux, Mac OS X, AmigaOS, BSD	ASCII	*LF*	`0A`	\n
Windows, DOS, OS/2, CP/M, TOS (Atari)	ASCII	*CR LF*	`0D 0A`	\r\n
Mac OS bis Version 9, Apple II	ASCII	*CR*	`0D`	\r
AIX OS, OS 390	EBCDIC	*NEL*	15	

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**
  Spezialbefehle zur Umschaltung zwischen mehreren Zeichensätzen,
  sog. *Escapesequenzen*

⇒ **Unicode** als System zur Codierung aller Zeichen aller bekannten Schriftsysteme
auch für tote Schriften/Sprachen

Unicode: Versionen und History

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

- 1991 1.0.0: europäisch, nahöstlich, indisch
- 1992 1.0.1: ostasiatisch (Han)
- 1993 akzeptiert als ISO/IEC-10646 Standard

  ...
- 2024 16.0.0: inzwischen 154 998 Zeichen
  - Sprachzeichen, Hieroglyphen etc.
  - Symbole: Satzzeichen, Währungen ($ ... ₿), Pfeile, mathematisch, technisch, Braille, Noten etc.
  - Emojis (3 790 aktuell) / Kombinationen

`home.unicode.org` `unicode.org/main.html` `de.wikipedia.org/wiki/Unicode` `symbl.cc/de/unicode/table`

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, `\u20AC` 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 `kermitproject.org/utf8.html`

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

<META ...
</head>
...
```

Unicode: UTF-8 Sampler

kermitproject.org/utf8.html

I Can Eat Glass

And from the sublime to the ridiculous, here is a certain phrase[1] in an assortment of languages:

1. **Sanskrit**: काचं शक्नोम्यत्तुम् । नोपहिनस्ति माम् ॥
2. **Sanskrit** *(standard transcription)*: kācaṃ śaknomyattum; nopahinasti mām.
3. **Classical Greek**: ὕαλον ϕαγεῖν δύναμαι· τοῦτο οὔ με βλάπτει.
4. **Greek** (monotonic): Μπορώ να φάω σπασμένα γυαλιά χωρίς να πάθω τίποτα.
5. **Greek** (polytonic): Μπορῶ νὰ φάω σπασμένα γυαλιὰ χωρὶς νὰ πάθω τίποτα.

37. **Manx Gaelic**: Foddym gee glonney agh cha jean eh gortaghey mee.
38. **Old Irish** *(Ogham)*: ᚛᚛ᚉᚑᚅᚔᚉᚉᚔᚋ ᚔᚈᚔ ᚍᚂᚐᚅᚑ ᚅᚔᚋᚌᚓᚅᚐ᚜
39. **Old Irish** *(Latin)*: Con·iccim ithi nglano. Ním·géna.
40. **Irish**: Is féidir liom gloinne a ithe. Ní dhéanann sí dochar ar bith dom.
41. **Ulster Gaelic**: Ithim-sa gloine agus ní miste damh é.
42. **Scottish Gaelic**: S urrainn dhomh gloinne ithe; cha ghoirtich i mi.
43. **Anglo-Saxon** *(Runes)*: ᛁᚳ᛫ᛗᚨᚷ᛫ᚷᛚᚨᛋ᛫ᛖᚩᛏᚪᚾ᛫ᚩᚾᛞ᛫ᚻᛁᛏ᛫ᚾᛖ᛫ᚻᛖᚪᚱᛗᛁᚪᚧ᛫ᛗᛖ᛬
44. **Anglo-Saxon** *(Latin)*: Ic mæg glæs eotan ond hit ne hearmiað me.
45. **Middle English**: Ich canne glas eten and hit hirtiþ me nouȝt.
46. **English**: I can eat glass and it doesn't hurt me.
47. **English** *(IPA)*: [aɪ kæn iːt glɑːs ænd ɪt dɐz nɒt hɜːt miː] (Received Pronunciation)
48. **English** *(Braille)*: ⠊⠀⠉⠁⠝⠀⠑⠁⠞⠀⠛⠇⠁⠎⠎⠀⠁⠝⠙⠀⠊⠞⠀⠙⠕⠑⠎⠝⠞⠀⠓⠥⠗⠞⠀⠍⠑
49. **Jamaican**: Mi kian niam glas han i neba hot mi.
50. **Lalland Scots / Doric**: Ah can eat gless, it disnae hurt us.
51. **Gothic** (4): 𐌼𐌰𐌲 𐌲𐌻𐌴𐍃 𐌹̈𐍄𐌰𐌽, 𐌽𐌹 𐌼𐌹𐍃 𐍅𐌿 𐌽𐌳𐌰𐌽 𐌱𐍂𐌹𐌲𐌲𐌹𐌸.
52. **Old Norse** *(Runes)*: ᛖᚴ ᚷᛖᛏ ᛖᛏᛁ ᚧ ᚷᛚᛖᚱ ᛘᚾ ᚦᛖᛋᛋ ᚨᚧ ᚡᛖ ᚱᚧᚨ ᛋᚨᚱ
53. **Old Norse** *(Latin)*: Ek get etið gler án þess að verða sár.

91. **Slovenian:** Lahko jem steklo, ne da bi mi škodovalo.
92. **Bosnian, Croatian, Montenegrin and Serbian** *(Latin)*: Ja mogu jesti staklo, i to mi ne šteti.
93. **Bosnian, Montenegrin and Serbian** *(Cyrillic)*: Ја могу јести стакло, и то ми не штети.
94. **Macedonian:** Можам да јадам стакло, а не ме штета.
95. **Russian**: Я могу есть стекло, оно мне не вредит.
96. **Belarusian** *(Cyrillic)*: Я магу есці шкло, яно мне не шкодзіць.
97. **Belarusian** *(Lacinka)*: Ja mahu jeści škło, jano mne ne škodzić.
98. **Ukrainian**: Я можу їсти скло, і воно мені не зашкодить.
99. **Bulgarian**: Мога да ям стъкло, то не ми вреди.
100. **Georgian**: მინას ვჭამ და არა მტკივა.
101. **Armenian**: Կրնամ ապակի ուտել և ինծի անհանգիստ չըներ։
102. **Albanian**: Unë mund të ha qelq dhe nuk më gjen gjë.
103. **Turkish**: Cam yiyebilirim, bana zararı dokunmaz.
104. **Turkish** *(Ottoman)*: جام ييه بلورم بڭا ضررى طوقونمز
105. **Tatar**: Алам да бар, пыяла, әмма бу ранит мине.
106. **Uzbek / O'zbekcha:** *(Roman)*: Men shisha yeyishim mumkin, ammo u menga zarar keltirmaydi.
107. **Uzbek / Ўзбекча** *(Cyrillic)*: Мен шиша ейишим мумкин, аммо у менга зарар келтирмайди.
108. **Bangla / Bengali**: আমি কাঁচ খেতে পারি, তাতে আমার কোনো ক্ষতি হয় না।
109. **Marathi** (masculine): मी काच खाऊ शकतो, मला ते दुखत नाही.
110. **Marathi** (feminine): मी काच खाऊ शकते, मला ते दुखत नाही.
111. **Kannada**: ನನಗೆ ಹಾನಿ ಆಗದೆ, ನಾನು ಗಜನ್ನು ತಿನಬಹುದು
112. **Hindi** (masculine): मैं काँच खा सकता हूँ और मुझे उससे कोई चोट नहीं पहुंचती.
113. **Hindi** (feminine): मैं काँच खा सकती हूँ और मुझे उससे कोई चोट नहीं पहुंचती.
114. **Malayalam**: എനിക്ക് ഗ്ലാസ് തിന്നാം. അതെന്നെ വേദനിപ്പിക്കില്ല.
115. **Tamil**: நான் கண்ணாடி சாப்பிடுவேன், அதனால் எனக்கு ஒரு கேடும் வராது.
116. **Telugu**: నేను గాజు తినగలను మరియు అలా చేసినా నాకు ఏమి ఇబ్బంది లేదు
117. **Sinhalese**: මට වීදුරු කෑමට හැකියි. එයින් මට කිසි හානියක් සිදු නොවේ.
118. **Urdu**(3): میں کانچ کھا سکتا ہوں اور مجھے تکلیف نہیں ہوتی ۔
119. **Pashto**(3): زه شيشه خوړلې شم، هغه ما نه خوږوي
120. **Farsi / Persian**(3): .من می توانم بدونِ احساس درد شيشه بخورم
121. **Arabic**(3): أنا قادر على أكل الزجاج و هذا لا يؤلمني.
122. **Maltese**: Nista' niekol il-ħġieġ u ma jagħmilli xejn.
123. **Hebrew**(3): אני יכול לאכול זכוכית וזה לא מזיק לי.
124. **Yiddish**(3): איך קען עסן גלאָז און עס טוט מיר נישט װײ.

139. **Burmese** (Unicode 4.0): က္ယ္ဝန္‌တော္‌၊က္ယ္ဝန္‌မ မ္ယက္‌စားနုိင္‌သည္‌။ ၎က္ရောင္‌့ ထိခုိက္‌မ္ဟု မရ္ဟိပာ။ (9)
140. **Burmese** (Unicode 5.0): ကျွန်တော် ကျွန်မ မှန်စားနိုင်တယ်။ ၎င်းကြောင့် ထိခိုက်မှုမရှိပါ။ (9)
141. **Vietnamese (quốc ngữ)**: Tôi có thể ăn thủy tinh mà không hại gì.
142. **Vietnamese (nôm)** (4): 些 𣎏 世 咹 水 晶 𦓡 空 𣎏 害 咦
143. **Khmer**: ខ្ញុំអាចញុំកញ្ចក់បាន ដោយគ្មានបញ្ហារ
144. **Lao**: ຂອ້ຍກິນແກ້ວໄດ້ໂດຍທີ່ມັນບໍ່ໄດ້ເຮັດໃຫ້ຂອ້ຍເຈັບ.
145. **Thai**: ฉันกินกระจกได้ แต่มันไม่ทำให้ฉันเจ็บ
146. **Mongolian** *(Cyrillic)*: Би шил идэй чадна, надад хортой биш
147. **Mongolian** *(Classic)* (5): ᠪᠢ ᠰᠢᠯᠢ ᠢᠳᠡᠶᠦ ᠴᠢᠳᠠᠨᠠ ᠂ ᠨᠠᠳᠤᠷ ᠬᠣᠤᠷᠠᠳᠠᠢ ᠪᠢᠰᠢ
148. **Nepali**: म काँच खान सक्छू र मलाई केहि नी हुन्‍न् ।
149. **Tibetan**: ཤེལ་སྒོ་ཟ་ནས་ང་ན་གི་མ་རེད།
150. **Chinese**: 我能吞下玻璃而不伤身体。
151. **Chinese** (Traditional): 我能吞下玻璃而不傷身體。
152. **Taiwanese**(6): Góa ē-tàng chia̍h po-lê, mā bē tio̍h-siong.
153. **Japanese**: 私はガラスを食べられます。それは私を傷つけません。
154. **Korean**: 나는 유리를 먹을 수 있어요. 그래도 아프지 않아요
155. **Bislama**: Mi save kakae glas, hemi no save katem mi.
156. **Hawaiian**: Hiki iaʻu ke ʻai i ke aniani; ʻaʻole nō lā au e ʻeha.
157. **Marquesan**: E koʻana e kai i te karahi, mea ʻā, ʻaʻe hauhau.
158. **Inuktitut** (10): ᐊᓕᒍᖅ ᓂᕆᔭᕌᖓᒃᑯ ᓱᕋᙱᑦᑐᓐᓇᖅᑐᖓ

Unicode: Beispieltabellen

Latin-Zeichen

- U+0000 bis U+007F: ASCII www.unicode.org/charts/PDF/U0000.pdf
- U+0100 bis U+017F: Latin-A www.unicode.org/charts/PDF/U0100.pdf
  Europäische Umlaute und Sonderzeichen
- ab U+0180 weitere Sonderzeichen: Latin-B, Latin-C usw.

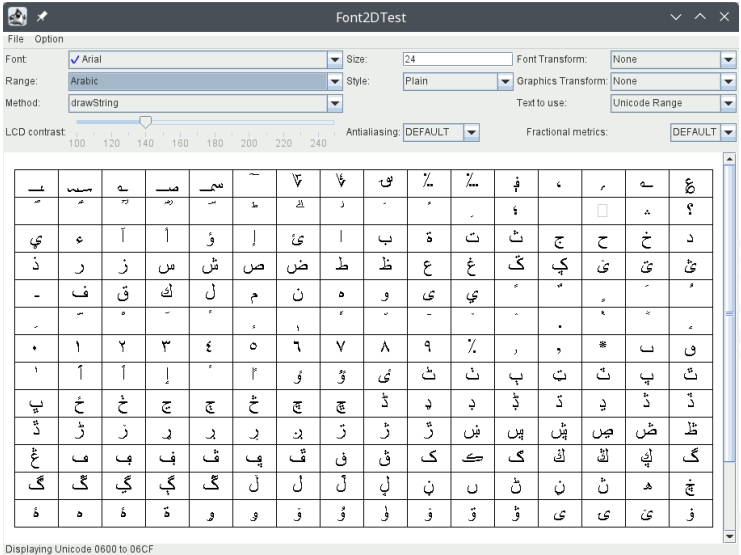
Symbole und Operatoren

- 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

Asiatische Sprachen: Chinesisch (traditional/simplified), Japanisch, Koreanisch

- U+3400 bis U+4DBF www.unicode.org/charts/PDF/U3400.pdf
- U+4E00 bis U+9FFF www.unicode.org/charts/PDF/U4E00.pdf

Unicode: Java2D Fontviewer



Oracle [JavaD]: JDK demos and samples
.../demo/jfc/Font2DTest

Unicode: Repräsentation?

- 16-bit für jedes Zeichen, bis zu 65 536 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 3.0 mehrere *Planes* zu je 65 536 Zeichen
- direkte Repräsentation aller Zeichen erfordert 32-bit/Zeichen
- vierfacher Speicherbedarf gegenüber ISO-8859-1

- bei Dateien ist möglichst kleine Dateigröße wichtig

$\Rightarrow$ Codierung als UTF-8 oder UTF-16

Zeichen	Unicode	Unicode binär	UTF-8 binär	UTF-8 hexadezimal
Buchstabe *y*	U+0079	00000000 01111001	01111001	79
Buchstabe *ä*	U+00E4	00000000 11100100	11000011 10100100	C3 A4
Zeichen für eingetragene Marke ®	U+00AE	00000000 10101110	11000010 10101110	C2 AE
Eurozeichen €	U+20AC	00100000 10101100	11100010 10000010 10101100	E2 82 AC
Violinschlüssel 𝄞	U+1D11E	00000001 11010001 00011110	11110000 10011101 10000100 10011110	F0 9D 84 9E

de.wikipedia.org/wiki/UTF-8

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

UTF-8: Codierung

Unicode-Bereich (hexadezimal)	UTF-Codierung (binär)	Anzahl (benutzt)
0000 0000 - 0000 007F	0*** ****	128
0000 0080 - 0000 07FF	110* **** 10** ****	1 920
0000 0800 - 0000 FFFF	1110 **** 10** **** 10** ****	63 488
0001 0000 - 0010 FFFF	1111 0*** 10** **** 10** **** 10** ****	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 Anz. Bytegruppen an
- Zeichen ergibt sich als Bitstring aus den $***\ldots*$
- 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 (Befehl: `locale`)
  Java: `java.util.Locale`
  Windows: Einstellung über System/Registry-Schlüssel

Textkonvertierungen

- `dos2unix`, `unix2dos`: Umwandeln von Dateien (z.B. Programm-Quelltexte) zwischen DOS/Windows und Unix/Linux: Codierung und Zeilenenden

`dos2unix -h`	Optionen anzeigen / Hilfe
`dos2unix -ascii -n a.txt b.txt`	nur Umbrüche (von `a.txt` nach `b.txt`)
`dos2unix -iso -n a.txt b.txt`	Umbrüche und ISO-8851-1
`unix2dos -850 -n a.txt b.txt`	Umbrüche und Codepage 850

- `iconv`: „Universalwerkzeug“ zur Umwandlung von Textcodierungen

`iconv -l`	Liste der unterstützten Codierungen
`iconv -f` ⟨*encoding*⟩ `...`	Codierung der Eingabedatei
`iconv -t` ⟨*encoding*⟩ `...`	Codierung der Ausgabedatei
`iconv -o` ⟨*filename*⟩ `...`	Name der Ausgabedatei

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

- Konvertierungsfunktionen in den meisten Texteditoren enthalten!

Base64-Codierung

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

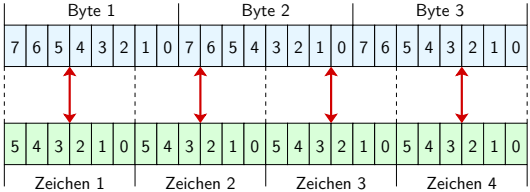
$\Rightarrow$ Verfahren zur Umcodierung der Datei in 7-bit ASCII notwendig

$\Rightarrow$ etabliert ist das **Base64** Verfahren (RFC 2045)

- alle E-Mail Dateianhänge und 8-bit Textdateien
- Umcodierung benutzt nur Buchstaben, Ziffern und drei Sonderzeichen
- Daten werden byteweise in ASCII Symbole umgesetzt

Base64-Codierung: Prinzip

1. Codierung von drei Bytes als vier 6-bit Zeichen



- $3 \times 8\text{-bit} \Leftrightarrow 4 \times 6\text{-bit}$
- 6-bit Binärwerte: $0 \ldots 63$
- nutzen 64 (von 128) 7-bit ASCII Symbolen

2.

Zahl	ASCII	Zuordnung der ASCII-Zeichen
$0 \ldots 25$	A...Z	
$26 \ldots 51$	a...z	
$52 \ldots 61$	0...9	
62	+	
63	/	
	=	Füllzeichen, falls Anz. Bytes nicht durch 3 teilbar
	CR	Zeilenumbruch (opt.), meistens nach 76 Zeichen

Base64-Codierung: Prinzip (cont.)

Text content	**M**								**a**								**n**							
ASCII	77								97								110							
Bit pattern	0	1	0	0	1	1	0	1	0	1	1	0	0	0	0	1	0	1	1	0	1	1	1	0
Index	19								22								5							
Base64-encoded	**T**								**W**								**F**							

- drei 8-bit Zeichen, neu 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-Codierung: Java

- in neueren Java Versionen direkt im JDK enthalten

  Module `java.base`, Package `java.util`: `Base64Encoder`, bzw. `Base64Decoder`

- diverse andere Packages
  - Apache Commons Codec (`commons.apache.org/proper/commons-codec`)

    `org.apache.commons.codec.binary.Base64InputStream`
    `org.apache.commons.codec.binary.Base64OutputStream`
  - JAXB (Java Architecture for XML Binding) in `javax.xml.bind.DatatypeConverter`
    `parseBase64Binary`, `printBase64Binary`
  - ...

Literatur

[Uni] The Unicode Consortium; Mountain View, CA.
`home.unicode.org`, `unicode.org/main.html`

[JavaI] Oracle Corporation: *The Java Tutorials – Trail: Internationalization*.
`docs.oracle.com/javase/tutorial/i18n`

[JavaD] Oracle Corporation: *Java SE Downloads*.
`www.oracle.com/java/technologies/downloads`

[Ull23] C. Ullenboom: *Java ist auch eine Insel – Einführung, Ausbildung, Praxis*.
17. Auflage, Rheinwerk Verlag GmbH, 2023. ISBN 978–3–8362–9544–4

16. Auflage Online verfügbar: `openbook.rheinwerk-verlag.de/javainsel`