



Was Informatiker:innen über Unicode wissen sollten

Anton Ehrmanntraut – 08.06.2022

Wie Texte binär
übertragen/abspeichern?

H e l l o W o r l d !

Wie Texte binär übertragen/abspeichern?

H	e	l	l	o		W	o	r	l	d	!
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
48	65	6c	6c	6f	20	57	6f	72	6c	64	21

Wie
über

H



48

ASCII Codepage

00	Null	20	space	40	@	60	`
01	Start of Header	21	!	41	A	61	a
02	Start of Text	22	"	42	B	62	b
03	End of Text	23	#	43	C	63	c
04	End of Transmission	24	\$	44	D	64	d
05	Enquiry	25	%	45	E	65	e
06	Acknowledge	26	&	46	F	66	f
07	Bell	27	'	47	G	67	g
08	Backspace	28	(	48	H	68	h
09	Horizontal Tab	29	)	49	I	69	i
0A	Line Feed	2A	*	4A	J	6A	j
0B	Vertical Tab	2B	+	4B	K	6B	k
0C	Form Feed	2C	,	4C	L	6C	l
0D	Carriage Return	2D	-	4D	M	6D	m
0E	Shift Out	2E	.	4E	N	6E	n
0F	Shift In	2F	/	4F	O	6F	o
10	Data Link Escape	30	0	50	P	70	p
11	Device Control 1	31	1	51	Q	71	q
12	Device Control 2	32	2	52	R	72	r
13	Device Control 3	33	3	53	S	73	s
14	Device Control 4	34	4	54	T	74	t
15	Negative Acknowledge	35	5	55	U	75	u
16	Synchronize	36	6	56	V	76	v
17	End of Transmission Block	37	7	57	W	77	w
18	Cancel	38	8	58	X	78	x
19	End of Medium	39	9	59	Y	79	y
1A	Substitute	3A	:	5A	Z	7A	z
1B	Escape	3B	;	5B	[	7B	{
1C	File Separator	3C	<	5C	\	7C	
1D	Group Separator	3D	=	5D	]	7D	}
1E	Record Separator	3E	>	5E	^	7E	~
1F	Unit Separator	3F	?	5F	_	7F	DEL

!



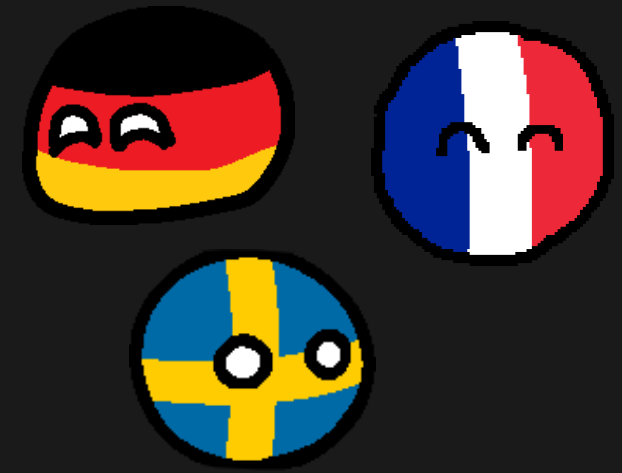
21

Wie Texte binär übertragen/abspeichern?

H	e	l	l	o		W	o	r	l	d	!
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
48	65	6c	6c	6f	20	57	6f	72	6c	64	21

in ASCII: 1 Codepoint = 1 Zeichen = 1 Byte

Codepages!



Codepage 819 - Latin 1 - ISO 8859-1

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F	
0-		0001	0002	0003	0004	0005	0006	0007	0008	0009	000A	000B	000C	000D	000E	000F	
1-		0010	0011	0012	0013	0014	0015	0016	0017	0018	0019	001A	001B	001C	001D	001E	001F
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-																	
9-																	
A-		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	®	¯		
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	

Codepages!



ASCII

neu!

Codepage 819 - Latin 1 - ISO 8859-1

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F	
0-		0001	0002	0003	0004	0005	0006	0007	0008	0009	000A	000B	000C	000D	000E	000F	
1-		0010	0011	0012	0013	0014	0015	0016	0017	0018	0019	001A	001B	001C	001D	001E	001F
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-																	
9-																	
A-		;	€	£	¤	¥	¦	§	¨	©	ª	«	¬	®	¯		
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	

Codepages!



ASCII

neu!

Codepage 819 - Latin 1 - ISO 8859-1

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-		0001	0002	0003	0004	0005	0006	0007	0008	0009	000A	000B	000C	000D	000E	000F
1-	0010	0011	0012	0013	0014	0015	0016	0017	0018	0019	001A	001B	001C	001D	001E	001F
2-	0020	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3-	0030	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>
4-	0040	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N
5-	0050	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^
6-	0060	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n
7-	0070	p	q	r	s	t	u	v	w	x	y	z	{		}	~
8-	0080															
9-	0090															
A-	00A0	;	€	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯
B-	00B0	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾
C-	00C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î
D-	00D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ
E-	00E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î
F-	00F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ

Re: New Harry Potter Book

От: Svetlana Pletnava <sveta@example.ru>

Кому: laura@example.fr

Тема: Re: New Harry Potter Book

Hey Laura,
that would be amazing - we cannot get new
Harry Potter book here in Russia yet! Here is
my address; it is in Russian so please copy it
carefully:

Россия Москва, 119415
пр.Вернадского, 37,
к.1817-1,
Плетневой Светлане

Thanks! Sveta

Слать

Re: New Harry Potter Book

От: Svetlana Pletnava <sveta@example.ru>

Кому: laura@example.fr

Тема: Re: New Harry Potter Book

Hey Laura,
that would be amazing - we cannot get new
Harry Potter book here in Russia yet! Here is
my address; it is in Russian so please copy it
carefully:

Россия Москва, 119415
пр.Вернадского, 37,
к.1817-1,
Плетневой Светлане

Thanks! Sveta

Re: New Harry Potter Book

De: Svetlana Pletnava <sveta@example.ru>

À: laura@example.fr

Sujet: Re: New Harry Potter Book

Hey Laura,
that would be amazing - we cannot get new
Harry Potter book here in Russia yet! Here is
my address; it is in Russian so please copy it
carefully:

òĭÓÓÉÑ ĭĭÓË×Á, 119415
ĐÒ.÷ÅÒĬÁÄÓËĬĬ, 37,
Ë.1817-1,
ðĬÄÔĬÄ×ÏË ó×ÅÒĬÁĬÄ

Thanks! Sveta

Re: New Harry Potter Book

От: Svetlana Pletnava <sveta@example.ru>

Кому: laura@example.fr

Тема: Re: New Harry Potter Book

Hey Laura,
that would be amazing - we could have the
Harry Potter book here in Russia. I'll send you
my address; it is in Russian. Please read it
carefully:

Россия Москва, 119415
пр.Вернадского, 37,
к.1817-1,
Плетневой Светлане

Thanks! Sveta

Re: New Harry Potter Book

De: Svetlana Pletnava <sveta@example.ru>

À: laura@example.fr

Sujet: Re: New Harry Potter Book

Hey Laura,
that would be amazing - we could have the
Harry Potter book here in Russia. I'll send you
my address; it is in Russian. Please read it
carefully:

Россия Москва, 119415
пр.Вернадского, 37,
к.1817-1,
Плетневой Светлане

Thanks! Sveta

KOI8-R	P	o	c	c	и	я
bytes	F2	Cf	D3	D3	C9	D2
ISO-8859-1	ò	ï	ó	ó	é	ñ

Re: New Harry Potter Book

От: Svetlana Pletnava <sveta@example.ru>

Кому: laura@example.fr

Тема: Re: New Harry Potter B

Hey Laura,
that woul
Harry Pot
my addre
carefully:

Россия М
пр.Верна
к.1817-1,
Плетнево

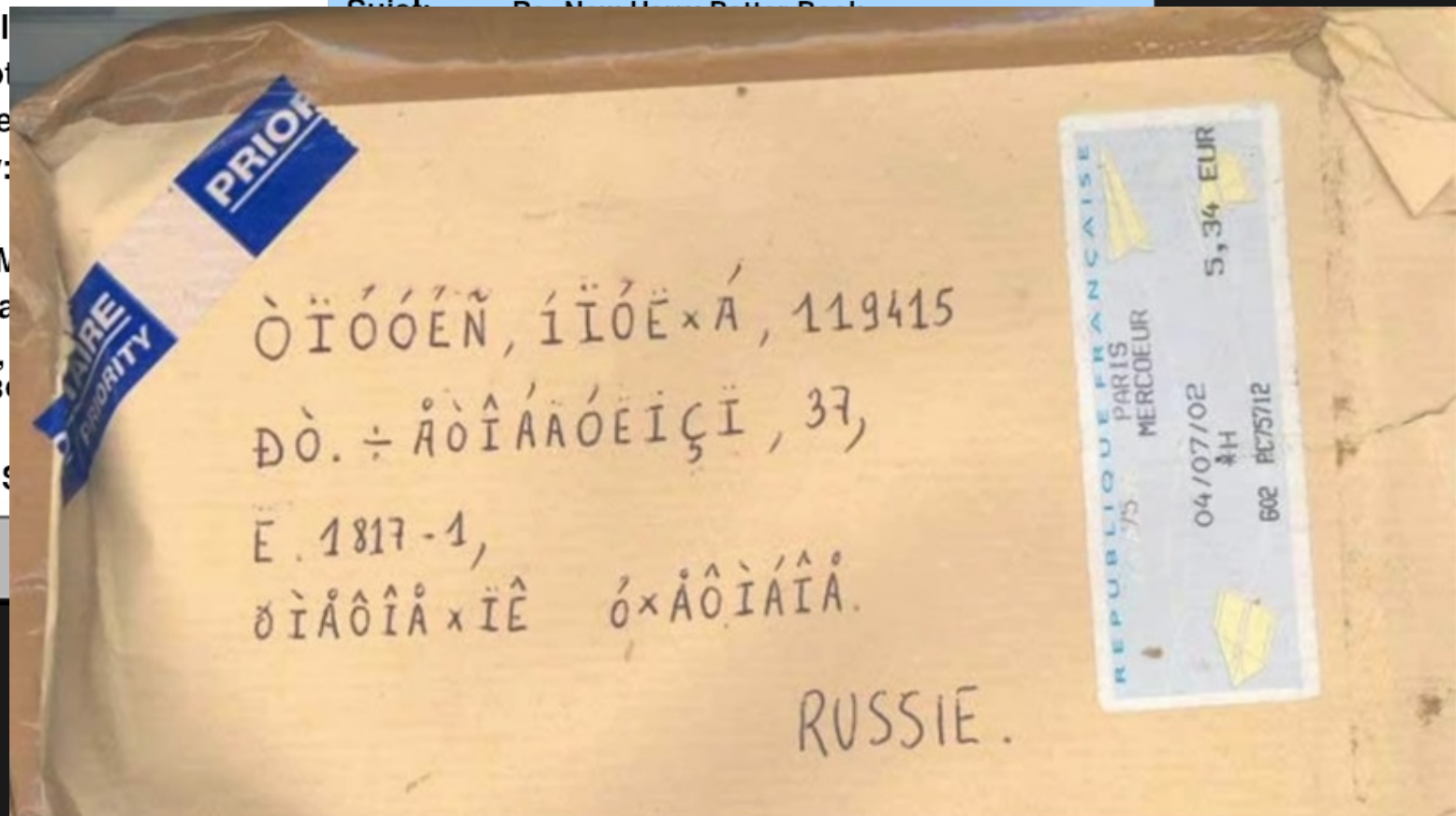
Thanks! S

Re: New Harry Potter Book

De: Svetlana Pletnava <sveta@example.ru>

À: laura@example.fr

Subject: Re: New Harry Potter Book



Re: New Harry Potter Book

От: Svetlana Pletnava <sveta@example.ru>

Кому: laura@example.fr

Тема: Re: New Harry Potter Book

Hey Laura,
that would be great!
Harry Potter book is on my address
carefully!

Россия Москва
пр.Вернадского
к.1817-1,
Плетнев

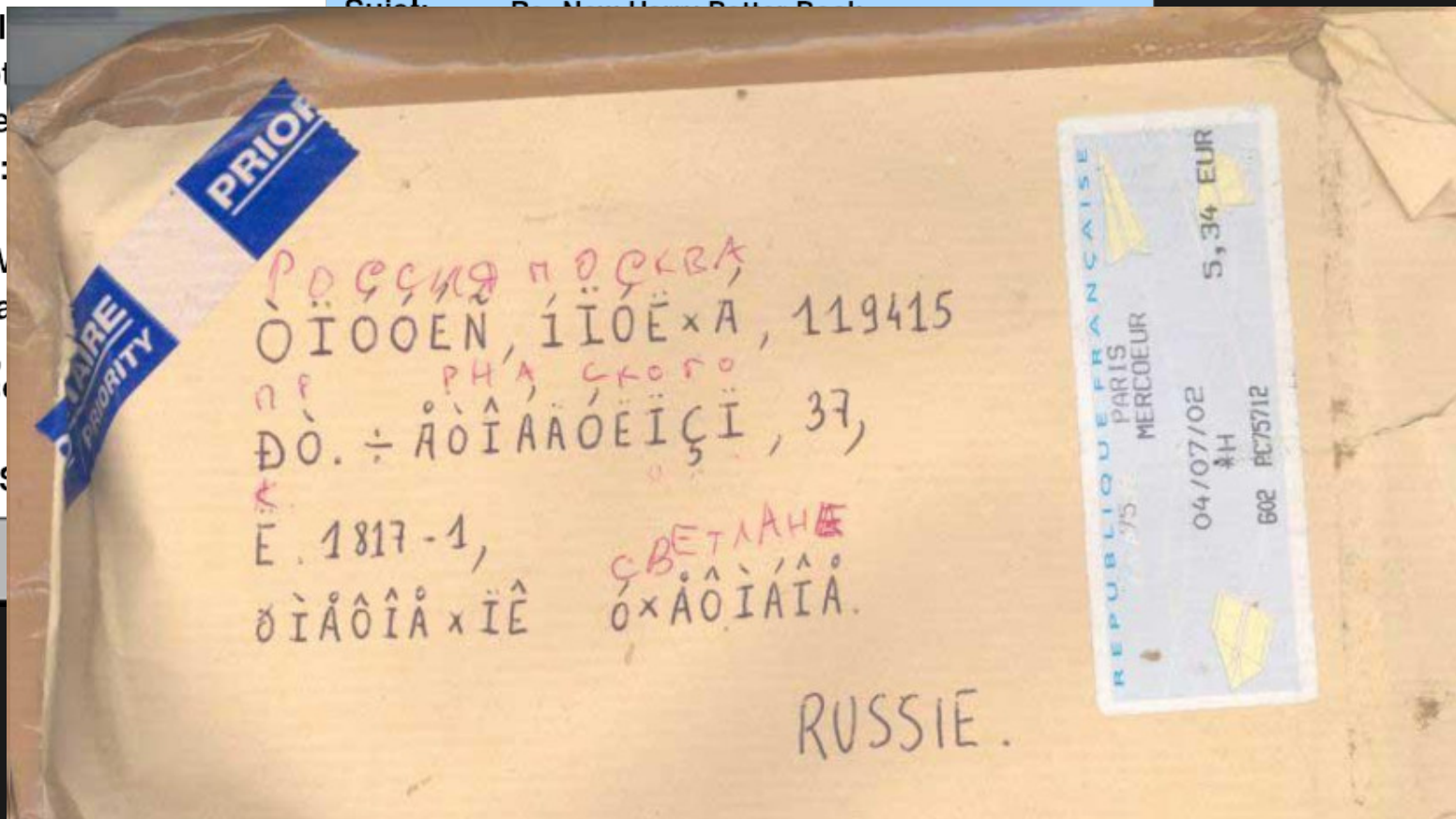
Thanks! S

Re: New Harry Potter Book

De: Svetlana Pletnava <sveta@example.ru>

À: laura@example.fr

Subject: Re: New Harry Potter Book



```
\usepackage[latin1]{inputenc}
```

```
<html><head>  
...  
<meta charset="iso-8851-1">
```

Keine Angabe der Codierung
⇒ kein Plain-Text!

```
Received: from mailmaster.uni-wuerzburg.de
```

```
...  
From: ...
```

```
Content-Type: text/plain; charset=Windows-1252
```

```
HTTP 200 OK
```

```
...
```

```
Content-Type: application/javascript;  
charset=utf-8
```

WELCOME TO



imgflip.com

Uni
UNICODE

Safari-Fahrplan

- Basics
- Codierung: UTF-8
- Codepoint vs „Zeichen“
- Normalisierungen
- Regex & Zeichenklassen
- *honorable Mentions*
- Unicode und Gesellschaft

Basics

Idee: ordne jedem ,atomaren Schriftzeichen'
eine Zahl zu → **Codepoint**

Basics

Idee: ordne jedem ,atomaren Schriftzeichen'
eine Zahl zu → **Codepoint**

a	→	U+0041	Latin Small Letter A
ü	→	U+00FC	Latin Small Letter U with Diaeresis
Й	→	U+0419	Cyrillic Capital Letter Short I
ओ	→	U+0913	Devanagari Letter O
我	→	U+6211	CJK Unified Ideograph-6211
🙄	→	U+1F928	Face with Raised Eyebrow

Basics

Idee: ordne jedem ,atomaren Schriftzeichen'
eine Zahl zu → **Codepoint**

a	→	U+0041	Latin Small Letter A
ü	→	U+00FC	Latin Small Letter U with Diaeresis
Й	→	U+0419	Cyrillic Capital Letter Short I
ओ	→	U+0913	Devanagari Letter 0
我	→	U+6211	CJK Unified Ideograph-6211
🙄	→	U+1F928	Face with Raised Eyebrow

144k Codepoints definiert
in U+000000 bis U+10FFFF

Basic Multilingual Plane

U+0000 bis U+FFFF

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	Latin script Non-Latin European scripts Middle Eastern and Southwest Asian scripts South and Central Asian scripts African scripts Southeast Asian scripts East Asian scripts American scripts Indonesian and Oceanic scripts Symbols Notational systems CJK characters UTF-16 surrogates Private use As of Unicode 14.0
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F	
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F	
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F	
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F	
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F	
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F	
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F	
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F	
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F	
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF	
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF	
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF	
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF	
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF	

Basic Multilingual Plane U+0000 bis U+FFFF

jede Box je 256 Zeichen, hier U+20xx

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	Latin script Non-Latin European scripts Middle Eastern and Southwest Asian scripts South and Central Asian scripts African scripts Southeast Asian scripts East Asian scripts American scripts Indonesian and Oceanic scripts Symbols Notational systems CJK characters UTF-16 surrogates Private use As of Unicode 14.0
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F	
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F	
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F	
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F	
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F	
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F	
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F	
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F	
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F	
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF	
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF	
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF	
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF	
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF	

Basic Multilingual Plane U+0000 bis U+FFFF

jede Box je 256 Zeichen, hier U+20xx

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

Latin script
 Non-Latin European scripts
 Middle Eastern and Southwest Asian scripts
 South and Central Asian scripts
 African scripts
 Southeast Asian scripts
 East Asian scripts
 American scripts
 Indonesian and Oceanic scripts
 Symbols
 Notational systems
 CJK characters
 UTF-16 surrogates
 Private use

As of Unicode 14.0

Unicode blocks and contained scripts [hide]					
Plane ↕	Block range ↕	Block name ↕	Code points ^[a] ↕	Assigned characters ↕	Scripts ^{[b][c][d][e][f]} ↕
0 BMP	U+0000..U+007F	Basic Latin ^[g]	128	128	Latin (52 characters), Common (76 characters)
0 BMP	U+0080..U+00FF	Latin-1 Supplement ^[h]	128	128	Latin (64 characters), Common (64 characters)
0 BMP	U+0100..U+017F	Latin Extended-A	128	128	Latin
0 BMP	U+0180..U+024F	Latin Extended-B	208	208	Latin
0 BMP	U+0250..U+02AF	IPA Extensions	96	96	Latin
0 BMP	U+02B0..U+02FF	Spacing Modifier Letters	80	80	Bopomofo (2 characters), Latin (14 characters), Common (64 characters)
0 BMP	U+0300..U+036F	Combining Diacritical Marks	112	112	Inherited
0 BMP	U+0370..U+03FF	Greek and Coptic	144	135	Coptic (14 characters), Greek (117 characters), Common (4 characters)
0 BMP	U+0400..U+04FF	Cyrillic	256	256	Cyrillic (254 characters), Inherited (2 characters)
0 BMP	U+0500..U+052F	Cyrillic Supplement	48	48	Cyrillic
0 BMP	U+0530..U+058F	Armenian	96	91	Armenian
0 BMP	U+0590..U+05FF	Hebrew	112	88	Hebrew
0 BMP	U+0600..U+06FF	Arabic	256	256	Arabic (238 characters), Common (6 characters), Inherited (12 characters)
0 BMP	U+0700..U+074F	Syriac	80	77	Syriac
0 BMP	U+0750..U+077F	Arabic Supplement	48	48	Arabic
0 BMP	U+0780..U+07FF	Thaana	64	50	Thaana

Basic Multilingual Plane

U+0000 bis U+FFFF

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

	Latin script
	Non-Latin European scripts
	Middle Eastern and Southwest Asian scripts
	South and Central Asian scripts
	African scripts
	Southeast Asian scripts
	East Asian scripts
	American scripts
	Indonesian and Oceanic scripts
	Symbols
	Notational systems
	CJK characters
	UTF-16 surrogates
	Private use

As of Unicode 14.0



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Current events](#)
[Random article](#)
[About Wikipedia](#)
[Contact us](#)
[Donate](#)

[Contribute](#)

[Help](#)
[Learn to edit](#)
[Community portal](#)
[Recent changes](#)
[Upload file](#)

[Tools](#)

[What links here](#)
[Related changes](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Cite this page](#)
[Wikidata item](#)

[Print/export](#)

[Download as PDF](#)
[Printable version](#)

[In other projects](#)

Article [Talk](#)

Georgian (Unicode block)

From Wikipedia, the free encyclopedia

See also: *Georgian Supplement (Unicode block)* and *Georgian*

Georgian is a **Unicode block** containing the **Mkhedruli** and **Asomtavruli** Georgian characters used to write Modern Georgian, Svan, and Mingrelian languages. Another lower case, **Nuskhuri**, is encoded in a separate **Georgian Supplement** block, which is used with the Asomtavruli to write the ecclesiastical **Khutsuri** Georgian script.

Block [\[edit \]](#)

Georgian ^{[1][2]}																
Official Unicode Consortium code chart  (PDF)																
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+10Ax	Ⴀ	Ⴁ	Ⴂ	Ⴃ	Ⴄ	Ⴅ	Ⴆ	Ⴇ	Ⴈ	Ⴉ	Ⴊ	Ⴋ	Ⴌ	Ⴍ	Ⴎ	Ⴏ
U+10Bx	Ⴐ	Ⴑ	Ⴒ	Ⴓ	Ⴔ	Ⴕ	Ⴖ	Ⴗ	Ⴘ	Ⴙ	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ
U+10Cx	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ		Ⴚ						Ⴚ		
U+10Dx	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ	Ⴚ	Ⴛ	Ⴜ	Ⴝ
U+10Ex	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ	Ⴚ	Ⴛ	Ⴜ	Ⴝ
U+10Fx	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ	Ⴚ	Ⴛ	Ⴜ	Ⴝ
Notes																
1.^ As of Unicode version 14.0																
2.^ Grey areas indicate non-assigned code points																

Georgian

Range	U+10A0..U+10FF (96 code points)
Plane	BMP
Scripts	Georgian (87 char.) Common (1 char.)
Major alphabets	Mkhedruli Asomtavruli
Assigned	88 code points
Unused	8 reserved code points

Unicode version history

1.0.0 (1991)	78 (+78)
3.2 (2002)	80 (+2)
4.1 (2005)	83 (+3)
6.1 (2012)	88 (+5)

Chart

[Code chart](#) ^(PDF)

Note: ^{[1][2]}

Basic Multilingual Plane U+0000 bis U+FFFF

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

- Latin script
 - Non-Latin European scripts
 - Middle Eastern and Southwest Asian scripts
 - South and Central Asian scripts
 - African scripts
 - Southeast Asian scripts
 - East Asian scripts
 - American scripts
 - Indonesian and Oceanic scripts
 - Symbols
 - Notational systems
 - CJK characters
 - UTF-16 surrogates
 - Private use
- As of Unicode 14.0



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Current events](#)
[Random article](#)
[About Wikipedia](#)
[Contact us](#)
[Donate](#)

[Contribute](#)

[Help](#)
[Learn to edit](#)
[Community portal](#)
[Recent changes](#)
[Upload file](#)

[Tools](#)

[What links here](#)
[Related changes](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Cite this page](#)
[Wikidata item](#)

[Print/export](#)

[Download as PDF](#)
[Printable version](#)

[In other projects](#)

Article [Talk](#)

Georgian (Unicode block)

From Wikipedia, the free encyclopedia

See also: [Georgian Supplement \(Unicode block\)](#) and [Georgian Supplement](#)

Georgian is a **Unicode block** containing the **Mkhedruli** and **Asomtavruli** Georgian characters used to write Modern Georgian, Svan, and Mingrelian languages. Another lower case, **Nuskhuri**, is encoded in a separate **Georgian Supplement** block, which is used with the Asomtavruli to write the ecclesiastical **Khutsuri** Georgian script.

Block [\[edit \]](#)

Georgian ^{[1][2]}																
Official Unicode Consortium code chart (PDF)																
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+10Ax	Ⴀ	Ⴁ	Ⴂ	Ⴃ	Ⴄ	Ⴅ	Ⴆ	Ⴇ	Ⴈ	Ⴉ	Ⴊ	Ⴋ	Ⴌ	Ⴍ	Ⴎ	Ⴏ
U+10Bx	Ⴐ	Ⴑ	Ⴒ	Ⴓ	Ⴔ	Ⴕ	Ⴖ	Ⴗ	Ⴘ	Ⴙ	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ
U+10Cx	Ⴀ	Ⴁ	Ⴂ	Ⴃ	Ⴄ	Ⴅ		Ⴈ						Ⴍ		
U+10Dx	Ⴐ	Ⴑ	Ⴒ	Ⴓ	Ⴔ	Ⴕ	Ⴖ	Ⴗ	Ⴘ	Ⴙ	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ
U+10Ex	Ⴀ	Ⴁ	Ⴂ	Ⴃ	Ⴄ	Ⴅ	Ⴆ	Ⴇ	Ⴈ	Ⴉ	Ⴊ	Ⴋ	Ⴌ	Ⴍ	Ⴎ	Ⴏ
U+10Fx	Ⴐ	Ⴑ	Ⴒ	Ⴓ	Ⴔ	Ⴕ	Ⴖ	Ⴗ	Ⴘ	Ⴙ	Ⴚ	Ⴛ	Ⴜ	Ⴝ	Ⴞ	Ⴟ

Notes

- [^] As of Unicode version 14.0
- [^] Grey areas indicate non-assigned code points

Georgian

Range	U+10A0..U+10FF (96 code points)
Plane	BMP
Scripts	Georgian (87 char.) Common (1 char.)
Major alphabets	Mkhedruli Asomtavruli
Assigned	88 code points
Unused	8 reserved code points

Unicode version history

1.0.0 (1991)	78 (+78)
3.2 (2002)	80 (+2)
4.1 (2005)	83 (+3)
6.1 (2012)	88 (+5)

Chart

[Code chart](#) ^(PDF)

Note: ^{[1][2]}

Basic Multilingual Plane

U+0000 bis U+FFFF

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F	
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F	
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F	
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F	
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F	
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F	
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F	
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F	
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F	
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF	
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF	
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF	
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF	
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF	

Latin script

Non-Latin European scripts

Middle Eastern and Southwest Asian scripts

South and Central Asian scripts

African scripts

Southeast Asian scripts

East Asian scripts

American scripts

Indonesian and Oceanic scripts

Symbols

Notational systems

CJK characters

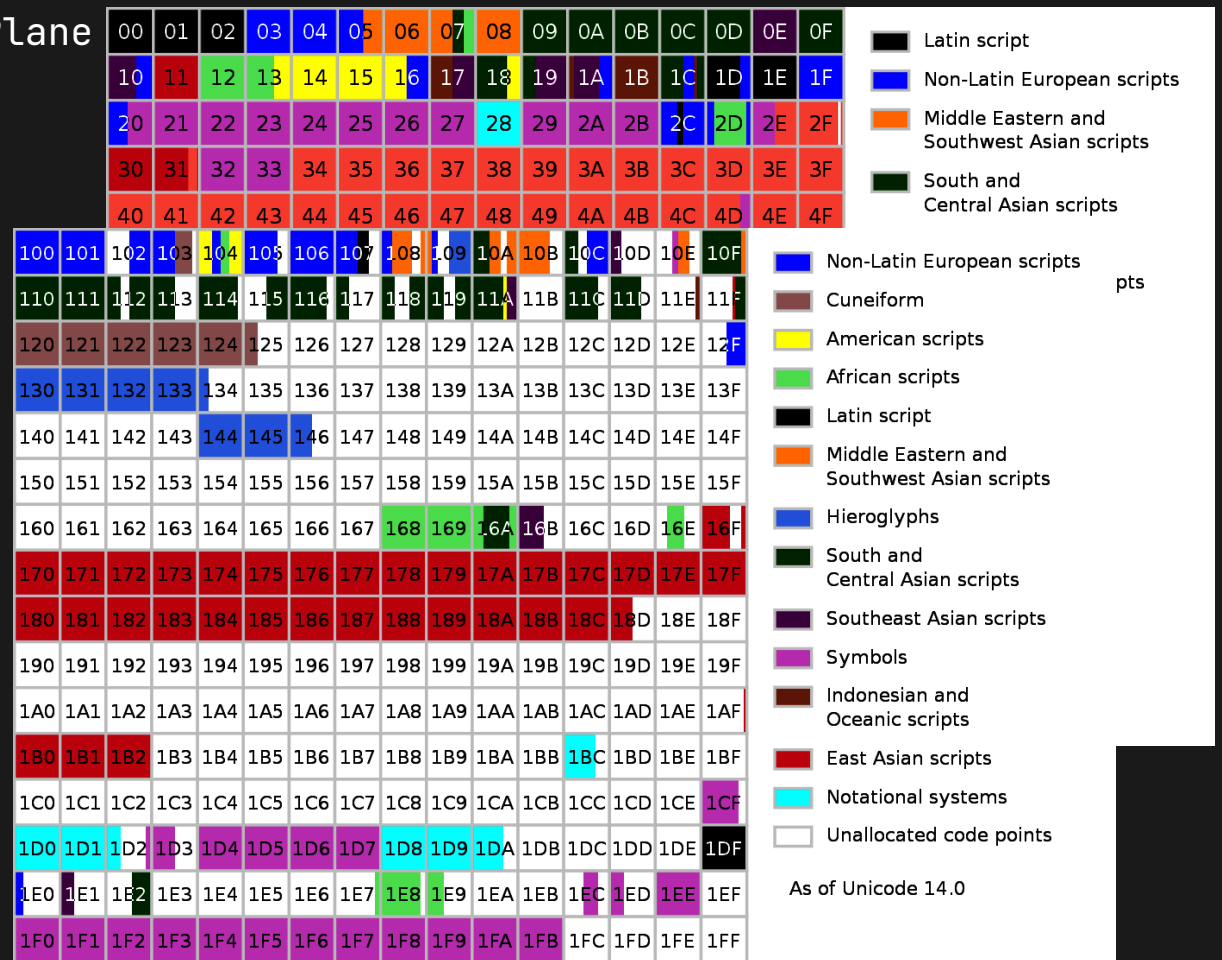
UTF-16 surrogates

Private use

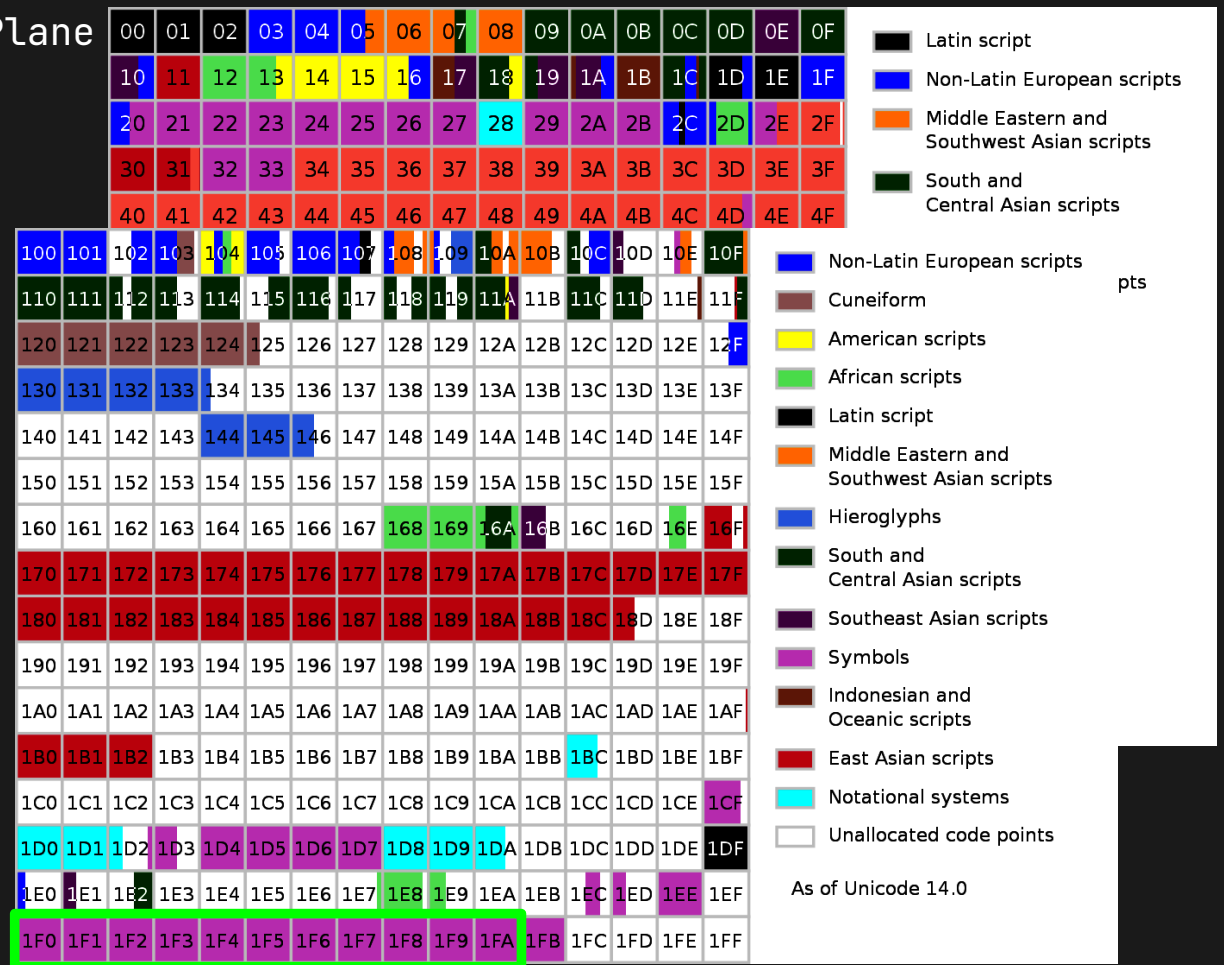
As of Unicode 14.0

Basic Multilingual Plane U+0000 bis U+FFFF

Supplementary Multilingual Plane U+10000 bis U+1FFFF



Basic Multilingual Plane U+0000 bis U+FFFF



Supplementary Multilingual Plane U+10000 bis U+1FFFF

Emoji Blöcke hier 🤪🥰 →

Basic Multilingual Plane U+0000 bis U+FFFF

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F

- Latin script
- Non-Latin European scripts
- Middle Eastern and Southwest Asian scripts
- South and Central Asian scripts

Supplementary Multilingual Plane U+10000 bis U+1FFFF

100	101	102	103	104	105	106	107	108	109	10A	10B	10C	10D	10E	10F
110	111	112	113	114	115	116	117	118	119	11A	11B	11C	11D	11E	11F
120	121	122	123	124	125	126	127	128	129	12A	12B	12C	12D	12E	12F
130	131	132	133	134	135	136	137	138	139	13A	13B	13C	13D	13E	13F
140	141	142	143	144	145	146	147	148	149	14A	14B	14C	14D	14E	14F

- Non-Latin European scripts
- Cuneiform
- American scripts
- African scripts
- Latin script
- Middle Eastern and

Supplementary Ideographic Plane U+20000 bis U+2FFFF

200	201	202	203	204	205	206	207	208	209	20A	20B	20C	20D	20E	20F
210	211	212	213	214	215	216	217	218	219	21A	21B	21C	21D	21E	21F
220	221	222	223	224	225	226	227	228	229	22A	22B	22C	22D	22E	22F
230	231	232	233	234	235	236	237	238	239	23A	23B	23C	23D	23E	23F
240	241	242	243	244	245	246	247	248	249	24A	24B	24C	24D	24E	24F
250	251	252	253	254	255	256	257	258	259	25A	25B	25C	25D	25E	25F
260	261	262	263	264	265	266	267	268	269	26A	26B	26C	26D	26E	26F
270	271	272	273	274	275	276	277	278	279	27A	27B	27C	27D	27E	27F
280	281	282	283	284	285	286	287	288	289	28A	28B	28C	28D	28E	28F
290	291	292	293	294	295	296	297	298	299	29A	29B	29C	29D	29E	29F
2A0	2A1	2A2	2A3	2A4	2A5	2A6	2A7	2A8	2A9	2AA	2AB	2AC	2AD	2AE	2AF
2B0	2B1	2B2	2B3	2B4	2B5	2B6	2B7	2B8	2B9	2BA	2BB	2BC	2BD	2BE	2BF
2C0	2C1	2C2	2C3	2C4	2C5	2C6	2C7	2C8	2C9	2CA	2CB	2CC	2CD	2CE	2CF
2D0	2D1	2D2	2D3	2D4	2D5	2D6	2D7	2D8	2D9	2DA	2DB	2DC	2DD	2DE	2DF
2E0	2E1	2E2	2E3	2E4	2E5	2E6	2E7	2E8	2E9	2EA	2EB	2EC	2ED	2EE	2EF
2F0	2F1	2F2	2F3	2F4	2F5	2F6	2F7	2F8	2F9	2FA	2FB	2FC	2FD	2FE	2FF

- CJK characters
- Unallocated code points

As of Unicode 14.0

Basic Multilingual Plane U+0000 bis U+FFFF

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F

- Latin script
- Non-Latin European scripts
- Middle Eastern and Southwest Asian scripts
- South and Central Asian scripts

Supplementary Multilingual Plane U+10000 bis U+1FFFF

100	101	102	103	104	105	106	107	108	109	10A	10B	10C	10D	10E	10F
110	111	112	113	114	115	116	117	118	119	11A	11B	11C	11D	11E	11F
120	121	122	123	124	125	126	127	128	129	12A	12B	12C	12D	12E	12F
130	131	132	133	134	135	136	137	138	139	13A	13B	13C	13D	13E	13F
140	141	142	143	144	145	146	147	148	149	14A	14B	14C	14D	14E	14F

- Non-Latin European scripts
- Cuneiform
- American scripts
- African scripts
- Latin script
- Middle Eastern and

Supplementary Ideographic Plane U+20000 bis U+2FFFF

200	201	202	203	204	205	206	207	208	209	20A	20B	20C	20D	20E	20F
210	211	212	213	214	215	216	217	218	219	21A	21B	21C	21D	21E	21F
220	221	222	223	224	225	226	227	228	229	22A	22B	22C	22D	22E	22F
230	231	232	233	234	235	236	237	238	239	23A	23B	23C	23D	23E	23F
240	241	242	243	244	245	246	247	248	249	24A	24B	24C	24D	24E	24F
250	251	252	253	254	255	256	257	258	259	25A	25B	25C	25D	25E	25F
260	261	262	263	264	265	266	267	268	269	26A	26B	26C	26D	26E	26F
270	271	272	273	274	275	276	277	278	279	27A	27B	27C	27D	27E	27F
280	281	282	283	284	285	286	287	288	289	28A	28B	28C	28D	28E	28F
290	291	292	293	294	295	296	297	298	299	29A	29B	29C	29D	29E	29F
2A0	2A1	2A2	2A3	2A4	2A5	2A6	2A7	2A8	2A9	2AA	2AB	2AC	2AD	2AE	2AF
2B0	2B1	2B2	2B3	2B4	2B5	2B6	2B7	2B8	2B9	2BA	2BB	2BC	2BD	2BE	2BF
2C0	2C1	2C2	2C3	2C4	2C5	2C6	2C7	2C8	2C9	2CA	2CB	2CC	2CD	2CE	2CF
2D0	2D1	2D2	2D3	2D4	2D5	2D6	2D7	2D8	2D9	2DA	2DB	2DC	2DD	2DE	2DF
2E0	2E1	2E2	2E3	2E4	2E5	2E6	2E7	2E8	2E9	2EA	2EB	2EC	2ED	2EE	2EF
2F0	2F1	2F2	2F3	2F4	2F5	2F6	2F7	2F8	2F9	2FA	2FB	2FC	2FD	2FE	2FF

- CJK characters
- Unallocated code points

As of Unicode 14.0

(nächsten Planes weitestgehend unbelegt)

Codierung

Unicode-String $\equiv$ Array an Codepoints

Codierung

Unicode-String $\equiv$ Array an Codepoints

"Марков: 🍷 für 3€"

Codierung

Unicode-String $\equiv$ Array an Codepoints

"Марков: 🥙 für 3€" $\equiv$ [U+41C CYRILLIC CAPITAL LETTER EM
U+430 CYRILLIC SMALL LETTER A
U+440 CYRILLIC SMALL LETTER ER
U+43A CYRILLIC SMALL LETTER KA
U+43E CYRILLIC SMALL LETTER O
U+432 CYRILLIC SMALL LETTER VE
U+3A COLON
U+20 SPACE
U+1F959 STUFFED FLATBREAD
U+20 SPACE
U+66 LATIN SMALL LETTER F
U+FC LATIN SMALL LETTER U WITH DIAERESIS
U+72 LATIN SMALL LETTER R
U+20 SPACE
U+33 DIGIT THREE
U+20AC EURO SIGN]

Codierung

Unicode-String $\equiv$ Array an Codepoints

"Марков: 🥙 für 3€" $\equiv$ [U+41C CYRILLIC CAPITAL LETTER EM
U+430 CYRILLIC SMALL LETTER A
U+440 CYRILLIC SMALL LETTER ER
U+43A CYRILLIC SMALL LETTER KA
U+43E CYRILLIC SMALL LETTER O
U+432 CYRILLIC SMALL LETTER VE
U+3A COLON
U+20 SPACE
U+1F959 STUFFED FLATBREAD
U+20 SPACE
U+66 LATIN SMALL LETTER F
U+FC LATIN SMALL LETTER U WITH DIAERESIS
U+72 LATIN SMALL LETTER R
U+20 SPACE
U+33 DIGIT THREE
U+20AC EURO SIGN]

Wie von Codepoint-Array zu Byte-Array?

Codierung: UTF-32

	U+1F959	00	01	f9	59
	U+20	00	00	00	20
f	U+66	00	00	00	66
ü	U+FC	00	00	00	fc
r	U+72	00	00	00	72
	U+20	00	00	00	20
3	U+33	00	00	00	33
€	U+20AC	00	00	20	ac

Codierung: UTF-32



U+1F959 00 01 f9 59

U+20 00 00 00 20

f

U+66 00 00 00 66

ü

U+FC 00 00 00 fc

r

U+72 00 00 00 72

U+20 00 00 00 20

3

U+33 00 00 00 33

€

U+20AC 00 00 20 ac

Codierung: UTF-32



U+1F959 00 01 f9 59

U+20 00 00 00 20

f

U+66 00 00 00 66

ü

U+FC 00 00 00 fc

r

U+72 00 00 00 72

U+20 00 00 00 20

3

U+33 00 00 00 33

€

U+20AC 00 00 20 ac



super simpel, $O(1)$ Zugriff,
schnelles Slicing, ...

Codierung: UTF-32

🍲	U+1F959	00	01	f9	59
	U+20	00	00	00	20
f	U+66	00	00	00	66
ü	U+FC	00	00	00	fc
r	U+72	00	00	00	72
	U+20	00	00	00	20
3	U+33	00	00	00	33
€	U+20AC	00	00	20	ac



super simpel, $O(1)$ Zugriff,
schnelles Slicing, ...



viele redundante 00, v.a. im
lateinischen Skript, nicht
null-terminated

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes



„a ray of sunshine in this
whole encoding mess“

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes

f U+66 000000 0000000000 01100110 $\leq$ U+7F
 $\Rightarrow$ 01100110 $\Rightarrow$ 1 Byte

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes

```
f    U+66      00000  000000000  01100110  ≤  U+7F
                                     ⇒  01100110  ⇒  1 Byte
```

ü U+FC 00000 00000000 11111100 ≤ U+7FF
 ⇒ 00011 111100 ⇒ 2 Byte

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes

```
f    U+66      00000  000000000  01100110  ≤  U+7F
                                     ⇒  01100110  ⇒  1 Byte
```

ü U+FC 00000 00000000 11111100 ≤ U+7FF
⇒ 11000011 10111100 ⇒ 2 Byte

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes

```
f    U+66      00000  000000000  01100110  ≤  U+7F
                                     ⇒  01100110  ⇒  1 Byte
```

ü U+FC 00000 00000000 11111100 ≤ U+7FF
 ⇒ 11000011 10111100 ⇒ 2 Byte

€ U+20AC 000000 00100000 10101100 ≤ U+FFFF
⇒ 11100010 10000010 10101100 ⇒ 3 Byte

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes

f U+66 00000 000000000 01100110 $\leq$ U+7F
 $\Rightarrow$ 01100110 $\Rightarrow$ 1 Byte

ü U+FC 00000 00000000 11111100 $\leq$ U+7FF
 $\Rightarrow$ 11000011 10111100 $\Rightarrow$ 2 Byte

€ U+20AC 00000 00100000 10101100 $\leq$ U+FFFF
 $\Rightarrow$ 11100010 10000010 10101100 $\Rightarrow$ 3 Byte

🍷 U+1F959 00001 11111001 01011001 $\geq$ U+10000
 $\Rightarrow$ 11110000 10011111 10101101 10011001 $\Rightarrow$ 4 Byte

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes

f U+66 00000 000000000 01100110 $\leq$ U+7F
 $\Rightarrow$ 01100110 $\Rightarrow$ 1 Byte

ü U+FC 00000 00000000 11111100 $\leq$ U+7FF
 $\Rightarrow$ 11000011 10111100 $\Rightarrow$ 2 Byte

€ U+20AC 00000 00100000 10101100 $\leq$ U+FFFF
 $\Rightarrow$ 11100010 10000010 10101100 $\Rightarrow$ 3 Byte

🍷 U+1F959 00001 11111001 01011001 $\geq$ U+10000
 $\Rightarrow$ 11110000 10011111 10101101 10011001 $\Rightarrow$ 4 Byte

Codierung: UTF-8

Eine *variable-width* Codierung für Unicode:
1 Codepoint $\Rightarrow$ 1-4 Bytes

f U+66 00000 000000000 01100110 $\leq$ U+7F
 $\Rightarrow$ 01100110 $\Rightarrow$ 1 Byte

ü U+007C  selbst-synchronisierend,
effizient(er), rückwärts-
kompatibel mit ASCII,
de-facto default

€ U+20AC  Codepoint-Indexierung nur
in $O(n)$

🍲 U+1F345 $>$ U+10000

$\Rightarrow$ 11110000 10011111 10101101 10011001 $\Rightarrow$ 4 Byte

Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

- Latin script
- Non-Latin European scripts
- Middle Eastern and Southwest Asian scripts
- South and Central Asian scripts
- African scripts
- Southeast Asian scripts
- East Asian scripts
- American scripts
- Indonesian and Oceanic scripts
- Symbols
- Notational systems
- CJK characters
- Private use
- Unallocated code points

in Java: `char-s`

Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F	
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F	
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F	
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F	
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F	
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F	
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F	
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F	
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F	
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF	
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF	
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF	
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF	
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF	

- Latin script
- Non-Latin European scripts
- Middle Eastern and Southwest Asian scripts
- South and Central Asian scripts
- African scripts
- Southeast Asian scripts
- East Asian scripts
- American scripts
- Indonesian and Oceanic scripts
- Symbols
- Notational systems
- CJK characters
- Private use
- Unallocated code points

in Java: `char-s`

Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	Latin script
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F	
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F	Non-Latin European scripts
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F	
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F	Middle Eastern and Southwest Asian scripts
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F	
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F	South and Central Asian scripts
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F	
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F	African scripts
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F	
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	Southeast Asian scripts
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF	
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF	East Asian scripts
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF	
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF	American scripts
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF	

UTF-16 surrogates

Private use

Unallocated code points

- Latin script
- Non-Latin European scripts
- Middle Eastern and Southwest Asian scripts
- South and Central Asian scripts
- African scripts
- Southeast Asian scripts
- East Asian scripts
- American scripts
- Indonesian and Oceanic scripts
- Symbols
- Notational systems
- CJK characters
- UTF-16 surrogates
- Private use
- Unallocated code points

in Java: `char-s`

Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

- Latin script
- Non-Latin European scripts
- Middle Eastern and Southwest Asian scripts
- South and Central Asian scripts
- African scripts
- Southeast Asian scripts
- East Asian scripts
- American scripts
- Indonesian and Oceanic scripts
- Symbols
- Notational systems
- CJK characters
- UTF-16 surrogates
- Private use
- Unallocated code points

100	101	102	103	104	105	106	107	108	109	10A
110	111	112	113	114	115	116	117	118	119	11A
120	121	122	123	124	125	126	127	128	129	12A
130	131	132	133	134	135	136	137	138	139	13A
140	141	142	143	144	145	146	147	148	149	14A
150	151	152	153	154	155	156	157	158	159	15A
160	161	162	163	164	165	166	167	168	169	16A
170	171	172	173	174	175	176	177	178	179	17A
180	181	182	183	184	185	186	187	188	189	18A
190	191	192	193	194	195	196	197	198	199	19A
1A0	1A1	1A2	1A3	1A4	1A5	1A6	1A7	1A8	1A9	1AA
1B0	1B1	1B2	1B3	1B4	1B5	1B6	1B7	1B8	1B9	1BA
1C0	1C1	1C2	1C3	1C4	1C5	1C6	1C7	1C8	1C9	1CA
1D0	1D1	1D2	1D3	1D4	1D5	1D6	1D7	1D8	1D9	1DA
1E0	1E1	1E2	1E3	1E4	1E5	1E6	1E7	1E8	1E9	1EA
1F0	1F1	1F2	1F3	1F4	1F5	1F6	1F7	1F8	1F9	1FA

in Java: char-s

surrogate pair



Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

f U+66 00000 000000000 01100110 $\leq$ U+FFFF
 $\Rightarrow$ 00000000 11111100 $\Rightarrow$ 2 Byte

€ U+20AC 00000 00100000 10101100
 $\Rightarrow$ 00010000 10101100

🍌 U+1F959 00001 11111001 01011001 $\geq$ U+10000
 $\Rightarrow$ 11011000 00111110 11011101 01011001 $\Rightarrow$ 4 Byte
 high surrogate in BMP low surrogate in BMP

Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

f U+66 00000 000000000 01100110 $\leq$ U+FFFF
 $\Rightarrow$ 00000000 11111100 $\Rightarrow$ 2 Byte

€ U+20AC 00000 00100000 10101100
 $\Rightarrow$ 00010000 10101100

🍌 U+1F959 00001 11111001 01011001 $\geq$ U+10000
 $\Rightarrow$ 11011000 00111110 11011101 01011001 $\Rightarrow$ 4 Byte
high surrogate in BMP low surrogate in BMP

in Java: char-s

Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

f U+66 00000 000000000 01100110 $\leq$ U+FFFF
 $\Rightarrow$ 00000000 11111100 $\Rightarrow$ 2 Byte

€ U+20AC 00000 00100000 10101100
 $\Rightarrow$ 00010000 10101100

🍌 U+1F959 00001 11111001 01011001 $\geq$ U+10000
 $\Rightarrow$ 11011000 00111110 11011101 01011001 $\Rightarrow$ 4 Byte
high surrogate in BMP low surrogate in BMP

in Java: char-s

Codierung: UTF-16 (deprecated)



Eine *variable-width* Codierung für Unicode:
Codepoint im BMP $\Rightarrow$ 2 Byte; ansonsten 4 Byte

f U+66 00000 00000000 01100110 $\leq$ U+FFFF
 $\Rightarrow$ 00000000 11111100 $\Rightarrow$ 2 Byte

€ U+20AC 00000 00100000 10101100
 $\Rightarrow$ 00010000 10101100

🌮 U+1F959 00001 11111001 01011001 $\geq$ U+10000
 $\Rightarrow$ 11011000 00111110 11011101 01011001 $\Rightarrow$ 4 Byte
high surrogate in BMP low surrogate in BMP

"🌮".length == 2 formal: code unit count
"🌮".codePointCount() == 1

in Java: char-s

```
\usepackage[utf8]{inputenc}  
(default seit 2018)
```

```
<html><head>  
...  
<meta charset="utf-8">
```

Nutzt Unicode (und am besten UTF-8)!

Immer Codierung angeben!

```
Received: from mailmaster.uni-wuerzburg.de  
...  
From: ...  
Content-Type: text/plain; charset=utf-8  
  
HTTP 200 OK  
...  
Content-Type: application/javascript;  
charset=utf-8
```

Python Coding Demo!

Kombinierende Zeichen



Андрей Андреевич Марков

Kombinierende Zeichen



Андрѐй Андрѐевич Ма́рков

Kombinierende Zeichen



Андрéй Андрéевич Мäркoв

U+435
Cyrillic Small
Letter Ie

U+301
Combining Acute Accent

e + é → é

Kombinierende Zeichen

e + é → é

s + é + é + é + p + é + ... → s é é é p é ...

Kombinierende Zeichen

e + ◌́ → é

s + ◌́ + ◌́ + ◌́ + ◌́ + p + ◌́ + ... → špóoky

त + ् + र + ि + स + ् → त्रिस्

Kombinierende Zeichen

e + ◌̇ → é

s + ◌̣ + ◌̇ + ◌̣ + p + ◌̇ + ... → śpōōky̐

t(a) cancel vowel r(a) i s(a) cancel vowel

त + ् + र + ि + स + ् → त्रिस्

(Hindi: *tris*, dreifach)

Kombinierende Zeichen

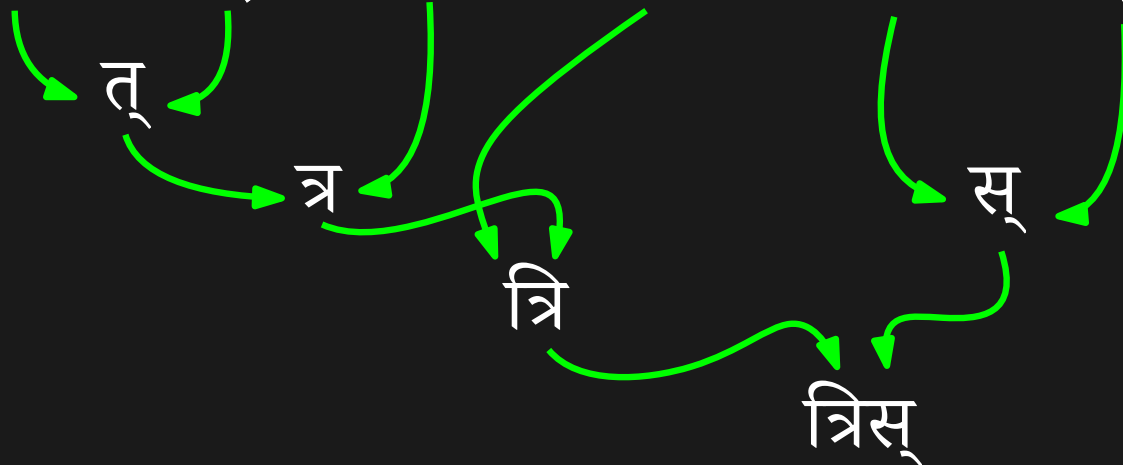
e + ◌̇ → é

s + ◌̇ + ◌̇ + ◌̇ + p + ◌̇ + ... → śpōōk̐y̐

t(a) cancel vowel r(a) i s(a) cancel vowel

त + ◌̇ + र + ि + स + ◌̇ → त्रिस्

(Hindi: *tris*, dreifach)



Kombinierende Zeichen

e + ◌́ → é

s + ◌́ + ◌́ + ◌́ + p + ◌́ + ... → špóoky

त + ् + र + ि + स + ् → त्रिस्

🚩 + [VS16] + [ZWJ] + 🌈 → 🏳️

Kombinierende Zeichen

e + ◌ → é

s + ◌ + ◌ + ◌ + ◌ + p + ◌ + ... → špóoky

त + ् + र + ि + स + ् → त्रिस्

🚩 + U+FE0F Variant Selector 16 [VS16] + U+200D Zero Width Joiner [ZWJ] + 🌈 → 🏳️‍🌈

The diagram illustrates the process of combining a flag emoji with a rainbow emoji to create a rainbow flag emoji. It shows the sequence: a flag emoji, followed by a plus sign, the text 'U+FE0F Variant Selector 16 [VS16]', another plus sign, the text 'U+200D Zero Width Joiner [ZWJ]', a third plus sign, a rainbow emoji, and finally an arrow pointing to the resulting rainbow flag emoji. Green curved arrows indicate the flow of the sequence, and a green dot marks the point where the joiner is applied.

Kombinierende Zeichen

e + ◌́ → é

s + ◌́ + ◌́ + ◌́ + p + ◌́ + ... → špóóký

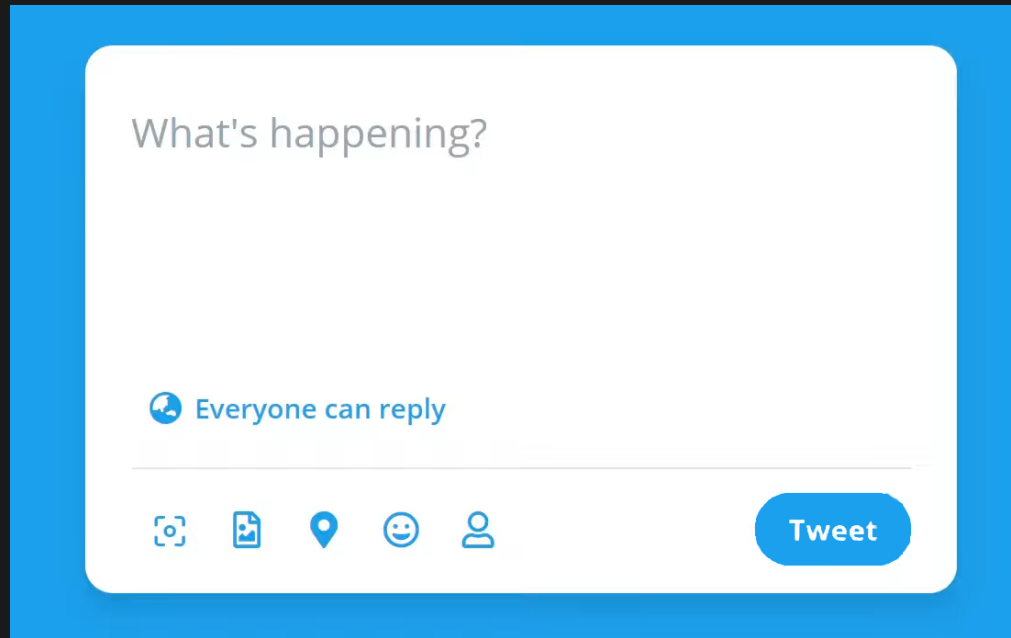
त + ् + र + ि + स + ् → त्रिस्

🚩 + [VS16] + [ZWJ] + 🌈 → 🏳️‍🌈

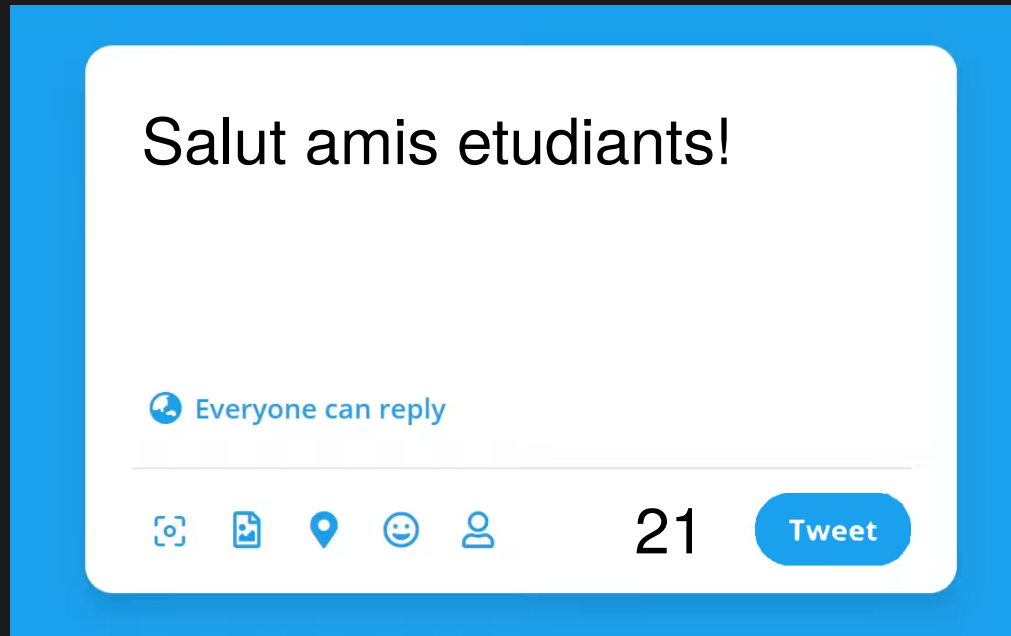
a + ◌̈ → ä ≠ ä

U+00E4 Latin Small
Letter A with
Diaeresis

Codepoint vs Zeichen: Graphemcluster



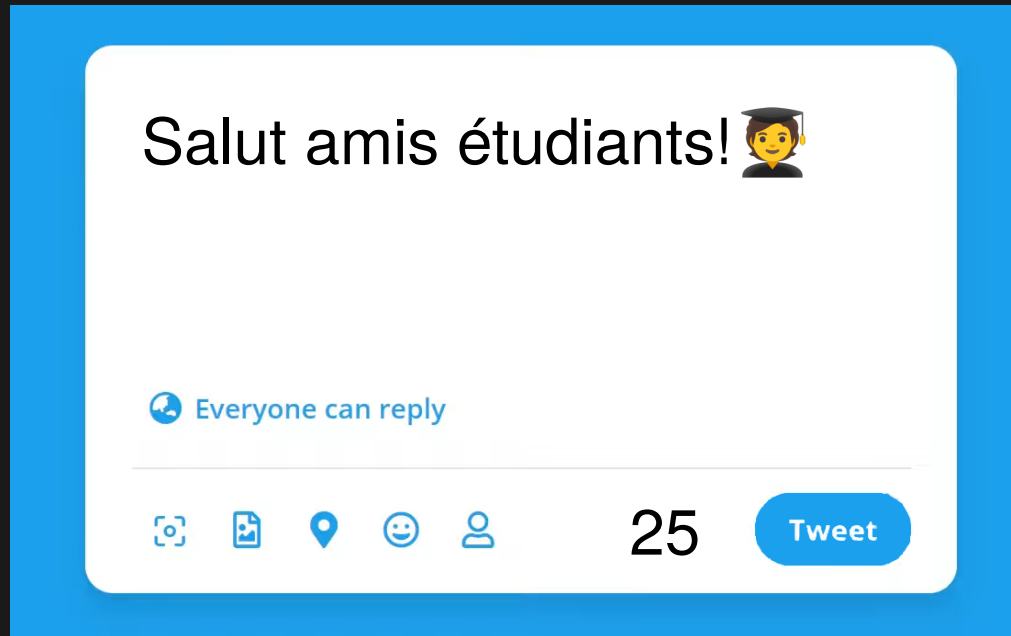
Codepoint vs Zeichen: Graphemcluster



Codepoint vs Zeichen: Graphemcluster



Codepoint vs Zeichen: Graphemcluster



Codepoint vs Zeichen: Graphemcluster

2 Codepoints!

3 Codepoints!



Codepoint vs Zeichen: Graphemcluster

2 Codepoints!

3 Codepoints!



```
len(tweet)  
= # Codepoints  
≠ # „Zeichen“
```

Codepoint vs Zeichen: Graphemcluster

2 Codepoints!

3 Codepoints!



```
len(tweet)  
= # Codepoints  
≠ # „Zeichen“
```

„grapheme clusters“

„user-perceived characters“

Codepoint vs Zeichen: Graphemcluster

2 Codepoints!

3 Codepoints!



```
len(tweet)  
= # Codepoints  
≠ # „Zeichen“
```

„grapheme clusters“

„user-perceived characters“

Merke: Was wollt ihr zählen?

Codepoint vs Zeichen: Graphemcluster

2 Codepoints!

3 Codepoints!



```
len(tweet)  
= # Codepoints  
≠ # „Zeichen“
```

„grapheme clusters“

„user-perceived characters“

Merke: Was wollt ihr zählen?

⇒ UAX #29 Unicode Text Segmentation
Beispiel in Scala mit ICU4J

U+00E4 Latin
Small Letter A

U+0308 Combining
Diaeresis

U+00E4 Latin Small
Letter A with
Diaeresis

a + ö → ä ≠ ä

U+00E4 Latin
Small Letter A

U+0308 Combining
Diaeresis

U+00E4 Latin Small
Letter A with
Diaeresis

a + ö → ä ≠ ä

Normalisierungen UAX #15

U+00E4 Latin
Small Letter A

U+0308 Combining
Diaeresis

U+00E4 Latin Small
Letter A with
Diaeresis

a + ö → ä ≠ ä

Normalisierungen UAX #15

source: "Ångström eöcrit..."

U+00E4 Latin
Small Letter A

U+0308 Combining
Diaeresis

U+00E4 Latin Small
Letter A with
Diaeresis

a + ö → ä ≠ ä

Normalisierungen UAX #15

source: "Ångström eócrit..."

(canonical) decomposition

NFD: "Aöngstroö eócrit..."

U+00E4 Latin
Small Letter A

U+0308 Combining
Diaeresis

U+00E4 Latin Small
Letter A with
Diaeresis

a + ö → ä ≠ ä

Normalisierungen UAX #15

source: "Ångström écrit..."

(canonical) decomposition

NFD: "Aöngstroööm écrit..."

(canonical) composition

NFC: "Ångström écrit..."

U+00E4 Latin
Small Letter A

U+0308 Combining
Diaeresis

U+00E4 Latin Small
Letter A with
Diaeresis

a + ö → ä ≠ ä

Normalisierungen UAX #15

source: "Ångström écrit..."

(canonical) decomposition

NFD: "Aöngstroööm écrit..."

(canonical) composition

NFC: "Ångström écrit..."

NFD ideal zur Volltextsuche

⇒ `search(nfd(needle), nfd(haystack))`

Compatibility Normalization

das geht an die *NLP-Community*

Compatibility Normalization

das geht an die *NLP-Community*

" $\frac{1}{2}$ *ℓ* **Bier**; 100 cm³; Henry IV"

Compatibility Normalization

das geht an die *NLP-Community*

" $\frac{1}{2}$ *ℓ* **Bier**; 100 cm³; Henry IV"

 verstehe
ich nicht

Compatibility Normalization

das geht an die *NLP-Community*

" $\frac{1}{2}$ ℓ **Bier**; 100 cm³; Henry IV"

→ "1/2 ℓ Bier; 100 cm³; Henry IV"

Compatibility Normalization

das geht an die *NLP-Community*

" $\frac{1}{2}$ *ℓ* **Bier**; 100 cm³; Henry IV"

→ "1/2 ℓ Bier; 100 cm³; Henry IV"

source: "für $\frac{1}{2}$ *ℓ* **Bier**..."

Compatibility Normalization

das geht an die *NLP-Community*

" $\frac{1}{2}$ ℓ **Bier**; 100 cm³; Henry IV"

→ "1/2 ℓ Bier; 100 cm³; Henry IV"

source: "für $\frac{1}{2}$ ℓ **Bier**..."



compatibility(!) decomposition

NFKD: "fuör 1/2 ℓ Bier..."

Compatibility Normalization

das geht an die *NLP-Community*

"½ ℓ **Bier**; 100 cm³; Henry IV"

→ "1/2 ℓ Bier; 100 cm³; Henry IV"

source: "für ½ ℓ **Bier**..."

compatibility(!) decomposition

NFKD: "fuöŕ 1/2 ℓ Bier..."

(canonical) composition

NFKC: "für 1/2 ℓ Bier..."

Compatibility Normalization

das geht an die *NLP-Community*

"½ ℓ **Bier**; 100 cm³; Henry IV"

→ "1/2 ℓ Bier; 100 cm³; Henry IV"

source: "für ½ ℓ **Bier**..."

compatibility(!) decomposition

NFKD: "fuör 1/2 ℓ Bier..."

(canonical) composition

NFKC: "für 1/2 ℓ Bier..."

NFKC ideal für NLP-Models: Zeichenvarianten in „verständliche“, allgemeine Normalform

Compatibility Normalization

das geht an die *NLP-Community*

"½ ℓ **Bier**; 100 cm³; Henry IV"

→ "1/2 ℓ Bier; 100 cm³; Henry IV"

source: "für ½ ℓ **Bier**..."

compatibility(!) decomposition

NFKD: "fuör 1/2 ℓ Bier..."

(„fuzzy“ Volltextsuche)

(canonical) composition

NFKC: "für 1/2 ℓ Bier..."

NFKC ideal für NLP-Models: Zeichenvarianten in „verständliche“, allgemeine Normalform

Task: finde alle großgeschriebenen Wörter

Task: finde alle großgeschrieben Wörter

```
"Info und Pizza".match(/[A-Z][a-z]*/g)
```

Task: finde alle großgeschrieben Wörter

```
"Info und Pizza".match(/[A-Z][a-z]*/g)
```

```
⇒ "Info", "Pizza"
```

Task: finde alle großgeschrieben Wörter

```
"Spaß und Pizza".match(/[A-Z][a-z]*/g)
```

Task: finde alle großgeschrieben Wörter

```
"Spaß und Pizza".match(/[A-Z][a-z]*/g)
```

⇒ "Pizza" 😞

Task: finde alle großgeschrieben Wörter

```
"Spaß und Pizza".match(/[A-Z][a-z]*/g)
```

⇒ "Pizza" 😞

```
"Spaß und Pizza. Πλάτων ist im Café trinken.".match(  
  /\p{General_Category=Uppercase_Letter}  
  \p{General_Category=Lowercase_Letter}*/gu)
```

Task: finde alle großgeschrieben Wörter

```
"Spaß und Pizza".match(/[A-Z][a-z]*/g)
```

⇒ "Pizza" 😞

```
"Spaß und Pizza. Πλάτων ist im Café trinken.".match(  
  /\p{General_Category=Uppercase_Letter}  
  \p{General_Category=Lowercase_Letter}*/gu)
```

Unicode-Zeichen mit Property X = Wert Y

Task: finde alle großgeschrieben Wörter

```
"Spaß und Pizza".match(/[A-Z][a-z]*/g)
```

⇒ "Pizza" 😞

```
"Spaß und Pizza. Πλάτων ist im Café trinken.".match(  
  /\p{General_Category=Uppercase_Letter}  
  \p{General_Category=Lowercase_Letter}*/gu)
```

⇒ "Spaß", "Pizza", "Πλάτων", "Café"

Sehr mächtig!

Regex / Properties

Wichtigste Property: **General_Category**

- **Letter** (L): uppercase (Lu), lowercase (Ll), titlecase (Lt), modifier (Lm), other (Lo)

Regex / Properties

Wichtigste Property: **General_Category**

- **Letter** (L): uppercase (Lu), lowercase (Ll), titlecase (Lt), modifier (Lm), other (Lo)



`\p{General_Category=Uppercase_Letter} = \p{gc=Lu} = \p{Lu} = \pLu`

Regex / Properties

Wichtigste Property: **General_Category**

- **Letter** (L): uppercase (Lu), lowercase (Ll), titlecase (Lt), modifier (Lm), other (Lo)
- **Mark** (L): nonspacing (Mn), combining (Mc), enclosing (Me) [u.a. Diakritika]
- **Number** (N): decimal (Nd), letter (Nl), other (No)
- **Punctuation** (P): connector (Pc), dash (Pd), initial quote (Pi), final quote (Pf), open (Ps), close (Pe), other (Po)
- **Symbol** (S): currency (Sc), modifier (Sk), math (Sm), other (So)
- **Separator** (Z): line (Zl), paragraph (Zp), space (Zs)
- **Other** (C): control (Cc), format (Cf), not assigned (Cn), private use (Co), surrogate (Cs)

Regex / Properties

Wichtigste Property: `General_Category`

Weitere Properties:

`Script`

(=Latin, Cyrillic, Devanagari, ...; z.B. `\p{sc=Greek}`)

`White_Space`

(=yes, no; `\p{White_Space}` vs. `\P{White_Space}`)

`Alphabetic`

(=yes, no)

`Emoji`

(=yes, no)

u.v.m.; schaut nach was bei eurer Lieblingssprache
unterstützt wird

honorable mentions

honorable mentions

Collation / Sortierung
(UAX #10) Y < Z < Å

honorable mentions

Collation / Sortierung
(UAX #10) Y < Z < Å

Casing
ü ↔ Ü

honorable mentions

Collation / Sortierung
(UAX #10) Y < Z < Å

Casing
ü ↔ Ü

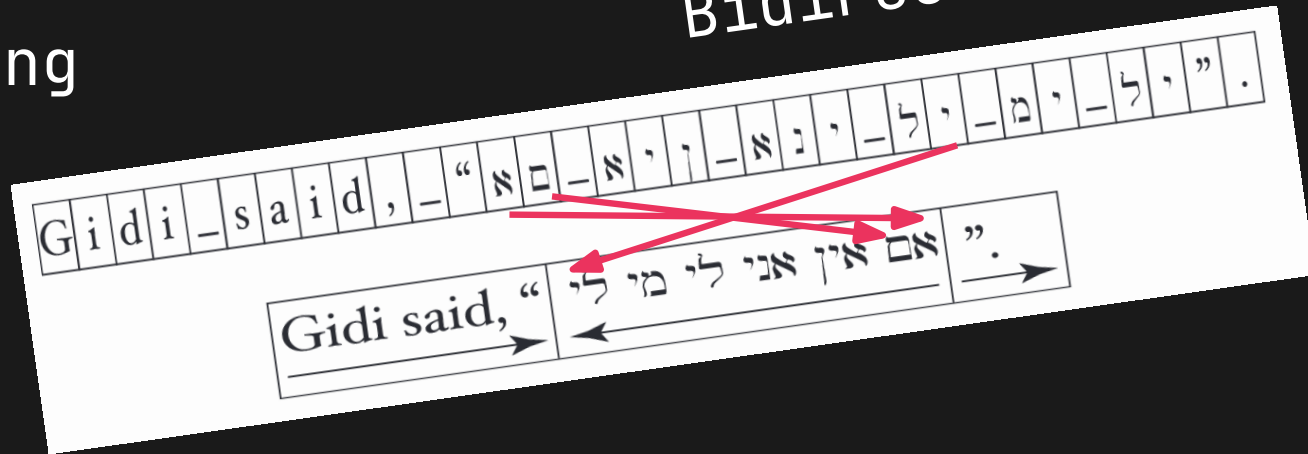
Line Breaking (UAX #14)
„Hier Zeile brechen?“, fragte er.

honorable mentions

Bidirectionality

Collation / Sortierung
(UAX #10) Y < Z < Å

Casing
ü ↔ Ü

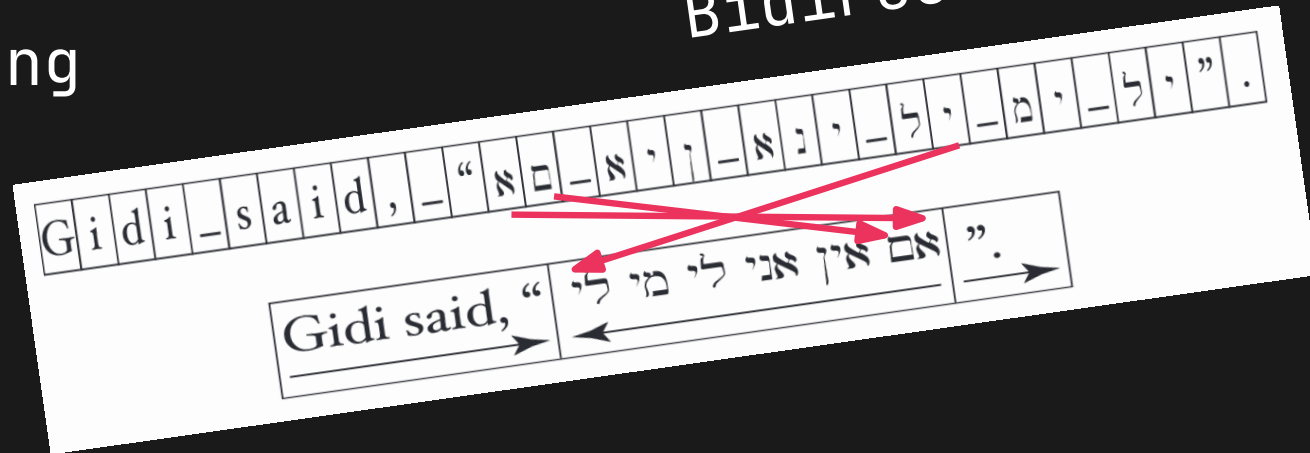


Line Breaking (UAX #14)
„Hier Zeile brechen?“, fragte er.

honorable mentions

Collation / Sortierung
(UAX #10) Y < Z < Å

Bidirectionality



Casing
ü ↔ Ü

Modelle für Schriften
der Welt!!!

لكن لا بد أن أوضح

শুভ সকাল

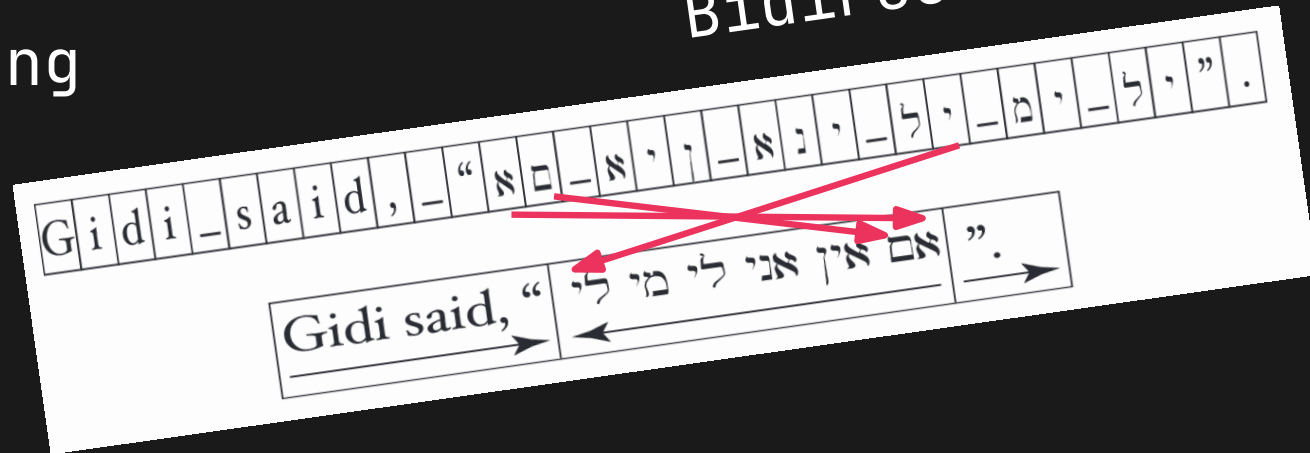
Tシャツを3枚購入しました

Line Breaking (UAX #14)
„Hier Zeile brechen?“, fragte er.

honorable mentions

Collation / Sortierung
(UAX #10) Y < Z < Å

Bidirectionality



Casing
ü ↔ Ü

Modelle für Schriften
der Welt!!!

Line Breaking (UAX #14)
„Hier Zeile brechen?“, fragte er.

لكن لا بد أن أوضح

শুভ সকাল

International
Components for
Unicode (ICU)

Tシャツを3枚購入しました

HarfBuzz text shaper

Best practices

- Unicode + UTF-8 verwenden
- Codierungen mitdenken
- nicht vergessen:
„Zeichen“ nicht gleich „Zeichen“

Best practices

- Unicode + UTF-8 verwenden
- Codierungen mitdenken
- nicht vergessen:
„Zeichen“ nicht gleich „Zeichen“
- (LuaLaTeX macht alles einfacher!)

Unicode, Sprache & Gesellschaft

MODEL VIEW CULTURE

A magazine about technology, culture and diversity.

[HOME](#) [RESOURCES](#) [ALL ISSUES](#) [AUTHORS](#) [NEWS](#) [SUPPORT](#)

ISSUE 18

I Can Text You A Pile of Poo, But I Can't Write My Name

We can't ignore the composition of the Unicode Consortium's members, directors, and officers -- the people who define the everyday writing systems of all languages across the globe.

—by [Aditya Mukerjee](#) on March 17th, 2015



Unicode, Sprache & Gesellschaft

MODEL VIEW CULTURE

A magazine about technology, culture and diversity.

HOME RESOURCES ALL ISSUES AUTHORS NEWS SUPPORT

ISSUE 18

I Can Text You A Pile of Poo, But I Can't Write My Name

We can't ignore the composition of the Unicode Consortium's members, directors, and officers -- the people who define the everyday writing systems of all languages across the globe.

—by [Aditya Mukerjee](#) on March 17th, 2015



আদিত্য মুখার্জি

Unicode, Sprache & Gesellschaft

CJK Unified Ideographs – Han-Unifikation

简体中文
Simplified Chinese

返

正體中文—臺灣
Traditional Chinese — Taiwan

返

繁體中文—香港
Traditional Chinese — Hong Kong

返

日本語
Japanese

返

한국어
Korean

返

Unicode, Sprache & Gesellschaft

CJK Unified Ideographs – Han-Unifikation

简体中文
Simplified Chinese

返

正體中文—臺灣
Traditional Chinese — Taiwan

返

繁體中文—香港
Traditional Chinese — Hong Kong

返

日本語
Japanese

返

한국어
Korean

返

5 mal der GLEICHE Codepoint U+8FD4 CJK
Unified Ideograph-8FD4

Unicode, Sprache & Gesellschaft

CJK Unified Ideographs

简体中文
Simplified Chinese

返

正體中文—臺灣
Traditional Chinese — Taiwan

返

繁體中文
Traditional Chinese

返

Unicode 14.0 character map showing various scripts and systems. The map is a grid of 16x16 blocks, each representing a 16x16 range of code points. The colors indicate different script families or systems. A legend on the right lists the categories:

- Latin script
- Non-Latin European scripts
- Middle Eastern and Southwest Asian scripts
- South and Central Asian scripts
- African scripts
- Southeast Asian scripts
- East Asian scripts
- American scripts
- Indonesian and Oceanic scripts
- Symbols
- Notational systems
- CJK characters
- UTF-16 surrogates
- Private use

As of Unicode 14.0

5 mal der GLEICHE Codepoint U+8FD4 CJK
Unified Ideograph-8FD4

Unicode, Sprache & Gesellschaft

CJK Unified Ideographs – Han-Unifikation

简体中文
Simplified Chinese

返

正體中文—臺灣
Traditional Chinese — Taiwan

返

繁體中文—香港
Traditional Chinese — Hong Kong

返

日本語
Japanese

返

한국어
Korean

返

5 mal der GLEICHE Codepoint U+8FD4 CJK
Unified Ideograph-8FD4

`<span lang="zh-Hant">返</span>` ⇒ 返

`<span lang="ko">返</span>` ⇒ 返

Unicode, Sprache & Gesellschaft

CJK Unified Ideographs – Han-Unifikation

Kontrafaktisches Argument #1:
„Greco Unification“ [Aditya Mukerjee]

Warum haben wir eigentlich a (Latein),
а (Kryllisch), α (Griechisch)
als separate Codepoints?

Unicode, Sprache & Gesellschaft

CJK Unified Ideographs – Han-Unifikation

Kontrafaktisches Argument #1:
„Greco Unification“ [Aditya Mukerjee]

Warum haben wir eigentlich a (Latein),
а (Kryllisch), α (Griechisch)
als separate Codepoints?

δεutsch ⇒ deutsch



WATCHING THE UNICODE PEOPLE TRY TO GOVERN THE INFINITE CHAOS OF HUMAN LANGUAGE WITH CONSISTENT TECHNICAL STANDARDS IS LIKE WATCHING HIGHWAY ENGINEERS TRY TO STEER A RIVER USING TRAFFIC SIGNS.



WATCHING THE UNICODE PEOPLE TRY TO GOVERN THE INFINITE CHAOS OF HUMAN LANGUAGE WITH CONSISTENT TECHNICAL STANDARDS IS LIKE WATCHING HIGHWAY ENGINEERS TRY TO STEER A RIVER USING TRAFFIC SIGNS.

Danke! 😊