Unicode Explained

Jukka K Korpela
Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>! (exclamation mark)</td>
<td>274, 380</td>
</tr>
<tr>
<td>&quot; (quotation mark)</td>
<td>7, 97, 384</td>
</tr>
<tr>
<td># (comments) in Unicode database files</td>
<td>222</td>
</tr>
<tr>
<td># (number sign)</td>
<td>374, 383</td>
</tr>
<tr>
<td>$ (dollar sign)</td>
<td>380, 396, 435</td>
</tr>
<tr>
<td>% (percent sign)</td>
<td>384</td>
</tr>
<tr>
<td>&amp; (ampersand)</td>
<td>377</td>
</tr>
<tr>
<td>' (apostrophe)</td>
<td>78, 377</td>
</tr>
<tr>
<td>vs. single quotation mark</td>
<td>418</td>
</tr>
<tr>
<td>&quot;\0&quot; null character (C)</td>
<td>550</td>
</tr>
<tr>
<td>() (parentheses)</td>
<td>376</td>
</tr>
<tr>
<td>* (asterisk)</td>
<td>377</td>
</tr>
<tr>
<td>+ (plus sign)</td>
<td>384</td>
</tr>
<tr>
<td>catenation operator</td>
<td>557</td>
</tr>
<tr>
<td>, (comma)</td>
<td>379</td>
</tr>
<tr>
<td>(hyphen-minus) (see hyphen-minus)</td>
<td></td>
</tr>
<tr>
<td>. (full stop)</td>
<td>374, 381</td>
</tr>
<tr>
<td>... (ellipsis points)</td>
<td>348, 421</td>
</tr>
<tr>
<td>.doc format</td>
<td>46</td>
</tr>
<tr>
<td>/ (division slash)</td>
<td>432</td>
</tr>
<tr>
<td>/ (solidus)</td>
<td>374, 386</td>
</tr>
<tr>
<td>division mark and</td>
<td>397</td>
</tr>
<tr>
<td>The 10 Unicode Design Principles</td>
<td>155</td>
</tr>
<tr>
<td>16-bit, 161, 207</td>
<td></td>
</tr>
<tr>
<td>Java and</td>
<td>562</td>
</tr>
<tr>
<td>7bit Content-Transfer-Encoding value</td>
<td>498</td>
</tr>
<tr>
<td>8-bit, 4, 552</td>
<td></td>
</tr>
<tr>
<td>ad hoc, defined by fonts</td>
<td>134</td>
</tr>
<tr>
<td>codes, 127–135</td>
<td></td>
</tr>
<tr>
<td>Content-Transfer-Encoding value</td>
<td>498</td>
</tr>
<tr>
<td>headers, using characters in</td>
<td>530</td>
</tr>
<tr>
<td>Vietnamese codes</td>
<td>140</td>
</tr>
<tr>
<td>: (colon)</td>
<td>379</td>
</tr>
<tr>
<td>; (semicolon)</td>
<td>386</td>
</tr>
<tr>
<td>in Unicode database files</td>
<td>222</td>
</tr>
<tr>
<td>&lt; (less-than sign)</td>
<td>382</td>
</tr>
<tr>
<td>= (equals sign)</td>
<td>380</td>
</tr>
<tr>
<td>&gt; (greater-than sign)</td>
<td>382</td>
</tr>
<tr>
<td>? (question mark)</td>
<td>384</td>
</tr>
<tr>
<td>@ (commercial at)</td>
<td>374, 380</td>
</tr>
<tr>
<td>] (square brackets)</td>
<td>376</td>
</tr>
<tr>
<td>\ (reverse solidus)</td>
<td>374, 386</td>
</tr>
<tr>
<td>&quot; (quotation mark) escape notation</td>
<td>538</td>
</tr>
<tr>
<td>' (apostrophe) escape notation</td>
<td>538</td>
</tr>
<tr>
<td>` (grave accent), 78, 374, 381</td>
<td></td>
</tr>
<tr>
<td>\ (backslash) escape notation</td>
<td>538</td>
</tr>
<tr>
<td>^ (circumflex), 374, 379</td>
<td></td>
</tr>
<tr>
<td>_ (low line, underscore)</td>
<td>383</td>
</tr>
<tr>
<td>in Unicode database files</td>
<td>223</td>
</tr>
<tr>
<td>\ (backslash) escape notation</td>
<td>538</td>
</tr>
<tr>
<td>\ (vertical tab) escape notation</td>
<td>538</td>
</tr>
<tr>
<td>\ (horizontal tab) escape notation</td>
<td>538</td>
</tr>
<tr>
<td>\ (return) escape notation</td>
<td>538</td>
</tr>
<tr>
<td>\ (tab) escape notation</td>
<td>538</td>
</tr>
<tr>
<td>\u escape notation</td>
<td>538</td>
</tr>
<tr>
<td>\v escape notation</td>
<td>538</td>
</tr>
<tr>
<td>\w escape notation</td>
<td>538</td>
</tr>
<tr>
<td>( ) (parentheses), 376</td>
<td></td>
</tr>
<tr>
<td>(pipe, vertical line)</td>
<td>144, 389</td>
</tr>
<tr>
<td>\ (tilde)</td>
<td>388</td>
</tr>
<tr>
<td>\ (inverted exclamation mark)</td>
<td>348, 396</td>
</tr>
<tr>
<td>€ (cent sign)</td>
<td>396, 435</td>
</tr>
<tr>
<td>£ (pound sign)</td>
<td>396, 435</td>
</tr>
<tr>
<td>¥ (yen sign)</td>
<td>397, 435</td>
</tr>
<tr>
<td>\ (broken bar)</td>
<td>399</td>
</tr>
<tr>
<td>§ (section sign)</td>
<td>400</td>
</tr>
</tbody>
</table>

We’d like to hear your suggestions for improving our indexes. Send email to index@oreilly.com.
\* (feminine ordinal indicator), 394
«» (chevrons/guillemets), 396
¬ (not sign), 398
® (registered sign), 400
° (degree sign), 397, 433
± (plus-minus sign), 398
µ (micro sign), 397
¶ (formatting marks), 83
¶ (pilcrow sign), 400
· (middle dot), 399, 431
º (masculine ordinal indicator), 394
degree sign and, 397
¿ (inverted question mark), 348, 396
Æ (U+00C6), 395
Ð (U+00D0), 395
× (multiplication sign), 274, 398, 431
Ø (U+00D8), 395
Þ (U+00DE), 395
ß (sharp s), 395
æ (U+00E6), 349, 395
δ (U+00F0), 395
÷ (division sign), 274, 397
ø (U+00F8), 395
þ (U+00FE), 395
œ, limitations of, 349
– (en dash), 72, 96, 125, 420
— (em dash), 125, 348, 349, 420
"" (smart quotes), 7, 97, 417
‰ (per mile sign), 384
€ (euro-currency sign), 435
₣ (franc sign), 435
€ (euro sign), 397, 435
№ (numero sign), 411
™ (trademark sign), 411
e (estimated symbol), 411
← →, (arrows), 13

A
“A” versions (code page versions), 555
abbreviations (codes), 214
abjads, 27
abstract character repertoire, 136
abstract characters, 11, 282
abstractions (character), 8
abugida, 27, 196
accented characters, 163, 347
Accept-Charset header, 497
Accept-Encoding header, 497
Accept-Language header, 497
accessibility, 464
ACK (acknowledge) character, 392
AddType instructions, 508
Afrikaans ISO 639-1 code, 364
age property, 215
AHex, 215
AI (Ambiguous) LineBreak property value, 271
AL (Arabic letter) Bidi Class value, 260
AL (Ordinary Alphabetic) LineBreak property value, 271
Albanian language, 346
allocation areas, 172
Alpha, 215
alphabetic order, collating, 253
alphabetic scripts, 27
European, 401
The Alphabets of Europe (Everson, Michael), 26
alphanumeric characters, 374
alphasyllabary, 196
alt attribute (HTML), 455, 519
Alt key, 74
Windows, using on, 84
Alt+n method, 87
Alt-0n method, 85
Alt-n method, 86
Alt-X method, 87
alternative names, 165
AltGr key, 74, 78
Alvestrand, Harald, 342
ambiguity among characters, 24
American National Standards Institute (see ANSI character set)
American Standard Code for Information Interchange (see ASCII)
amperand (&), 377
AMSTeX, 105
AN (Arabic number) Bidi Class value, 260
angular brackets, 421–424
annexes (Unicode standard), 289
annotations, 16
ANSI character set, 118
Windows Latin 1, 125
Apache servers, 508
language negotiation and, 521
API (Application Programming Interface), 554
apostrophe (‘), 78, 377
vs. single quotation mark, 418
appearance of characters, 9
application dependent, using virtual keys, 80
application media type, 494
Application Programming Interface (API), 554
application type, 483
ar (Arabic) ISO 639-1 code, 364
Arab (Arabic) script code, 336
Arabic digits (0–9), 375
Arabic language
encodings, 148
ISO 639-1 code, 364
long and short codes for, 336
Arabic presentation forms, 476
ArabicShaping.txt file, 217
Armenian encodings, 148
Armenian scripts, 402
Armn (Armenian) script code, 336
arrows (→ ←), 13
ASCII (American Standard Code for Information Interchange), 37, 45, 117, 374–393
Internet and, 505
storage requirements and, 318
ASCII apostrophe, 377
ASCII Hex Digit, 215
ASCII quotation mark ("), 384
assigned code points, 177
asterisk (*), 377
at sign (@) (see commercial at)
atomic (units of text), 8, 12
attributes (HTML/XML), 360
audio media type, 494
authoring software, 506
auto-synchronization, 295
AutoCorrect (MS Word), 94
auxiliary characters data fields, 343
auxiliary keys, 74
azerty keyboard, 349

B
B (paragraph separator) Bidi Class value, 260
B encoding, 530
B2 LineBreak property value, 271
BA LineBreak property value, 271
BabelPad, 111
backquote (’) (see grave accent)
backslash (\), 374, 386
(see also reverse solidus)
CSS, uses for, 544
escape notations and, 537
backspace (BS) character, 392
backtick (‘) (see grave accent)
Bali script code, 336
Baltic encodings, 148
bar (\), 389
base characters, 224
Base64, 304, 306, 498
basic characters (exemplarCharacters) data fields, 343
Basic Latin, 37, 173, 298, 374–393
illogical division into blocks and, 205
invariance of, 240
rows and blocks, 173
structure of database files and, 222
vs. ISO-8859-1, 299
Basic Multilingual Plane (BMP), 139, 160, 171, 207, 296
CESU-8 encoding, 305
noncharacter code points, 176
storage requirements and, 318
Basque language, 346
Batk script code, 336
BB LineBreak property value, 271
bc (Bidi Class), 215
BCP (Best Current Practice), 319
BE (big-endian), 300
BEL (bell), 392
Beng (Bengali) script code, 336
Best Current Practice (BCP), 319
Bidi C property, 215
Bidi Class, 215
BiDi class, directionality property values, 260
Bidi M property, 215
Bidi Mirrored, 215
Bidi Mirroring Glyph, 215
BidiMirroring.txt, 215
bidirectionality, 259, 286
Bido Class, mirroring, 262
big-endian (BE), 300
Big5 encoding for Chinese, 141
Big5-HKSCS encoding for Chinese, 141
binary collations, 306
binary Content-Transfer-Encoding value, 498
Binary Ordered Compression for Unicode (BOCU-1), 305
BIPM (Bureau International des Poids et Mesures), 430
bit, 46
bitmap fonts, 33
BK * LineBreak property value, 271
blanked (ignorable) settings for collation elements, 255
Blis script code, 336
blk (Block), 215
blocked character, 241
blocks, 172, 215, 434
illogical division into, 205
internal structure of, 174–176
Blocks.txt file, 215
bmg (Bidi Mirroring Glyph), 215
BMP (Basic Multilingual Plane), 139, 160, 171, 207, 296
CESU-8 encoding and, 305
noncharacter code points, 176
storage requirements and, 318
BN (boundary neutral) Bidi Class value, 260
BOCU-1 (Binary Ordered Compression for Unicode), 305
bold type face, 32, 461
BOM (byte order mark), 301, 511
Bopo (Bopomofo) script code, 336
<br> tag (HTML), 469
braces ({ }), 376
Brah script code, 336
Brai (Braille) script code, 336
Breton language, 346
British (U.K.) keyboards, 76
broken bar (¦), 399
browsers, 361
directionality and, 263
encodings, viewing, 54
entity references and, 109
font support for, 58
HTTP headers and, 508
language negotiation and, 523
languages, setting preferences for, 329
settings, 64
BS (backspace) character, 392
Bugi script code, 336
Buhd script code, 336
Bureau International des Poids et Mesures (BIPM), 430
button-like symbols, 456–457
buttons for character input, 578
byte order, 300–303
byte order mark (BOM), 511
bytes, 4, 46

C (Common) rule, 246
C programming language, 306, 377, 547
escape notations and, 537
Unicode, 551
wide characters and, 553
C0/C1 controls, 389–393
HTML characters, 540
XML characters, 540
calendar subtype, 484
CAN (cancel) character, 392
Canadian Multilingual keyboards, 76
Canonical Combining Class (ccc), 215, 232, 241
canonical decomposable, 234
canonical decompositions, 231–234
canonical equivalence, 99, 224, 228
canonical mappings, 226, 227
Cans script code, 336
capital letters, 244
caret (^) (see circumflex)
carriage return (CR), 390, 392, 426
Cascading Style Sheets (CSS) (see CSS)
Case Charts, 247
case folding, 215, 246, 588
case mappings and, 245
d vs. normalization, 238
case mappings, 245
case ordering, collating, 253
case properties, 244–248
case sensitivity, 585
CaseFolding.txt, 220, 246
Catalan language, 346
catalog, 215
catenation operator (.), 557
CB * (Contingent Break Opportunity) LineBreak property value, 271
Cc General Category value, 212
ccc (Canonical Combining Class), 215, 232, 241
CE (Composition Exclusion), 215, 242
cent sign (¢), 396, 435
Central European encodings, 148
CESU-8 (Compatibility Encoding Scheme for UTF-16: 8-bit), 304, 305
cf (Case Folding), 215
Cf General Category value, 212
CGI (Common Gateway Interface), 579
CGI scripts, 526
CGPM (General Conference on Weight and Measures), 430
Cham script code, 336
char data type (C), 547
character codes, 4, 65, 334–339
directionality, 259
using, 145–150
Windows (Microsoft), 126
character data types, 546
character encodings, 65, 482
charset information, 495
forms, 136, 285
on the Web, 503
schemes, 136, 285
character folding (see case folding)
character level, expressing information, 458–461
character map
BabelPad, 111
shortcuts, defining, 97
character maps, 5, 43, 88–93
case, 245
vs. visual mappings, 247
viewing, 247
Character Reference Open (CRO), 539
character references, 42, 108, 539
character repertoire, 66, 339, 487
character sequences, writing special characters, 71
character sets, 48, 117
can lever-oriented processing, 573
characters, 3–67, 6–27, 65
accessibility and, 464
adding to Unicode, 207
ambiguity among, 24
choosing, effects on, 289
classification, 210–213
compatibility (see compatibility characters)
avoiding, 236
definitions of, 20
deprecated/obsolete, 181
directionality, 259
displaying, 34
finding information about, 187–192
identity, 15, 162–169
input and output (I/O), 573–578
inserting symbols in MS Word, 81
Internet message headers and, 497
languages and, 323–370
level, 445–485
misleading names of, 204
names of, 165
numbering, 39–43
octet sequences, 43
Perl and, 556
processing of, 14
properties of, 209–291
repertoires, 36–38
spacing between, 100
strictness of, 23
typing, 73
URLs and, 533
usage, 371–439
vs. font variation, 32
vs. images, 12
writing, 69–115
charCount(int) method, 565
charset, 108
charValue() method, 565
Che (Cherokee) script code, 336
chevrons («»), 396
The Chicago Manual of Style, 26
Chinese language, 369–370
encodings, 141, 148
ISO 639-1 code, 364
Chinese National Standard GB 180303, 313
Chinese/Japanese/Korean (see CJK)
<circle> tag, 231
circled mapping, 474
circumflex accent (^), 374, 379
Cirt script code, 336
CJK (Chinese/Japanese/Korean), 27, 102, 369–370
16-bits and, 161
cultural biases and, 201
codings for, 140–142
properties and, 214
CJKV (Chinese/Japanese/Korean/Vietnamese), 142
CL LineBreak property value, 271
classification of characters, 210–213
CLDR (Common Locale Data Repository), 22, 60, 73, 251, 535, 601–605
programming and, 535
technical definitions of character requirements and, 343
CLDR database, 27
clipboard, 81, 91
cliptext, 81
closed collection, 36
CM * LineBreak property value, 271
Cn General Category value, 212
Co General Category value, 212
code charts, 19, 188
code number, 40, 66, 538
code points, 40, 170, 176–181
  order vs. collating order, 252
  unassigned, 282
code position, 40, 170
code set value, 40
code units, 136, 179, 282
code value, 40
coded character representation, 283
coded character set, 135
coded combinations of diacritic marks, 404
coding space, 170
collating order, 248–256, 252, 359
collation charts, 250
Collation Element Table, 254
collation elements, 254
collection of characters, 15
colon (:) (,), 379
color of text, 460
Combining Class, 215
combining diacritical marks, 225
comma (,), 379
Comma Separated Values (CSV), 424, 484
command menus, 75, 81–84
  writing special characters, 71
comments in Unicode database files, 222
commercial at (@), 374, 380
Common (Zyzy) script code, 335
common character classes, 344
Common Gateway Interface (CGI), 579
Common Locale Data Repository (CLDR) (see CLDR)
common parentheses (()), 376
Comp Ex, 215
compareTo(Character) method, 565
<compat> tag, 229, 231
compat mapping, 474
compatibility characters, 234
  avoiding, 236
  ligatures, 237
compatibility decomposable characters, 234, 235
  definition of, 243
compatibility decompositions, 32, 224, 231–234
Compatibility Encoding Scheme for UTF-16: 8-bit (CESU-8), 304, 305
compatibility equivalence, 233
compatibility formatting tags, 230
compatibility mappings, 226–231, 472
complexity of Unicode, 198
composite characters, 226
Composition Exclusion (CE), 215, 242
compositions, 224
counter languages, 535–545
counter professionals, 8
conformance requirements, 159, 281–289
Connolly, Dan, 9
corresponding scripts, 27
Consortium (see Unicode Consortium)
content negotiation, 515–527
Content-Encoding header, 497
Content-Language header, 497
Content-Transfer-Encoding header, 497
Content-Type header, 147, 498, 504
  MIME headers and, 496
Content-type header
  HTTP headers and, 504–508
contextual forms, 29
continuation line convention, 425
continuity of design principles, 157
control characters, 261, 389–393
  HTML characters, 540
  XML characters, 540
core code points, 178
core conversions, 303
converters, 143, 156
Copt script code, 336
copying (Ctrl-C), 81
copying text formatting, 81
copyright sign (©), 399
core character classes, 344
Cornish language, 346
correct writing (orthography), 371
CP 437 code pages, 128
CP 850 code pages, 128
Cprt (Cypriot) script code, 336
CR (carriage return), 390, 392, 426
CR LineBreak property value, 271
criticism of Unicode, 198–206
CRO (Character Reference Open), 539
Croatian language, 346
cross-mapping tables, 143
CS (common number separator) Bidi Class value, 260
Cs General Category value, 212
CSS (Cascading Style Sheets), 62, 71, 361, 450, 484, 543–545
character/visual mappings and, 248
directionality in, 263
escape notations in, 103
Internet protocols and, 488
CSV (Comma Separated Values), 424, 484
Ctrl key, 75
Ctrl-C (copying), 81
Ctrl-Q (Emacs), 88
Ctrl-V (paste), 91
cultural background, 11
cultural biases, 200
currency sign (¤), 396
currency symbols, 396, 435
cursive, 32
Cyrillic blocks, 174
Cyrillic encodings, 148, 489
Cyrillic KOI8 encodings, 132
Cyrillic script, 402
Cyril (Cyrillic) script code, 336
Cyr script code, 336
Czech language, 346
D
Daniels and Bright (The World’s Writing Systems), 26
Danish language, 346
dashes, 216, 418–421
data, 3–67
data link escape (DLE) character, 392
data transfer, 427
database files, 222
DC (Dublin Core), 361
DC1 (device control one) character, 392
DC2 (device control two) character, 392
DC3 (device control three) character, 392
DC4 (device control four) character, 392
de (German) ISO 639-1 code, 364
dead keys, 76
decoding UTF-8, 298
decomposable characters, 226
decomposed characters, 225
decomposition mappings, 226
Decomposition Type (dt), 216
decompositions, 224–237
default boundary rules, 257
default casing operations, 288
Default Ignorable Code Point (DI), 216
Default Unicode Collation Element Table (DUCET), 254
degree Celsius (°C), 433
degree sign (°), 397
DEL (delete) character, 393
Dep (Deprecated), 216
deprecated characters, 181
derived properties, 213
DerivedAge.txt file, 215
DerivedCoreProperties.txt file, 215
design principles, 153–169
Deva (Devanagari) script code, 336
device controls, 390
DI (Default Ignorable Code Point), 216
Dia (Diacritic), 216
diacritic marks, 76, 402–411
combining, 99, 405
cultural biases and, 200
French characters and, 349
impact of, 224–226
Latin 1 Supplement block and, 394
transcription and, 356
unified, 157
variation of, 409
diacritic order, collating, 253
dictionary order, 249
digit(char,int) method, 565
digraphs, 163, 182
dingbats, 134, 438
dir attribute (HTML), 263
direction property (CSS), 264
directionality, 257–264, 324
directory subtype of text, 484
discretionary hyphen (see soft hyphen)
division sign (+), 274, 397
division slash (/), 432
DLE (data link escape) character, 392
dm (Decomposition Mapping), 216
dns (Domain Name System) data, 484
document character set, 109, 539
dollar sign ($), 380, 396, 435
DOM (Document Object Model), 559
domain names, 531–533
DOS systems, 86
code pages, 128
dot (.) (see full stop) 475
dotted numbers, 475
double diacritics, 409
dsrt (Deseret script code), 336
dt (Decomposition Type), 216
dublin core (dc), 361
DUCET (Default Unicode Collation Element Table), 254
Dutch language, 346
ISO 639-1 code, 364
dynamic compositions, 156, 224

e

East Asian Characters (see CJK) 216, 290
East Asian Width (ea), 216, 290
EBCDIC, 131, 304, 312
ECMA, 120
ECMA (European Computer Manufacturers’ Association)
ECMAScript (JavaScript) and, 559
ECMAScript (see JavaScript)
ecmascript subtype, 485
editors, 111, 427
efficiency, 153–155
Egyd script code, 336
Egyh script code, 336
Egyp script code, 336
el (Greek) ISO 639-1 code, 364
ellipses points (...), 348, 421
EM (end of medium) character, 392
em dash (—), 125, 348, 349, 420
Emacs
Ctrl-Q method, 88
email, 50, 373
defaults and, 489
headers, 491–493, 492, 499
Internet and, 146
protocol headers and, 527
sending Unicode, 53
eMBEDded fonts, 35
eMBEDded information, 328
eMERgency breaks, 268
en (English) ISO 639-1 code, 364
EN (European number) Bidi Class values, 260
en dash (–), 72, 96, 125, 420
spacing between characters and, 101
encoding negotiation, 527
encodings, 49–56, 117–150, 123, 138, 293–320

algorithms for UTF-8, 298
ASCII, 119
auto-detecting, 317
B, 530
choosing, 317–320
common choices of, 148
conversions between, 303
information about, 488
Internet, 146
Q, 529
schemes, 293

English language, 346
ISO 639-1 code, 364
ENQ (enquiry) character, 392
enriched text, 485
entity references, 107–110, 539
enumeration, 215
eo (Esperanto) ISO 639-1 code, 364
EOF indicator (C), 549
EOT (end of transmission) character, 392
equals sign (=), 380
equals(Object) method, 565
equivalent sequences, 156
ES (European number separator) Bidi Class value, 260
es (Spanish) ISO 639-1 code, 365
ESC (escape), 118, 393
escape (ESC), 393
escape mechanisms/notations/sequences, 102–110, 145, 386, 537
Esperanto language, 346
ISO 639-1 code, 364
estimated symbol (€), 411
Estonian language, 346
ET (European number terminator) Bidi Class value, 260
ETB (end of transmission block) character, 392
Ethi (Ethiopic) script code, 336
ETX (end of text) character, 392
EU (European Union), 519
EUC-JP encoding for Japanese, 142
EUC-TW encoding for Chinese, 141
euro sign (€), 397, 435
euro-currency sign (€), 435
eurocents, 396
European alphabetic scripts, 401
European Computer Manufacturers’ Association (ECMA), 120
ECMAScript (JavaScript) and, 559
European Union (EU), 519
EX (Exclamation/Interrogation) LineBreak
property value, 271
exceptNamesList.txt, 222
exclamation mark (!), 274, 380
exemplarCharacters data fields, 343
Ext (Extender), 216
Extensible HyperText Markup Language
(XHTML), 109, 511
attributes for, 360
Extensible Markup Language (see XML)
external identifiers, 583

F
F (Full) rule, 246
failures, displaying characters, 34
fantasy (font), 32
Faroese language, 346
Farsi (Persian) encodings, 148
FC NFKC, 216
feminine ordinal indicator (*ª), 394
FF (form feed), 392, 426
figure space (U+2007), 476
file separator (FS) character, 393
File Transfer Protocol (FTP), 490
file, submitting with input fields, 582
<final> tag, 231
Finnish language, 346
Firefox (see Mozilla Firefox)
fixed width, 412
fixed-width space characters, 100, 475
folding (see case folding)
Font properties extension, 35
font variation, 461
font-family value, 63
<font> tag (HTML), 476
<font> tag, 231
fonts, 29, 31–33, 193–198, 488
8-bit codes, defined by, 134
choosing, 71
embedding, 35
finding, 62
implementation of, 32
IPA on computers and, 356
languages and, 368–370
printers, 61
substitutions, 60
Unicode vs. tricks, 197
usage, 194
web authoring and, 62
working with, 36–64
foreign names, 7
form feed (FF), 392, 426
formal names, 165
formally defined properties, 213
format code points, 178
formatting controls, 390
formatting marks (¶), 83
forms, 79, 314, 578
processing data, 579–582
FORTRAN, 546
Fortress programming language, 537
four-letter codes, 334
fr (French) ISO 639-1 code, 364
<fraction> tag, 231
fractions, 429, 476
franc sign (₣), 435
Free Recode converter, 144
French language, 346, 349–350
ISO 639-1 code, 364
Frisian language, 346
FS (file separator) character, 393
FTP (File Transfer Protocol), 490
Full Composition Exclusion, 215
full stop (.), 374, 381
fullwidth ASCII characters, 325
fundamentally similar characters, 227
furigana, 437

G
Galician language, 346
GB 18030 (Chinese Unicode), 304, 313
GB 18030, encoding for Chinese, 141
GB2312 encoding for Chinese, 141
GBK encoding for Chinese, 141
gc (General Category), 210–213, 216, 290
GCB (Grapheme Cluster Break), 216
General Category (gc), 210–213, 290
General Conference on Weight and Measures
(CGPM), 430
general punctuation, 411–424
generalized URL encodings, 316
generic currency symbol, 396
generic font names, 63
Geok script code, 336
Geor (Georgian) script code, 336
Georgian encodings, 148
Georgian scripts, 402

Index | 643
German language, 346
ISO 639-1 code, 364
GET method, 579
getDirectionality Java function, 564
getNumericValue Java function, 565
getType Java function, 564
GL * LineBreak property value, 271
Glag script code, 336
globalization, 324
glyphs, 11, 16, 29, 66, 391
allowed variation of, 30
images, 30
mapping, 196
operations on, 196
unification across variation, 157
Google searches, 314
Goth (Gothic) script code, 336
Gr Base (Grapheme Base), 216
Gr Ext (Grapheme Extend), 216
Gr Link, 217
grammar checks, 327, 358
Grapheme Cluster Break (GCB), 216
grapheme cluster text element, 256
Grapheme Extend/Link, 216
GraphemeBreakProperty.txt file, 256
graphic code points, 178
graphic symbols, 13, 391
grave accent (‘), 78, 374, 381
greater-than sign (>), 382
Greek language, 336
encodings, 148
ISO 639-1 code, 364
keyboards, 76
Greek script, 401
Greenlandic language, 346
Grek (Greek) script code, 336
grep program, using regular expressions, 594
GS (group separator) character, 393
guillemets («»), 396
Gujr (Gugarati) script code, 336
Guru (Gurmukhi) script code, 336
Hang (Hangul) script code, 336
Hangul Syllable Type (hst), 217
HangulSyllabletype.txt file, 217
Hani (Han) script code, 337
Hano (Hanunoo) script code, 337
Hans script code, 337
Hant script code, 337
hash sign (#), 374, 383
he (Hebrew) ISO 639-1 code, 364
headers, 147, 488, 491–493
HTTP, 51, 504–508, 507
MIME, 50, 493
non-ASCII characters in, 530
protocol, 527–531
Hebr (Hebrew) script code, 337
Hebrew language
encodings, 149
ISO 639-1 code, 364
Hebrew script code, 337
heuristic recognition, 328, 512
Hex, 217
hexadecimal notation, 40
hi (Hindi) ISO 639-1 code, 364
higher-level protocol, 445
Hindi language, ISO 639-1 code, 364
Hira (Hiragana) script code, 337
hiragana characters, 142
Hmng script code, 337
horizontal tabulation HT (TAB) character, 390, 392
horizontal writing direction, 257
HP Roman-8, 127
Hrkt (Katakana_Or_Hiragana) script code, 337
hst (Hangul Syllable Type), 217
HT (horizontal tabulation) TAB character, 390, 392
HTML (HyperText Markup Language), 71, 450
attributes for, 360
character/visual mappings and, 248
characters in, 539–543
CSS, using with, 544
directionality in, 263
entity references, 109
escape notations in, 103
HTTP and, 507
JavaScript and, 558
notations for characters and, 107–110
processing form data, 579–582
registered subtypes of text and, 485
translations and, 524
XHTML and, 511
HTTP headers, 51, 504–508, 507
servers, configuring, 508
HTTP_ACCEPT_LANGUAGE environment
variable, 526
Hung script code, 337
Hungarian language, 346
HY (Hyphen) LineBreak property value, 271
hybrid formulas, using \times, 398
hypertext links, 81
HyperText Markup Language (see HTML)
Hyphen property, 217
hyphen-minus (−), 382, 418
computer languages and, 419
IDNs and, 531
hyphens, 382, 418–421
HZ encoding for Chinese, 141
I/O (Input and Output), 573–578
i18n (internationalization), 323
IAB (Internet Architecture Board), 135
“IBM Classes for Unicode”, 598
iCalendar format, 484
Icelandic language, 346
iconv converter, 144
ICU (International Components for Unicode),
535, 598
ICU4C, 598
ICU4J, 598
ID (Ideographic) LineBreak property value, 271
IDC (ID Continue), 217, 586
identifiers, 17, 375, 583–598
alternative syntax, 592
CSS, 545
Java, 562
pattern characters and, 586
syntax, 587–592
identities (recognized), 11
Ideo (Ideographic), 217
ideographic scripts, 27
ideographic telegraph symbols for months, 476
IDN (Internationalized Domain Names), 313,
531
IDNA (Internationalizing Domain Names in
Applications), 531
IDS (ID Start), 217
IDSB (IDS Binary Operator), 217
IDST (IDS Trinary Operator), 217
IEC (International Electrotechnical
Commission), 120
IETF (Internet Engineering Task Force), 317
IETF Policy on Character Sets and Languages
(RFC 2277, BCP 18), 139
ignorable (blanked) settings for collation
elements, 255
image media type, 494
images vs. characters, 12, 452–457
IME (Input Method Editors), 102
<img> tag (HTML), 455
implementations of fonts, 32
IN (Inseparable) LineBreak property value, 271
indexes, 18
using numbers as, 41
Inds script code, 337
inefficiency of Unicode, 198
information technology (IT), 323–339
informative properties, 209, 215, 221
inherently multilingual, 516
Inherited (Qaai) script code, 335
<i>initial</i> tag, 231
Input and Output (I/O), 573–578
input fields (HTML), 579
Input Method Editors (IME), 102
input methods, 226
Insert menu (MS Word), 83
interface (CGI), 579
Interlinear annotation characters, 437
internal identifiers, 583
internal markers (MS Word), 83
internal structure of blocks, 174–176
International Components for Unicode (ICU),
535, 598
International Electrotechnical Commission
(IEC), 120
international group discussion, 528
International Organization for Standardization
(ISO), 119
International Phonetic Alphabet (see IPA)
International Reference Version (IRV), 119
International System of Units (see SI)
internationalization (i18n), 323
Internationalized Domain Names (IDN), 313,
531
Internationalized Domain Names in
Applications (IDNA), 531
Internet, 146, 365
character encodings and, 503–515
UTF-8 and, 319, 505
Internet Architecture Board (IAB), 135
Internet Engineering Task Force (IETF), 317
Internet Explorer, viewing different encodings, 54
Internet media types (see MIME)
Internet message format, 496
Internet message headers, 488
Internet protocols, 425, 487–533
interpretation of coded characters, 283
interpreting UTF-8, 298
inverted exclamation mark (!), 348, 396
inverted question mark (?), 348, 396
invisible overflow line breaks, 269
IPA (International Phonetic Alphabet), 16, 20, 435
phonetic transcription in, 355
Irish language, 346
IRV (International Reference Version), 119
IS (Infix Separator) LineBreak property value, 271
isce (ISO Comment), 217
isDefined(char) method, 565
isDigit(char) method, 566
isHighSurrogate(char) method, 566
isISOControl(char) method, 566
isLetter(char) method, 566
isLetterOrDigit(char) method, 566
isLowerCase Java function, 566
isLowSurrogate(char) method, 566
isMirrored Java function, 564
ISO (International Organization for Standardization), 119
ISO 10646, 159, 539
ISO 3166, 365
ISO 639, 363–365
ISO 646, 119
ISO 8859 codes, 122
ISO Comment, 217
ISO Latin 1, 37, 51, 122–124, 123, 173, 318
French characters and, 349
Internet and, 505
language coverage of, 345
Spanish characters and, 348
vs. UTF-8, 299
ISO-2022 encoding for Chinese, 141
ISO-2022-JP encoding for Japanese, 142
ISO-8859-1 (see ISO Latin 1)
<isolated> tag, 231
isSpace(char) method, 566
isTitleCase(char), 566
isUnicodeIdentifierPart Java function, 564
isUnicodeIdentifierStart Java function, 564
isUpperCase Java function, 565
isWhitespace Java function, 566
it (Italian) ISO 639-1 code, 364
Italian (Old_Italic) script code, 337
Italian language, 346
ISO 639-1 code, 364
italic type face, 32, 461

J
ja (Japanese) ISO 639-1 code, 364
Japanese language
encodings, 142, 149
ISO 639-1 code, 364
Java programming language, 306, 560–565
I/O, 576–578
Java script code, 337
JavaScript, 542, 558–560
javascript subtype, 485
jg (Joining Group), 217
JL (Hangul L Jamo) LineBreak property value, 271
Join C (Join Control), 217
Joining Group (jg), 217
Joining Type (jt), 217
Joint Technical Committee (JTC), 161
JT (Hangul T Jamo) LineBreak property value, 271
jt (Joining Type), 217
JTC (Joint Technical Committee), 161
JV (Hangul V Jamo) LineBreak property value, 271

K
Kali script code, 337
Kana (Katakana) script code, 337
kanji characters, 142
Kannada (Knda) script code, 337
katakana characters, 142
key combinations, writing special characters, 71
keyboards
variations/settings, 73–77
virtual, 77–80
Khar script code, 337
Khmr (Khmer) script code, 337
Knda (Kannada) script code, 337
ko (Korean) ISO 639-1 code, 365
KOI8-R, 127
Korean language
encodings, 142, 149
ISO 639-1 code, 365

L
(see also ISO Latin 1)
Latin 1 Supplement block, 393–400
Latin Extended Additional block, 401
Latin Extended-A block, 400
Latin Extended-B block, 401
Latin language, 346
ISO 639-1 code, 365
Latin scripts, 28, 331
Latinization, 351
Latn (Latin) script code, 337
Latvian language, 346
LB (see line breaks)
Lb (Line Break), 218, 290
lc (Lowercase Mapping), 218
LDML (Locale Data Markup Language), 607
LE (little-endian), 293, 300
LED displays, 78
left bracket (‘), 376
left curly bracket (‘), 376
left-to-right writing, 257
legacy data/software, 143, 253
Lepc script code, 337
less-than sign (<), 382
letterlike symbols, 411
levels of text, 445–485
lexicographic order, 249
LF (linefeed), 118, 392, 426
LF * (Line Feed) LineBreak property value, 271
libiconv library, 145
library routines (Java), 563–565
ligatures, 163, 195
Limb script code, 337
limitations due to convertibility, 158
Lina script code, 337
Linb script code, 337
Line Break (lb), 290
line breaks, 11, 50, 218
coding line structure and, 426
disallowing and allowing, 416
preventing, 478–479
principles of, 267
properties, 264–281
rules, 270, 275–277
tailoring, 279
line separator (LS), 426, 469
line structure control, 424–427
line-breaking algorithm, 277
line-oriented processing, 573
linear text, 451–452
LineBreak property, 270
LineBreak.txt file, 272–274
linefeed (LF), 118, 426
lining digits, 375
link texts, 519–521
literal characters, 586
Lithuanian language, 346
little-endian (LE), 293, 300
LI General Category value, 211
Lm General Category value, 212
Lo General Category value, 212
Locale Data Markup Language (LDML), 607
locales, 599–610
localization (110n), 324
LOE (Logical Order Exception), 218
logical AND operator (&), 377
logical order, 156
Logical Order Exception (LOE), 218
logical symbols, 397–398
low line (_), 383
Lower (Lowercase) property, 218
lowercase letters, 244
    case folding and, 246
    vs. uppercase, 375
Lowercase Mapping (lc), 218
LRE (left-to-right embedding) Bidi Class value, 260
LRE (left-to-right embedding) control character, 261
LRO (left-to-right) Bidi Class value, 260
LS (line separator), 426, 469
Lt General Category value, 211
Lu General Category value, 211
Luxemburgish language, 346

M
Mac Roman, 129, 500
Macintosh (Mac) computers, 129
macro invocations, 109
macron below (U+0331), 409
major version number, 169
majuscules, 244
Maltese language, 346
Mand script code, 337
Manx Gaelic language, 346
mapping tables, 143
markup-significant characters, 110
masculine ordinal indicator (º), 394
degree sign and, 397
Math property, 218, 290
mathematical notations, 451–454
mathematical operators, 289
mathematical symbols, 397, 427
MathWorld, 452
Maya script code, 337
MBCS (Multibyte Character Sets), 555
Mc General Category value, 212
Me General Category value, 212
media type negotiation, 527
media types, 481, 494
<medial> tag, 231
menu (commands), 81–84
Mero script code, 337
message media type, 494
<meta> tag, 510
metadata, 358–368
methods (Java), 561
methods, writing characters, 69–73
micro sign (µ), 397
Microsoft Windows, 77 (see Windows)
    Alt key, using on, 84
    character maps, 91
Microsoft Word, 45
    character maps in, 89–91
    character/visual mappings and, 247
    default replacements in, 93–97
    fonts, working with, 56
    Format Font command, adjusting spacing
    and, 415
    formatting marks (¶), 83
    French, writing with, 349
    languages and, 326–329
    line breaks and, 266
    not sign (¬), displaying, 398
    quotation marks and, 384
    regular expressions and, 594
    Save As option and, 50
    shortcuts, defining, 97
    Spanish, writing with, 348
    special characters and, 71
    symbols, inserting, 81, 420
    virtual keyboards and, 79
middle dot (.), 399, 431
MIME (Multipurpose Internet Mail Extensions), 50, 493–515
MIME-Version header, 498
minor version number, 169
mirroring characters, 262
Mlym (Malayalam) script code, 337
Mn General Category value, 212
mnemonic identifier, 16, 583
model media type, 494
modern style digits, 375
Modified UTF-8, 304, 306
modifier keys, 76
Mong (Mongolian) script code, 337
monospace (font), 31
most significant bit (MSB), 46
Mozilla Firefox, 54, 329
    Chinese characters and, 370
Mozilla Thunderbird, 52, 71, 147, 492
insertion menu in, 81
Unicode, sending, 54
MS DOS (see DOS)
MS Office, 58, 61
MS Word (see Microsoft Word)
MSB (most significant bit), 46
multi-octet encoding, 48
Multibyte Character Sets (MBCS), 555
multilingual applications, 7
multilingual web sites, 515–527
multipart media type, 494
multiple uses (semantics) of characters, 374
multiplication sign (×), 274, 398, 431
Multipurpose Internet Mail Extensions (see MIME)
multiviews method (Apache), 522
Mymr script code, 337

N
na (Name) property, 218
na1 (Unicode Name), 218
NAK (negative acknowledge) character, 392
Nameprep algorithm, 532
names (Unicode characters), 165–167
<narrow> tag, 231, 476
narrow no-break space (U+202F), 415, 476
national variants of ASCII, 120
NChar (Noncharacter Code Point), 218
Nd General Category value, 212
negative acknowledge (NAK) character, 392
negative kerning, 269
negative transfer, 545
NEL (next line), 426
Netscape, 558
neutral directionality, 261
next line (NEL), 426
NFC (Normalization Form C), 238
W3C normalization and, 243
NFC QC (NFC Quick Check), 218
NFD (Normalization Form D), 238, 240
NFD QC (NFD Quick Check), 218
NFDK (Normalization Form KD), 240
NFKC (Normalization Form KC), 238, 240, 242
NFKC QC (NFKC Quick Check), 218
NFKD (Normalization Form KD), 238
NFKD QC (NFKD Quick Check), 218
Nkoo script code, 337
nl (Dutch) ISO 639-1 code, 364
NL (Next Line) LineBreak property value, 271
NL General Category value, 212
No (General Category value), 212
no-break space (U+00A0), 122, 265, 400, 412, 476
<nobreak> tag, 231, 476
non-ASCII characters, 530
programming languages and, 536
non-breaking characters, 476
non-breaking hyphen (U+011) character, 265, 421, 449, 476
non-collapsing (space characters), 413
non-ignorable setting for collation elements, 255
non-overrideable properties (normative), 222
non-starter decomposition characters, 242
Noncharacter Code Point (NChar), 218
noncharacter code points, 176–178
normal (non-compatibility) characters, 234
normal (regular) type face, 32
normal form, 237
Normal.dot file, 97
normalization, 237–244, 588
normalization forms, 286
NormalizationCorrections.txt, 216
normative properties, 209, 215, 221
normative references, 286
Norwegian language, 346
not sign (¬), 398
notations, 7, 168
Notepad (Windows), 44, 71, 447
NS (Non Starter) LineBreak property value, 271
NSM (non-spacing mark) Bidi Class value, 260
nt (Numeric Type), 218
NU (Numeric) LineBreak property value, 271
NUL byte (zero byte) convention, 550–551
NUL character, 306, 392
null pointer, 550
number forms block, 429
number sign (#), 374, 383
numbering characters, 39
numbers, 17, 164
as indexes, 41
numeric character references, 109
Numeric Type (nt), 218
Numeric Value (nv), 219
numero sign (№), 411
nv (Numeric Value), 219
O

O with a stroke (Ø, ø), 395
OAlpha (Other Alphabetic), 219
objects (Java), 561
obsolete characters, 181
octet, 4, 66
   UTF-8 ranges, 299
octet sequences, 43
ODI (Other Default Ignorable Code Point), 219
OE (Outlook Express), 54, 146
   non-ASCII characters and, 530
OEM (original equipment manufacturer) code
   pages, 86, 128
offline data, 147
Ogam script code, 337
OGr Ext (Other Grapheme Extend), 219
OIDC (Other ID Continue), 219
OIDS (Other ID Start), 219
OLower (Other Lowercase), 219
OMath (Other Math), 219
ON (other neutrals) Bidi Class value, 260
OP (Opening Punctuation) LineBreak property
   value, 271
open collections, 36
OpenType, 33, 368
Opera browser, 55
Optional Hyphen, 83
original equipment manufacturer (OEM) code
   pages, 86, 128
Orkh script code, 338
orthography, 371
Orya (Orlya) script code, 338
Osma (Osmanya) script code, 338
Other Alphabetic (OAlpha), 219
Other Default Ignorable Code Point (ODI), 219
Other Grapheme Extend (OGr Ext), 219
Other ID Continue (OIDC), 219
Other ID Start (OIDS), 219
Other Lowercase (OLower), 219
Other Math (OMath), 219
Other Uppercase (OUpper), 219
OUpper (Other Uppercase), 219
Outlook Express (OE), 54, 146
   non-ASCII characters and, 530
overlined text, 461
override characters, 261

P

<p> tag (HTML), 469
pair table for line breaking, 278
PanEuropean character set, 36
paragraph breaks (¶), 83
paragraph separator (B) Bidi Class value, 260
paragraph separator (PS), 426, 469
parentheses (()), 376
parenthesized letters, 475
parenthesized numbers, 475
parenthesized symbols, 476
parityfec subtype, 485
paste (Ctrl-V), 91
Pat Syn (Pattern Syntax), 219
Pat WS (Pattern White Space), 219
pattern characters, 586
pattern syntax, 592
Pattern Syntax Characters, 586
Pattern Whitespace Characters, 587
patterns, 585
Pe General Category value, 212
Pd General Category value, 212
PDF (pop directional format) Bidi Class value,
   260
PDF (Portable Document Format), 45
Pe General Category value, 212
percent sign (%), 384
per mille sign (‰), 384
period (.), see full stop
Perl, 526, 556–558
I/O, 574
   regular expressions and, 593
Perm (Old Permic) script code, 338
PF General Category value, 212
Phag script code, 338
Phnx script code, 338
phonemes, 403
phonetic characters, 435
phonetic transcription, 355
PHP, 560
physical records (lines), 425
physical symbols, 397–398
Pi General Category value, 212
pilcrow sign (¶), 400
pipe (|), 144, 389
Pird script code, 338
pl (Polish) ISO 639-1 code, 365
plain text, 44, 193, 446, 482
database files and, 222
design principles and, 156
email, sending, 53
line breaks and, 264
plain subtype for, 485
planes, 171
plus sign (+), 384
catenation operator and, 557
plus-minus sign (±), 398
PO (Postfix) LineBreak property value, 271
Po General Category value, 212
pointers (C), 550
Polish language, 346
ISO 639-1 code, 365
Portable Document Format (PDF), 45
Portuguese language, 346
ISO 639-1 code, 365
POSIX specifications, 603
post composition version precomposed characters, 242
POST method, 579
Postel, Jon, 372
PostScript, 103
PostScript Type 1, 33
pound sign (£), 396, 435
PR (Prefix) LineBreak property value, 271
precomposed characters, 225, 242
preparedness principle, 566–573
primary composite character, 241
primary language codes, 366
printer fonts, 61
private use code points, 177, 180
program commands, 80–88
programming, 213, 535–610
  hyphen-minus and, 419
  line structuring, 424
  markup, generating, 541
  quotation marks and, 385
  regular expressions and, 593
  tildes and, 388
properties, 16, 209–291, 287
case, 244–248
  overview of, 213–223
  summary of, 214–221
property-driven processing, 569–572
PropertyAliases.txt file, 214
PropertyNames.txt file, 214
PropList.txt file, 215
proportional attribute, 32
provisional properties, 214
prs.fallenstein.rst subtype, 485
prs.lines.tags subtype, 485
PS (paragraph separator), 426, 469
Ps General Category value, 212
pt (Portuguese) ISO 639-1 code, 365
punctuation, 326, 395
  General Punctuation block, 411
PuTTY, 507
Q
Q encoding, 529
Qaai (Inherited) script code, 335
Qabx script code, 338
QMark (Quotation Mark), 219
QP (Quoted Printable), 304, 308
  quoted-printable Content-Transfer-Encoding value, 498
QU LineBreak property value, 271
quasi-markup, 448
question mark (?), 384
quotation mark ("), 384
Quotation Mark (QMark), 219
quotation marks (""), 97, 417
quotations, 7
Quoted Printable (QP), 304, 308
  quoted-printable Content-Transfer-Encoding value, 498
R
R (right-to-left) Bidi Class value, 260
Radical, 219
raster fonts, 33
raw format (email), 492
README.TXT files, 490
Real-time Transport Protocol (RTP), 485
recoding, converting character data, 143
recognized identities, 11
record separator (RS) character, 393
records (lines), 425
red subtype, 485
REFC (Reference Close), 539
Reference Close (REFC), 539
Regional and Language Options control panel, 78
registered sign (®), 400
regular (normal) type face, 32
regular expressions (regexp), 593–598
relcom.* groups, 133
rendering/recognizing all Unicode characters, 199
repertoires, 63, 546
ASCII, 118
character, 36–38
characters outside, 454–457
requirements on, 145
replacement characters, 437
replacements on the fly, 93–98
Request for Comments (RFC), 45
reserved code points, 178
reStructuredText markup syntax, 485
reverse solidus (\), 374, 386
RFC (Request for Comments), 45
RFC 1766, 365
RFC 2046, 482
RFC 2822, 496
RFC 3023, 507
RFC 3066, 365
rfc822-headers subtype, 485
Rhaeto-Romanic language, 346
Rich Text Format (see RTF)
richtext subtype, 485
right bracket (]), 376
right curly bracket ({), 376
right-to-left writing, 257
RLE (right-to-left embedding) Bidi Class value, 260
RLE (right-to-left embedding) control character, 261
RLM (right-to-left mark) control character, 261
RLO code (Bidi Class value), 261
Roman numerals, 429, 475
Romanian keyboards, 76
Romanian language, 346
romanization, 351
Roro script code, 338
rows, 172
RS (record separator) character, 393
RTF (Rich Text Format), 45, 104, 447
escape notations in, 103
rtf subtype for, 485
RTP (Real-time Transport Protocol), 485
ruby, 437
Runr (Runic) script code, 338
Russian (ru) ISO 639-1 code, 365

S
S (segment separator) Bidi Class value, 261
S (Simple) rule, 246
SA LineBreak property value, 271
sans serif, 31
Sara script code, 338
Save As option, 46
converting plain text with, 449
Microsoft Word and, 50
Notepad and, 49
Save function, selecting encodings, 49
saving as Unicode, 140
SB (Sentence Break), 219
sc (Script), 219
Sc General Category value, 212
scalar values, 294
scc (Special Case Condition), 220
Scottish Gaelic language, 347
screen, 78
screen fonts, 33
script-specific precomposed characters, 242
scripts, 19, 27, 330–339
unification of, 157
SCSU (Standard Compression Scheme for Unicode), 304
SD (Soft Dotted), 220
search engines, 506
secondary code (second subtag), 366
section sign (§), 400
security threats, using IDNs, 532
selection table, 88
(see also character maps)
self-synchronizing, 295
semantic disambiguation, 203
semantic values, 11
semantics, 155
semicolon (;), 386
in Unicode database files, 222
Sentence Break (SB), 219
sentence text element, 256
SentenceBreakProperty.txt file, 256
sequences of characters, 228
serialization, 137
serif, 31
servers, 508
language negotiation and, 521
sfc (Simple Case Folding), 220
SG * (Surrogates) LineBreak property value, 271
SGML (Standard Generalized Markup Language), 107–110, 539
characters and, 465
sgml subtype for, 485
sharp s (ß), 395
sharp sign (#), 374, 383
Shaw (Shavian) script code, 338
Shift key, 75, 78
shift out (SO) character, 392
shift-trimmed setting for collation elements, 255
shifted setting for collation elements, 255
Shift_JIS encoding for Japanese, 142
shortcuts, defining, 97
“Show Formatting” mode (MS Word), 83
SI (International System of Units), 20, 384
symbols and, 397
SI (shift in) character, 392
SI notations, 430–434
signatures (sig), 529
Simonsen, K., 105
Simple Case Folding (sfc), 220
Simple Lowercase Mapping (slc), 220
Simple Titlecase Mapping (stc), 220
Simple Uppercase Mapping (suc), 220
simplified writing system (Chinese), 369
single quotation mark, 418
single-octet encoding, 47
singleton decomposition characters, 242
Sinh (Sinhala) script code, 338
size of characters, 460
Sk General Category value, 212
slash (/), 374
slc (Simple Lowercase Mapping), 220
Slovak language, 347
Slovenian language, 347
Sm General Category value, 212
<small> tag, 231, 477
smart quotes (“”), 7, 97
SO (shift out) character, 392
So General Category value, 212
Soft Dotted (SD), 220
soft hyphen, 122, 265, 396, 420
“SOH (start of heading) character, 392
solidus (/), 374, 386
division mark and, 397
Sorbian language, 347
sorting, 248–256
sound values of letters, 324
sounds in writing systems, 354
source (email), 492
SP * (Space) LineBreak property value, 272
space between characters, 100
space character (U+0020), 387, 412–417
spacebar, 75
spacing between characters, 460
spacing modifier letters, 434
<span> tag (HTML), 263
Spanish (es) ISO 639-1 code, 365
Spanish keyboards, 76
Spanish language, 347–348
Special block, 437
Special Case Condition (scc), 220
special characters, 69–115, 399
special Characters menu, 83
SpecialCasing.txt file, 218, 245
speech synthesis, 358
spelling checks, 327, 358
spoofing, 332
<square> tag, 231, 477
square brackets ([ ]), 376
stability of design principles, 157
Standard Compression Scheme for Unicode (SCSU), 304
Standard Generalized Markup Language (SGML), 107, 539
characters and, 465
sgml subtype for, 485
standards, 183–192
standards of characters, 16
starter characters, 241
stc (Simple Titlecase Mapping), 220
stdin (standard stream input), 144
stdout (standard stream output), 144
STerm, 220
storage requirements for encodings, 318
straight quotes (“”), 7, 97
strictness of characters, 23
String class (Java), 562
string data, 545–565
string data type, 546
string matching, 357
strings, 8, 182
C programming language and, 548
Java and, 561
JavaScript, 558
Perl and, 556
sorting, 249
structure, 153–208
STX (start of text) character, 392
style attribute (HTML), 544
style elements (HTML), 544
styled text in RTF, 447
<sub> tag, 231
SUB (substitute) character, 392
Subject header, 498
subscripts, 428, 461–464
subtypes of text, 484
<sub> tag (HTML), 477
suc (Simple Uppercase Mapping), 220
Sun Microsystems, 537
<super> tag, 231, 477
superscripts, 395, 428, 461–464
surrogate code points, 178, 179
surrogate pairs, 296
Swedish language, 347
SY (Symbols Allowing Breaks) LineBreak property value, 272
syllabic scripts, 27
Sylo script code, 338
symbols
character, 81
currency, 396
graphic, 13
letterlike, 411
logical, mathematical, and physical, 397–398
Symbols and Punctuation (code charts), 19
SYN (synchronous idle) character, 392
syntax characters, 586
Syrc (Syriac) script code, 338
Syre script code, 338
Syri script code, 338
Syrm script code, 338
system fonts, 33
systematic methods, writing characters, 70
Sámi language, 347

T
T (Turkic) rule, 247
t140 subtype, 485
Tab Separated Values (TSV), 424, 482
tab-separated-values subtype for, 485
table-driven processing, 569–572
tabulated data, 375
Tagb (Tagbanwa) script code, 338
tags (language), 367

tailoring line breaks, 279
Tale (Tai_Le) script code, 338
Talu script code, 338
Tami (Tamil) script code, 338
tc (Titlecase Mapping), 220
TCVN 8-bit codes, 141
<td> tag (HTML), 479
technical symbols, 427
Telnet, 507
Telu (Telugu) script code, 338
Teng script code, 338
Term (Terminal Punctuation), 220
terms (Unicode), 181
TES (see Transfer Encoding Syntax)
TeX, 71, 105
escape notations in, 103
text boundaries, 256
text elements, 182
text media type, 495
text type, 482, 483
Tfng script code, 338
Tglg (Tagalog) script code, 338
Thaa (Thaana) script code, 338
Thai encodings, 149
Thai script code, 338
The Unicode standard, Version 4.0, 183
thin space, 416
thorn, 395
Thunderbird (Mozilla) (see Mozilla Thunderbird)
Tibetan mark delimiter tsheg bstar (U+0F0C), 476
Tibt (Tibetan) script code, 338
tilde (~), 388
titlecase, 244
Titlecase Mapping (tc), 220
toLowerCase Java function, 564
tone marks, 354
Tools menu (MS Word), 94
toTitleCase Java function, 565
toUpperCase Java function, 565
trademark (™), 411
transcoding, 352
tools, 143
transcriptions, 350–357
transfer encoding negotiation, 527
Transfer Encoding Syntax (TES), 136, 137
heading for, 498
Transfer-Encoding header, 498
translations, producing for the web, 524
transliterations, 350–357
troff subtype, 485
TrueType, 33
truncation, 269
TSV (Tab Separated Values), 424, 482
Turkish encodings, 149
Turkish language, 347
two-letter code for General Category values, 211
Type 1, 33
type faces, 32
type text, 482
type-map method (Apache), 522
typeface, 32
typographic discrepancies, 72
typography, 7, 372

U
U+nnnn convention, 19
U.S. International keyboards, 78
UAX (Unicode Standard Annex), 256
  line break rules, 270
uc (Uppercase Mapping), 220
UCA (Unicode Collation Algorithm), 253–256
UCD (Unicode Character Database), 213
UCS (Universal Character Set), 159, 168, 293
UCS Sequence Identifiers (USI), 168
UCS-2, 296, 304
UCS-4, 295, 304
Ugar (Ugaritic) script code, 338
Uldeo (Unified Ideograph), 220
unambiguity, 153–155
unassigned code points, 178, 180, 282
underlined text, 461
underscore (_), 383
underscore (_) in Unicode database files, 223
Unibook, 188
Unicode 1 Name (na1), 218
Unicode 2.0 repertoire, 38
Unicode 4.0.1, 183
Unicode 4.1.0, 184
Unicode algorithms, 287
Unicode Case Charts, 247
Unicode Collation Algorithm (UCA), 253–256
Unicode Consortium, 160
  collation charts and, 250
Unicode Encoded logo, 514
Unicode fonts, 197
Unicode Radical Stroke Count (URS), 220
Unicode scalar values, 294
Unicode Sequence Identifier (USI), 405
Unicode Standard Annex (UAX), 256
  line break rules, 270
Unicode standard annexes, 289
Unicode Technical Note (UTN), 305
Unicode Technical Report (UTR), 136
  normalization vs. folding, 238
Unicode Transformation Format (UTF), 293
Unicode versions, 169, 183
UnicodeBlock.of Java function, 564
Unicodedata.txt file, 191, 213
UnicodeData.txt file
  canonical mapping and, 229
  database files and, 223
unification, 156, 202, 359
unified diacritics, 157
Unified Ideograph (Uldeo), 220
Uniform Resource Locators (see URL)
  uniformity, 155
UniPad, 111
Uniscribe (Microsoft), 42
unit symbols, 434
units of text (characters), 8
Universal Character Set (UCS), 159, 168, 293
universality, 153
Uniview, 210
update version number, 169
Upper property, 220
uppercase letters, 244
  case folding and, 246
  vs. lowercase, 375
Uppercase Mapping (uc), 220
upward arrow head (^) (see circumflex)
uri-list subtype, 485
URL (Uniform Resource Locators), 304, 314–316, 531–533, 539
URS (Unicode Radical Stroke Count), 220
US (unit separator) character, 393
US-ASCII, 118
user characters, 11
user interfaces, 367
USI, 168
USI (Unicode Sequence Identifier), 405
UTF (Unicode Transformation Format), 293
UTF-1, 304, 312
UTF-16, 139, 293, 296, 304 properties of, 297
UTF-16BE, 304
UTF-16LE, 304
UTF-32, 139, 293, 295, 304
UTF-32BE, 304
UTF-32LE, 304
UTF-7, 139, 304, 311
UTF-8, 48, 135, 293, 298–300, 304, 581 cultural biases and, 201 database files and, 222 encoding algorithm, 298 Internet (see Internet) vs. ISO-8859-1, 299
UTF-EBCDIC, 304, 312
UTN (Unicode Technical Note), 305
UTR (Unicode Technical Report), 136 normalization vs. folding, 238
Uuencode, 304, 310
U+0007 (\a) escape notation, 538
U+0008 (\b) escape notation, 538
U+0009 (\t) escape notation, 538
U+000A (LF) (see linefeed)
U+000B (VT) (see vertical tabulation)
U+000B (\v) escape notation, 538
U+000C (FF) (see form feed)
U+000C (\f) escape notation, 538
U+000C (CF) (see form feed)
U+000D (CR) (see carriage return)
U+000D (\r) escape notation, 538
U+0020 space character, 387, 412
U+0021 exclamation mark (!), 380
U+0022 (\") escape notation, 538
U+0022 (\") escape notation, 538
U+0023 number sign (#), 383
U+0024 dollar sign ($), 380, 396, 435
U+0025 percent sign (%), 384
U+0026 ampersand (&), 377
U+0027 (\') escape notation, 538
U+0027 apostrophe ('), 377
U+002A asterisk (*), 377
U+002C comma (,), 379
U+002D (-) (see hyphen-minus)
U+002E full stop (.), 381
U+002F solidus (/), 386
U+003A colon (:), 379
U+003B semicolon (;), 386
U+003C less-than sign (<), 382
U+003D equals sign (=), 380
U+003E greater-than sign (>), 382
U+0040 commercial at (@), 380
U+005C (\/) escape notation, 538
U+005C reverse solidus (/), 386
U+005E circumflex accent (^), 379
U+007E tilde (~), 388
U+0085 next line (NEL), 426
U+00A0 (no-break space), 122, 265, 400, 412, 476
U+00A2 cent sign (¢), 396, 435
U+00A3 pound sign (£), 396, 435
U+00A4 currency sign (¤), 396
U+00A5 yen sign (¥), 397, 435
U+00A6 broken bar (¡), 399
U+00A7 section sign (§), 400
U+00A9 copyright sign (©), 399
U+00AA feminine ordinal indicator (ª), 394
U+00AC not sign (¬), 398
U+00AE registered sign (®), 400
U+00B0 degree sign (°), 397
U+00B1 plus-minus sign (+), 398
U+00B5 micro sign (µ), 397
U+00B6 pilcrow sign (¶), 400
U+00B7 middle dot (.), 399, 431
U+00BA masculine ordinal indicator (º), 394
U+00C6 (Æ), 395
U+00C6 (æ), 395
U+00D0 (Ð), 395
U+00D7 multiplication sign (×), 398, 431
U+00D8 (Ø), 395
U+00D9 (Ø), 395
U+00DE (Þ), 395
U+00DF (ß), 395
U+00F0 (ð), 395
U+00F8 (ø), 395
U+00FE (þ), 395
U+00F0C (Tibetan mark delimiter tsheg bstar), 476
U+2007 (figure space), 476
U+2011 (non-breaking hyphen) (see non-breaking hyphen)
U+2026 (ellipses points) (see ellipsis points)
U+2028 line separator (LS), 426, 469
U+2029 paragraph separator (PS), 426, 469
U+202F (narrow no-break space), 415, 476
U+2030 per mille sign (‰), 384
U+20A0 euro-currency (€), 435
U+20A3 franc sign (₣), 435
U+20AC euro sign (€), 435
U+2103 degree Celsius (°C), 433
U+2116 numero sign (№), 411
U+2122 trademark sign (™), 411
U+212E estimated symbol (₽), 411
U+2215 division slash (/), 432
U+FEFF (zero-width no-break space), 416
U+nnnn convention
database files and, 223
hexadecimal notation and, 41

V
Vaii script code, 338
values (semantic), 11
variable (collation element), 254
variable weighting tag, 255
variables, 48
Variation Selector (VS), 194, 220
variations in shapes of characters, 10
varying-width digits, 375
versions of Unicode, 169, 183
<vertical> tag, 231, 477
vertical forms, 325
vertical line (|), 389
vertical tabulation (VT), 392, 426
vertical writing, 325
video media type, 495
Vietnamese (vi) language
   ISO 639-1 code, 365
Vietnamese language
   encodings, 140, 149
virtual keyboards, 71, 77–80
   UniPad, 111
Virtual Reality Modeling Language (VRML), 494
VISCI 8-bit codes, 141
visible overflow breaks, 269
Visp script code, 339
visual mappings, 247
vowels, 196
VPN 8-bit codes, 141
VRML (Virtual Reality Modeling Language), 494
VS (Variation Selector), 220
VT (vertical tabulation), 392, 426
vulgar fractions, 395, 476

W
“W” versions (Unicode versions), 555
W3C (World Wide Web Consortium), 63
normalization, 243
WAII (Web Accessibility Initiative), 362
WB (Word Break), 220, 290
<wbr> tag (HTML), 269
weak directionality, 261
Web Accessibility Initiative (WAII), 362
web authoring
   entity references and, 107
   fonts, 62
web browsers (see browsers)
Web Embedding Fonts Tool (WEFT), 35
web pages, viewing in different encodings, 54
Welsh language, 347
West European (ISO), 123
Western European encodings, 149
Western European languages, 78
WGL4 (Windows Glyph List 4), 36, 38, 373
whitespace characters, 387, 586
<wide> tag, 231, 477
wide characters, 553
wildcards (regular expressions), 594
Win32 APIs, 554
Windows (Microsoft), 77
   character codes, 126
   Latin 1, 125
Windows character code, 85
Windows Cyrillic, 132
Windows Glyph List 4 (see WGL4)
Windows Latin 1, 125
   French characters and, 349
   Spanish characters and, 348
Windows Latin 1 repertoire, 38
Windows XP, 84
   installing fonts on, 57
windows-1252 encoding, 505
windows-1258 (Windows Vietnamese) 8-bit codes, 141
Wingdings symbols, 134
WinLatin1, 125
WJ (word joiner) character, 195, 272, 416, 479
   line breaking and, 265
Word (see Microsoft Word)
Word Break (WB), 220, 290
word joiner (see WJ)
word processors, 427
   formatting with, 449
   languages and, 326
word text element, 256
WordBreakProperty.txt file, 256
WordPad (Windows), 71, 84
World Wide Web Consortium (W3C), 63
normalization, 243
The World’s Writing Systems (Daniels and
Bright), 26
writing systems, 27, 64, 323–339, 324
WS (whitespace) Bidi Class value, 261
WSpace (White Space), 221
X
XHTML (Extensible HyperText Markup
Language), 109, 511
attributes for, 360
XIDC (XID Continue), 221
XIDS (XID Start), 221
XML (Extensible Markup Language), 71
attributes for, 360–362
character entities, 110
character/visual mappings and, 248
characters and, 465
characters in, 539–543
escape notations in, 103
identifiers, 588–592
notations for characters, 107–110
XHTML and, 511
xml subtypes for, 485
xml-external parsed entity subtype, 485
XO NFC, 221
XO NFD, 221
XO NFKC, 221
XO NFKD, 221
Xpeo script code, 339
Xsux script code, 339
XX LineBreak property value, 272
Y
yen sign (¥), 397, 435
Yiii (Yi) script code, 339
yuan, 397
Z
zero byte (NUL byte) convention, 550–551
zero width joiner/non-joiner, 195
zero-width no-break space (ZWNBSP), 416
zero-width space (ZWSP), 265, 414
zh (Chinese) ISO 639-1 code, 364
Zl General Category value, 212
Zvon database, 187
ZW * LineBreak property value, 272
ZWNBSP (zero-width no-break space), 416
ZWSP (zero-width space), 414
ZWSP (zero-width space) character, 266
Zxxx script code, 339
Zyyy (Common) script code, 335
Zzzz script code, 339