

Index

A

- AAT (Apple Advanced Typography), 675
 - baseline adjustment, 681
 - caret positioning, 681–682
 - glyphs
 - compound, 680
 - selection/placement, 678–682
 - transformations, 679
 - kerning, 680
 - optical alignment, 681
 - tracking, 680–681
- abbreviations, expanding for sorting, 608
- absolute value symbol, 486
- abstract character repertoire, 42, 80, 721
- accent stacking, 721
- accented letters, 15, 721
- accents, 721
 - acute accent, 721
 - grave accent, 752
- accents, sorting and, 604
- accounting numerals, 431–432, 721
- activate Arabic form shaping character, 476
- activate symmetric swapping character, 475
- addak, 316, 721
- additive-multiplicative notation, 430, 433, 722
- additive notation, 429, 722
- al-lakuna, 329, 722
- alef maksura, 281, 722
- algorithmic conversions, 580
- algorithmically derived names, 153–154
- algorithms
 - bi-di, 261–272
 - SCSU, 555–557
- alifu, 295, 722
- alignment, optical (AAT), 681
- allkeys.txt file, 619–623, 722
 - global characteristics for sort, 623
 - mappings, 619–620
 - ordering, collation elements, 621
- alphabet, 722
- alphabetic numerals, 427–428
- Alphabetic Presentations Forms block, 495
- alphabetic property, 164
- alphabets
 - Arabic, 280–290
 - Armenian, 249–251
 - Coptic, 242–243

- alphabets, *continued*
- Cyrillic, 243–249
 - Deseret, 422–423
 - Georgian, 251–253
 - Gothic, 421–422
 - Greek, 237–242
 - Hebrew, 276–279
 - historical, 417
 - IPA (International Phonetic Alphabet), 226–228
 - Japanese, 378–385
 - Latin, 216–219
 - Macedonian, 247
 - Middle Eastern, characteristics, 255–256
 - Mongolian, 407–411
 - Ogham, 419–420
 - Old Italic, 420
 - Runic, 418–419
 - Russian, 245–247
 - Serbian, 247
 - Syriac, 290–294
 - Thaana, 294–296
 - Western, characteristics of, 214–216
- alphasyllabary, 722
- alphasyllabic, 722
- alternate weighting, 624–626, 722
- American National Standards Institute. *See* ANSI
- American Standard Code for Information Interchange. *See* ASCII
- American Standards Association (ASA), 32, 722
- angkhankhu, 332, 723
- annotation techniques, East Asian languages, 400–402
- ano teleia, 723
- ANSI (American National Standards Institute), 32, 722
- ANSI X3.4-1967, 723
- anusvara, 723
- Bengali, 313
 - Devanagari, 306
 - Gujarati, 317
 - Malayalam, 327
 - Sinhala, 329
- apostrophe, 236, 451–452
- applications. *See* software
- Arabic alphabet, 280–285
- alef maksura, 281, 722
 - character forms, 126
 - compared to Mongolian, 407
 - cursive joiners, 286–288
 - font technology, 677
 - hamza, 280, 752
 - hindi numerals, 754
 - kashida, 282, 763, 795
 - line breaking and, 665–666
 - kashida justification, 763
 - lam-alef ligature, 284, 764
 - left-joining letters, 289, 764
 - madda, 281, 768
 - matching abstract to presentation, 678
 - points, 781
 - right-joining letters, 289, 786
 - shadda, 281, 789
 - tatweel, 282, 795
 - teh marbuta, 795
- Arabic block, 285–286
- Arabic contextual shaping properties, 173–174
- Arabic form shaping, 476, 723
- Arabic-Indic numerals, 723
- Arabic numerals, 723
- Arabic percent symbol (%), 485
- Arabic Presentation Forms A block, 127, 290, 495
- Arabic Presentation Forms B block, 126, 288–289, 495
- ArabicShaping.txt file, 141, 677, 723
- Aramaic, 291
- architecture
- BMP (Basic Multilingual Plane), 95–99
 - encoding space, 93
 - glyph model, 62–66
 - planes, 93–95
- Armenian alphabet, 249–251
- ech-yiwn ligature, 250, 743
 - hyphen, 251, 446

- Army (US), FIELDATA code, 32–33
- arrays, 723
- boundary analysis, state machines, 659–662
 - inversion lists, 504–506
 - set operations, 506–512
 - line starts, 652
 - optimizing, 531–533
 - run arrays, 688
- arrow characters, 487
- arrow keys, 692–695
- in bidirectional text, 694
- Arrows block, 493
- articulation marks, 491, 723
- ASA. *See* American Standards Association
- ascenders, 723
- ASCII (American Standard Code for Information Interchange), 8, 724
- 8-bit character encoding, 37–38
 - character representation, 36
 - compatibility with UTF-8, 195–196
 - extended ASCII, 746
 - FIELDATA and, 32–33
 - sorting, 596–597
 - compared to EBCDIC, 35
 - Unicode compatibility, 217–218
- ASCII hex digit property, 150, 164, 724
- asomtavruli, 251, 724
- aspiration, 226–227, 724
- ATSUI (Apple Type Services for Unicode Imaging), 675, 682, 718, 724
- attachment points, 683, 724
- augmentation dot, 491, 724
- B**
- backing store, 724
- backspaces, 724
- making accented letters with, 36, 112
- backward compatibility, 105
- backward levels, 724. *See also* French, accents, sorting and
- Bangla. *See* Bengali alphabet
- banks, 28–29, 725
- bantoc, 337, 725
- baryoosan, 337, 725
- Base64, 725
- MIME and, 710–712
- base characters, 725
- baseline, 725
- adjustment, 681, 725
- basic multilingual plane. *See* BMP
- Baudot, Emile, 28, 725
- Baudot code, 28–29, 725
- BCD (Binary Coded Decimal Information Code), 35, 725
- BCDIC. *See* BCD
- beams, 492, 726
- Becker, Joe, 52–53
- BEL, 30, 726
- Bengali script, 312–314
- issnar, 313, 761
- Bengali block, 314
- beta files, Character Database, obtaining, 140
- bi-di algorithm, 261–262, 726
- directional marks, 267–269
 - embedding levels, 668
 - explicit override characters, 269–270
 - implementing, 667–672
 - inherent directionality, 262–265
 - line breaks, 272
 - mirroring characters, 271
 - neutral directionality, 265–266
 - numbers, 266–267
 - paragraph breaks, 272
- bi-di category, 172–173, 726
- bi-di control property, 149, 164, 726
- bicameral script, 726
- BidiMirroring.txt file, 141, 726
- bidirectional formatting characters, 474–475
- bidirectional layout properties, 172
- bidirectional scripts
- computer representation issues with, 256–261
 - text-editing applications, 272–275
- bidirectional text, 726
- arrow keys and, 694
 - selection drawing and, 695–696

- bidirectional text layout algorithm. *See* bi-di algorithm
- Big5, 50, 726
- big-endian, 192, 726
- Binary Coded Decimal Information Code (BCD). *See* BCD
- binary comparisons, 596, 614, 705, 727
- binary order, sorting and, 599
- binary properties, 149
- binary trees, 530, 727
- bindi, 316, 727
- bit codes, state and, 31
- bit maps, 727
- bit patterns, character representation, 14
- bitmap fonts, 673–674, 727
- bitwise comparisons. *See* binary comparison
- blanked, 727
- Block Elements block, 496
- block storage, 686, 727
- blocks, character assignment, 155–156
 - Arabic, 280–290
 - Armenian, 249–251
 - Bengali, 312–314
 - Bopomofo, 376–378
 - Canadian Aboriginal Syllables, 415–417
 - Cherokee, 413–415
 - CJK Compatibility Ideographs, 368–369
 - CJK Compatibility Ideographs Supplement block, 369
 - CJK Radicals Supplement, 370
 - CJK Symbols and Punctuation, 440
 - CJK Unified Ideographs, 367
 - CJK Unified Ideographs Extension A, 367–368
 - CJK Unified Ideographs Extension B, 368
 - Coptic, 242–243
 - Cyrillic, 243–249
 - Deseret, 422–423
 - Devanagari, 300–312
 - Ethiopic, 411–413
 - General Punctuation, 437–440
 - Georgian, 251–253
 - Gothic, 421–422
 - Greek, 237–242
 - Greek Extended, 242
 - Gujarati, 316–318
 - Gurmukhi, 314–316
 - Hangul Compatibility Jamo, 392
 - Hangul Jamo, 390–391
 - Hangul Syllables, 392–393
 - Hebrew, 276–279
 - Hiragana, 385–386
 - Indic, 297–300
 - IPA (International Phonetic Alphabet), 226–228
 - Japanese, 378–385
 - Kanbun, 386–387
 - Kangxi Radicals, 369–370
 - Kannada, 324–326
 - Katakana, 386
 - Katakana Phonetic Extensions, 386
 - Khmer, 335–338
 - Lao, 333–335
 - Latin, 216–219
 - Latin-1, 219–220
 - Latin Extended A, 220–222
 - Latin Extended Additional, 225–226
 - Latin Extended B, 222–225
 - Malayalam, 326–328
 - Mongolian, 407–411
 - Myanmar, 338–340
 - Ogham, 419–420
 - Old Italic, 420
 - Oriya, 318–319
 - Philippine, 344–345
 - Runic, 418–419
 - Russian, 245–247
 - Serbian, 247
 - Sinhala, 328–330
 - Syriac, 290–294
 - Tamil, 320–323
 - Telugu, 323–324
 - Thaana, 294–296
 - Thai, 331–333
 - Tibetan, 340–344
 - Unified Canadian Aboriginal Syllabics, 417
 - Yi, 402–404
- Blocks.txt, 141, 727

- BMP (Basic Multilingual Plane), 53, 94, 727
 architecture, 95–99
 non-BMP code points, 191
- BOCU (Binary Ordered Compression for Unicode), 207–208, 727
- BOM (byte-order mark), 194, 728
- Bopomofo, 376–378
- bottom-joining vowels, 728
- boundary analysis, 728
 dictionary and, 662–663
 Lao, 662
 line breaking and, 654
 pair tables, implementation and, 657–659
 state machines, implementation and, 659–662
 Thai, 662
- boustrophedon writing, 214, 728
- box-drawing characters, 496, 728
- Boyer, Robert S., 640
- Boyer-Moore algorithm, 728
 searches and, 640–644
 Unicode and, 644–645
- brackets, 454–455
- Brahmi scripts, 297–298
- Braille, 492–493
- breaking characters, 654–655
- breathing marks, 238–239, 728
- breve, 728
- bskur yig mgo, 456
- buffers, line layout, 668
- Buhid script, 344
- bullets, 455–456
- Burmese. *See* Myanmar script
- byte order, 728
- byte-order mark, 194, 728
- C**
- C/C++ programming languages, 714–715
- C programming language, illegal values (planes) and, 101–102
- C0 and C1 areas, 37–38, 39, 728–729
- calligraphy, Arabic alphabet, 282
- Canadian Aboriginal Syllables, 415–417
- candrabindu, 729
 Bengali, 313
 Devanagari, 306
 Gujarati, 317
- canonical accent ordering, 120–123
- canonical composites, 119, 729
 in allkeys.txt file, 621
- canonical composition, normalization, 567–568
 composition exclusion, 571–573
 sequence of Hangul Jamo, 570–571
 single pair of characters, 568–570
 on strings, 573–575
- canonical decomposition, 77, 118–120, 729
 compared to compatibility decomposition, 124
 Hangul, precomposed syllables, 130–131
 Hangul characters, 562–563
 normalization, 559–565
 singleton, 127–128
 supplementary-plane characters, 560–562
- canonical ordering, 79, 729
- canonical reordering, normalization, 563–564
- canonical representation, 729
- canonically equivalent, 729
- cantillation marks, 729
 Arabic, 280
 Hebrew, 278
- capitals, 729
- carets, 730
 insertion caret, 693
 positioning, AAT, 681–682
- caron, 730, 752
- case
 capital, 729
 sorting and, 600–602
- case folding, 730
- case folding, conversion, 588
- case-insensitive comparison, 730
- case mapping, 730
 single characters, 585–587
 strings, 587–588
- cased letters, 730
- CaseFolding.txt, 141, 169–170, 730

- category buffer, line layout, 668
- Cc (general category property), 163, 730
- CCCI (Chinese Character Code for Information Interchange), 50, 730
- CCITT (Consultative Committee for International Telephone and Telegraph), 30, 730
- cedilla, 231, 730
- CESU-8 encoding scheme, 199, 730
- Cf (formatting character), 163, 731
- CGJ (combining grapheme joiner), 731
- character buffer, line layout, 668
- character codes, 731
- Character Database, 84–85, 139, 799
 - obtaining current version, 140
 - structure, 146–147
 - UnicodeData.txt file (Unicode Web site), 145–148
- character encoding form level, 43, 80
- character encoding forms, 185–186, 731, 792
- character encoding scheme level, 43–44, 80
- character encoding schemes, 185–186, 788
- character generator chip, 63, 731
- character-glyph model, 62, 672–673
- character mapping tables, 675, 731
- character names, 152
 - algorithmically derived, 153–154
 - control characters, 154–155
- character-oriented display, 731
- character repertoire, 731
- character semantics, encoding, 83–85
- character set, 731
- character shaping control characters, 464–471
- character strings, comparing, conformity requirements, 110
- characters, 63, 731. *See also* encoding systems
 - alternate selection strategies, 75–76
 - appearance specification, 71
 - bit patterns, 14
 - blocks and scripts, 155–156
 - circled, 125
 - combining character sequences, layout and, 666
 - combining characters, 655
 - comparing, combining classes, 123
 - contracting character sequences in sorting, 605–606
 - drawing, Unicode standard compliance, 61
 - expanding, sorting and, 606
 - formatting bidirectional text, 474–475
 - glue characters, 654–655
 - glyphs, 13
 - Grapheme_Base, 137
 - Grapheme_Extend, 138
 - Grapheme_Link, 137
 - line-breaking properties, 654
 - logical order memory storage, 259–260
 - mapping
 - exception tables, 521–523
 - inversion maps, 512–514
 - multiple character key exception tables, 523–527
 - single to multiple values, 520–521
 - tries as exception tables, 527–530
 - tries as main lookup table, 530–533
 - two-level compact arrays, 519–520
 - mapping to weight value sequences, 606
 - mappings between Unicode and other standards, obtaining, 140
 - noncharacter code points, 100–103
 - number of, Chinese, 45
 - order of, 68–70
 - positioning, 66–70
 - properties, 85
 - general category, 156–163
 - layout, 171–176
 - letters, 166–170
 - normalization, 176–183
 - numerals, 170–171
 - other categories, 163–166
 - representation schemes
 - Baudot code, 28–29
 - FIELDATA, 32–33
 - Hollerith code, 33–34
 - ISO, 36–40

- Morse code, 27
- Murray's code, 27
- by numbers, 14–18
- sorting, 15–16
- terminology, 41–45
- sources of, 71–72
- support for, 106
- tagging, 479–481
- testing for category membership, 501–504
 - inversion lists, 504–506
 - set operations, 506–512
 - visual order memory storage, 259–260
- CharMapML (Character Mapping Markup Language), 731
- charset, 732
- Cherokee alphabet, 413–414
- Cherokee block, 414–415
- Chinese
 - character unification, 74–75
 - characters, 732
 - number of, 45
 - encoding standards, 45–47
 - HZ, 754
 - kangxi radical, 763
 - kangxi Zidian, 763
 - simplified Chinese, 789
 - Traditional Chinese, 797
 - transcribing to Latin alphabet, 223
 - Unicode compared to ISO 10646, 54–55
- Chinese Character Code for Information Interchange, 50
- choon, 732
- choseong, 129, 391, 732
- chu Han, 349, 732
- chu Nom, 358, 732
- Chu Yin, 377
- circled characters, 125
- circled numbers, 435
- circumflex accent, 732
- CJK (Chinese, Japanese, Korean), 732
- CJK Compatibility block, 125–126, 485, 494
- CJK Compatibility Forms block, 126, 495
- CJK Compatibility Ideographs block, 368–369, 495
- CJK Compatibility Ideographs Supplement block, 369
- CJK ideographs, non-BMP code points, 192
- CJK Joint Research Group, 732
- CJK Miscellaneous Area (BMP), 96, 732
- CJK Radicals Supplemental block, 370
- CJK scripts. *See* East Asian alphabets
- CJK Symbols and Punctuation block, 440
- CJK Unified Ideographs, 367–368
 - character names, 154
- CJK Unified Ideographs Extension A, 367–368
- CJK Unified Ideographs Extension A block, 804
- CJK Unified Ideographs Extension B, 368
- CJK Unified Ideographs Extension B block, 804
- CJK_Unified_Ideograph property, 166
- CJKV (Chinese, Japanese, Korean, and Vietnamese), 732
- CJKV scripts. *See* East Asian alphabets
- CJKV Unified Ideographs Area (BMP), 98, 732
- CL area, 733
- closing punctuation, 733
 - general category property, 161
- cmap table (character mapping table), 733
- Cn (general category property), 163
- cn (noncharacter code), 733
- CNS 11643 standard, 50, 733
- Co (general category property), 163
- co (private-use character), 733
- code charts, 733
 - additional information with, 83–84
 - ISO 10646 comments, 155
- code pages, 733
- code points, 43, 732
 - character code, 731
 - collation elements and, 645
 - noncharacters, 100–103, 481–482
 - sorting, 599
 - value, 733
- code set, 733
- code-switching scheme, 733

- code units, 43, 733
- coded character set level, 43, 80, 733
- coeng, 733
- collation element, 607, 733, 734
 - allkeys.txt file, 621
 - alternate weighting and, 624
 - code points and, 645
 - searching and, 639
- collation keys, 734, 790
- collation strength, 734
 - sorting and, 632–635
- combining character sequences, 734
 - advantages and disadvantages, 229–234
 - canonical accent ordering, 120–123
 - caveats, 117–118
 - converting to/from Unicode, 584
 - Devanagari, 310–312
 - Hangul, 129
 - layout and, 666
 - overview, 111–113
 - proper display of, 116
- combining characters, 734
 - line breaking, 655
- combining classes, 122–123, 734
 - properties, 179
- Combining Diacritical Marks block, 228–234
 - Greek alphabet, 240–241
- combining grapheme joiner, 136, 473–474, 734
- combining half mark, 495, 734, 752
- combining marks, 78, 229, 734. *See also*
 - diacritical marks
 - directionality, 264
 - East Asian characters, 395
 - in isolation, 234–235
 - rendering, problems with, 230–234
- combining spacing mark, 734
 - general category property, 159
- compact array, 735
- compact arrays, mapping characters, 514–519
- compatibility
 - backward support, 105
 - issues with, 51–57
 - legacy encoding systems, 73
 - levels of, 647–649
 - Unicode with ASCII, 217–218
 - UTF-8 and ASCII, 195–196
- Compatibility Area (BMP), 99, 735
- compatibility characters, 77–80, 119, 735
- compatibility composites, 79, 735
 - allkeys.txt, 622
 - compatibility decompositions, 124–127
 - W3C and, 702
- compatibility decomposition, 77, 124–127, 735
 - normalization, 565–567
- Comp_Ex (Full_Composition_Exclusion), 734
- compliance with Unicode standard. *See*
 - conformity
- composite character, 735
- composite key, 634, 735
- composition, 735
- composition exclusion, 571–573
 - list, 179–181, 735
- CompositionExclusions.txt (UNIDATA
 - directory), 142, 735
- compound characters, radicals, 356–357
- compound fonts, 684–685, 735
- compound glyphs, 680, 736
 - OpenType, 683
- compression schemes
 - BOCU, 207–208
 - SCSU, 204–207
 - UTF-8, 202–204
- concatenating, files, problems with, 195
- conformity, 736
 - overview, 104
 - support requirements
 - comparing character strings, 110
 - drawing text on output devices, 108–110
 - interpreting text, 107
 - outputting text, 106–107
 - passing text through, 108
 - version support, 104–106
- conjoining jamo (Hangul), 130–131, 736

- conjunct consonants, 736
 Bengali, 313
 Devanagari, 303
 Devanagari block, 309
 Gujarati, 317
 Gurmukhi, 315
 Kannada, 325
 Khmer, 337
 Malayalam, 327
 Myanmar, 339
 Sinhala, 329
 Tamil, 321
 Telugu, 323–324
 Thai, 332
- connector punctuation, 736
- consonant conjunct, 736
- consonant gemination, 281, 736
 Hiragana, 381
- consonant stacks, 341–343, 736
- consonantal, 736
- consonants. *See also* conjunct consonants
 Devanagari, 302–306
 Ethiopic, 412
 Hangul, 388
 Tibetan, 341–342
- Consultative Committee for International
 Telephone and Telegraph (CCITT), 30
- Content-Transfer-Encoding field, MIME,
 710, 736
- context-sensitive weighting, 736
 sorting and, 606–607
 UCA and, 617
- contextual shaping, 64–65, 737
- contours, 737
- contracting character sequences, 737
 allkeys.txt and, 620
 sorting, 605–606
 UCA and, 617
- control-character names, 154–155
- control characters, 460–461, 737
 directionality, 264–265
 general category property, 163
- control code, 30, 737
- control functions, 37, 737
- control picture, 737
- Control Pictures block, 494
- control points, 737
- conversion
- case folding, 588
- case mapping
- single characters, 585–587
- strings, 587–588
- converters, choosing, 584
- line breaks, 584–585
- transliteration, 589–593
- between Unicode and other standards,
 577–578
- algorithmic conversions, 580
- character order, 584
- combining character sequences, 584
- glyph encodings, 584
- handling exceptional conditions,
 581–583
- ISO 2022 encodings, 580
- multibyte encodings, 579
- normalizing, 583
- obtaining conversion information,
 578
- single-byte encodings, 578–579
- stateful encodings, 579–580
- Coptic alphabet, 242–243
- CR area, 37, 737
- CR (carriage return), 29, 457, 737
- CRLF sequence, 737
- as new-line function, 457–459
- grapheme clusters, 136
- line breaking, 654
- Croatian, mapping to Serbian, 223–224
- Croatian diagraphs, 223
- Cs (surrogate character), 163, 738
- currency symbol, 738
- general category property, 162
- currency symbols, 483–484
- cursive, 738
- cursive joiners, 286–288, 738
- Syriac alphabet, 294

- Cyrillic alphabet, 243–247. *See also* Russian alphabet
 variations in, 247
- Cyrillic block, 248–249
- Cyrillic Supplementary block, 249
- D**
- dagesh, 277, 738
- dakuten, 380, 738, 772
 Katakana, 382
- danda, 306
- dash, 738
- dash property, 164
 PropList.txt, 150
- dash punctuation, 738
 general category property, 161
- dasia, 239, 738, 787
- dead keys, 691, 738
- dead languages, 191
- decimal-digit numbers, 738
 general category property, 160
- decimal-digit values, 170, 738
- decomposition, 77, 738–739
 properties, 177
 type, 739
 type properties, 177–178, 739
- default-ignorable code point property, 165, 739
- degree symbol (°), 484
- DEL, 28, 30, 739
- dependent vowels, 302, 739
- deprecated, 739
 characters, 475–476
 property, 150, 166
- derived
 names, 153–154
 normalization properties, 182–183
 property files, 143–144, 739
- derived property, 143, 740
- DerivedAge.txt, 143, 740
- DerivedBidiClass.txt, 143, 740
- DerivedBinaryProperties.txt, 144, 740
- DerivedCombiningClass.txt, 144, 740
- DerivedCoreProperties.txt, 144, 166, 740
- DerivedDecompositionType.txt, 144, 740
- DerivedJoiningGroup.txt, 144, 740
- DerivedJoiningType.txt, 144, 740
- DerivedLineBreak.txt, 144, 740
- DerivedNormalizationProperties.txt, 144, 575, 740
 and case mapping, 631
- DerivedNumericType.txt file, 144, 741
- DerivedNumericValues.txt file, 144, 741
- DerivedProperties.html file, 144, 741
- descenders, 741
- Deseret alphabet, 422–423
- Devanagari
 anusvara, 723
 candrabindu, 729
 eyelash ra, 747
 repha, 785
 stem, 792
- Devanagari block, 307–312
- Dhivehi, 294
- diacritic property, 150, 165
- diacritical marks. *See also* combining marks
 Armenian alphabet, 250
 ASCII, 36
 combining character sequences, 111–113
 double, 123
 Greek alphabet, 238
 Gujarati, 317
 Hebrew, 277
 Hiragana, 380
 in isolation, 234–235
 issues with, 228–234
 noncombining, 235–237
 Oriya, 319
 positioning, 66–70
 Russian, 245–246
 Syriac block, 293
- diaeresis, 741
 compared to umlaut, 218
 Greek alphabet, 239
- dialytika, 239, 741, 797
- dialytika-tonos, 241, 741

- dictionaries
 boundary analysis, 662–663
 Han characters, 356
 Kangxi Zidian, 357
- digit properties, 170–171
 digit values, 170–171, 741
 digits, 741. *See also* numbers
 digraphs, 741
- Croatian, 157, 223–224
 Dutch, 224
 Spanish, 605
- dingbats, 741
 Dingbats block, 496
- diphthongs, 388, 741
- directional marks, 267–269
- directional runs, 263, 667–668, 741
- directionality, 742
 strong, 267, 792
 weak, 267, 806
- discontiguous selections, 695–696, 742
- display cell, 742
- disunify, 742
- document structure, 11
- dot leaders, 455, 742
- dots, 455–456
- double-byte mode (SCSU), 548–549
- double danda, 306, 742
- double diacritics, 123, 742
- Draft Unicode Technical Reports (DUTR),
 86, 89, 742
- drawing, Box Drawing block, 496
- dual-joining letter, 289, 678, 743
- DUTR (Draft Unicode Technical Reports),
 86, 743
- dynamic composition, 743
- E**
- early normalization, 133, 743
- East Asian scripts
 Bopomofo, 376–378
 characteristics, 347–349
 CJK Compatibility Ideographs,
 368–369
- CJK Compatibility Ideographs Supplement
 block, 369
- CJK Radicals Supplemental block, 370
- CJK Unified Ideographs, 367
- CJK Unified Ideographs Extension A,
 367–368
- CJK Unified Ideographs Extension B,
 368
- full-width characters, 393–396
- Han characters, 349–360
 character set development, 363–367
 ideographic description sequence,
 370–372
 unavailable ideographs, display
 alternatives, 370–372
- Hangul Compatibility Jamo block, 392
- Hangul Jamo block, 390–391
- Hangul Syllables area, 392–393
- Hiragana block, 385–386
- Japanese, 378–385
- Kanbun block, 386–387
- Kangxi Radicals block, 369–370
- Katakana block, 386
- Katakana Phonetic Extensions block, 386
- quotation marks, 450–451
- vertical text layout, 396–400
- Yi, 402–404
- East Asian characters, combining marks, 395
- East Asian languages
 annotation techniques, 400–402
 characters, number of, 45
 encoding schemes, 47–51
 encoding standards, 45–47
 input methods, 692
- East Asian Width layout properties, 174–175,
 743
- EastAsianWidth.txt, 142, 743
- EBCDIC (Extended Binary Coded Decimal
 Information Code), 35, 743
- ech-yiwn ligature, Armenian, 250, 743
- ECMA-6, 36, 743
- ECMA-35, 37, 743
- ECMA-94, 38, 744

- ECMA (European Computer Manufacturers' Association), 36, 743
- ECMAScript, 715, 744
- editing software, grapheme clusters and, 134
- editing text, 651, 744
 - arrow keys, 692–695
 - optimization, 685–690
- Egyptian (ancient), 242–243
- elision character, 744
- ellipsis character, 744
 - typographer's ellipsis, 455, 798
- em, 441, 744
- em dash, 445, 744
- em quad, 441, 744
- em space, 441, 744
- embedding levels, 668, 744
- en dash, 444, 744
- en quad, 441, 744
- en space, 441, 744
- Enclosed Alphanumerics block, 125
- Enclosed CJK Letters and Months block, 125, 494
- enclosing marks, 744
 - general category property, 159–160
- encoding forms
 - CESU-8, 199
 - definition, 43
 - transformations between, 536–537
 - UTF-7, 201–202
 - UTF-8, 195–199
 - UTF-16, 189–192
 - UTF-32, 188–189
 - UTF-EBCDIC, 200–201
- encoding space, 43, 745
 - architecture, 93
 - history, 186–187
 - ISO 10646, 187
 - planes, 93–95
 - UTF-32, 188–189
- encoding standard, 745
- encoding systems, 8–10
 - Baudot code, 28–29
 - compatibility issues, 51–57
 - East Asian languages, 45–51
 - EUC (Extended UNIX Code), 48–50
 - FIELDATA code, 32–33
 - Hangul, 129
 - Hollerith code, 33–34
 - HZ, 50
 - ISO, 36–40
 - legacy, compatibility with, 73
 - Morse code, 27
 - Murray's code, 27
 - normalized forms, 132–134
 - punched cards, 34–35
 - Shift-JIS encoding system, 47–48
 - terminology, 41–45
 - vendor-defined, obtaining mappings for, 140
- endian-ness, 192–195, 745
- ending punctuation, 745
- general category, 161
- erotimatiko, 240, 745
- errors, conformity requirements and, 107
- escape sequence, 37–38, 745
- escaping mechanism, W3C and, 701
- Estrangelo, 291, 745
- Ethiopic script, 411–413
 - numbers, 433–434
- Ethiopic block, 413
- ethnic groups, language and, 5
- EUC (Extended UNIX Code), 48–50, 745
- Evans, James, 416
- exception tables, 746
 - key closure, 526
 - mapping characters, 520–527
 - tries as, 527–530
- excluded characters (Normalization Form C)
 - nonstarter decompositions, 180–181
 - post-composition conversion characters, 181
 - script specific, 180
 - singleton decompositions, 180
- expanding characters, 746
 - allkeys.txt and, 620
 - sorting and, 606
 - UCA and, 617
- Expands_On_NFC property, 182, 746

- Expands_On_NFD property, 182, 746
Expands_On_NFKC property, 182, 746
Expands_On_NFKD property, 182, 746
explicit override characters, 269–270, 746
expressions, regular expressions, 646–649, 784
extended ASCII, 746
Extended UNIX Code (EUC), 48–50, 747
Extender property, 150, 165, 747
eyelash ra, 305, 747
- F**
- fancy text, 10, 747
feather marks, 747
feet symbol (´), 485
FF (form feed), 460, 747
FIELDATA code, 32–33, 747
FIGS, 28, 747
figure dash, 444
figure space, 441, 464
files
 concatenating, problems with, 195
 detecting Unicode, 208–209
fili, 295, 747
final form, 283, 747
final-quote punctuation, 747
 general category property, 161
first-series consonants, 336, 747
fixed-width spaces, 441–442
FNC property, 182, 748
fongman, 332, 456, 748
fonts, 748
 Arabic, 677
 ascenders, 723
 bitmap, 673–674, 727
 compound, 684–685
 descenders, 741
 font technologies, 673–675
 glyphs and, 673
 kerning pairs and, 677
 last resort font, 685, 764
 Letterlike Symbols, 125
 outline, 673–674, 778
 PostScript, 674
 proportionally-spaced, 783
 TrueType technology, 674–675
 typeface, 797
 virtual, 684–685, 805
forfeda, 419, 748
format effector, 29, 748
formatting characters, 461–472, 748
 bidirectional text, 474–475
 general category property, 163
 Mongolian, 409–410
 Syriac Abbreviation Mark, 293–294
four-per-em space, 441
fourth-level difference, 748
fourth-level weight, 748
fraction slash, 487–488, 748
fractions, 127
 numbers, 435–436
French, accents
 sorting and, 604
 UCA sorting and, 617
French secondary, 748
FTP (File Transfer Protocol), 706, 748
full-width characters, 393–396
full-width presentation, word breaks and, 656–657
Full_Composition_Exclusion property, 182, 749
furigana, 749
futhark, 418, 749
- G**
- G0 and G1 areas, 37, 749
G source, 367, 749
Gabuli Tana, 295
Gakushu Kanji, 45, 749
gap storage, 686, 749
 polarity switching, 690
 run positions and, 688–689
Garshuni, 291, 749
GB 2312 (Chinese), 46, 750
GDEF table, OpenType, 683
Ge’ez. *See* Ethiopic alphabet
gemination, 750
general category, 156–163, 750

- general category properties
 - letters, 156–159
 - marks, 159–160
 - miscellaneous, 162–163
 - numbers, 160
 - punctuation, 160–161
 - separators, 162
 - symbols, 161–162
- General Punctuation block, 437–440
- General Scripts Area (BMP), 96, 750
- Geometric Shapes block, 496
- Georgian alphabet, 251–253
 - asomtavruli, 251, 724
 - khutsuri, 251, 763
 - mkhedruli, 251, 769
 - nuskhuri, 251, 776
- German
 - expanding characters, 606
 - sorting and, 599
 - umlaut, 799
- geta mark, 371–372, 750
- GL area, 37, 750
- glottal stop, 750
- glue characters, 655, 750
- glyphic variants, 751
- glyphs, 13–14, 750
 - choosing between alternates, 75–76
 - combining character sequences, proper
 - display of, 116–117
 - compound glyphs, 680
 - difference from characters, 62–66
 - fonts and, 673
 - indexes, 675, 750
 - mapping, line breaking and, 665
 - metamorphosis table, 679, 750
 - minimal selection, 676–678
 - selection, 750
 - selection and positioning, 672–685
 - AAT, 678–682
 - OpenType, 682–684
 - selection control characters, 464–471
 - substitution tables, 750
 - OpenType, 682
 - transformations, 679
- Variation Selectors block, 471–472
- Gothic alphabet, 421–422
- GPOS table, OpenType, 683
- GR area, 37, 751
- grapheme clusters, 134–138, 751, 803
 - definition, 135–138
 - properties, 183
 - text rendering, 136–137
 - user-defined, 136
- grapheme joiner, 473–474, 751
- Grapheme_Base property, 137, 183, 751
- Grapheme_Extend property, 138, 183, 751
- Grapheme_Link property, 137, 150, 183, 751
- graphemes, 118, 751
- graphic characters, 752
- Greek alphabet, 237–240
 - ano teleia, 240, 723
 - breathing marks, 238–239, 728
 - dasia, 239, 738, 787
 - dialytika, 239, 741
 - dialytika-tonos, 741
 - erotimatiko, 240, 745
 - koronis, 239, 763
 - monotonic Greek, 239, 770
 - oxia, 238, 778
 - perispomeni, 238, 779
 - polytonic Greek, 239, 781
 - prosgegrammeni, 240, 783
 - psili, 239, 783
 - rough-breathing mark, 239, 787
 - smooth-breathing mark, 239, 790
 - tonos, 239, 796
 - trema, 239, 797
 - varia, 238, 804
 - ypogegrammeni, 240, 808
- Greek block, 240–242
- Greek Extended block, 242
- GSUB table, 682, 752
- guillemet, 447, 752
- Gujarati script, 316–318
- Gujarati block, 318
- Guoyin, 377
- Gurmukhi script, 314–316
 - bindi, 316, 727
 - tippi, 316, 796
- Gurmukhi block, 316

H

- H source, 368, 752
- hacek, 752
- hair space, 442
- halant, 752
- half-form, 304, 752
- half mark, 752
- half-width characters, 393–396, 752
- Half-width and Full-width Forms block, 125, 396, 495
- hamza, 280, 752
- Han characters, 349–360, 753
 - allkeys.txt, 622
 - Bopomofo, 376–377
 - character set development, 363–367
 - geta mark, 371–372
 - ideographic variation indicator, 372
 - Korean, 387
 - as numerals, 430–433
 - ordering of in Unicode standard, 367
 - properties, 183–184
 - radicals, 356–357
 - Ruby (interlinear annotation), 400–402
 - unavailable ideographs
 - display alternatives, 370–372
 - ideographic description sequence, 372–376
 - variant forms, 360–362
 - word breaks, 656–657
- Han unification, 54, 362–367, 753
- handakuten, 380, 753, 769
- Hangul, 50–51, 128–129, 387–390, 753.
 - See also* Korean
 - canonical composition, 570–571
 - canonical decomposition, 562–563
 - character names, 154
 - choseong, 391, 732
 - combining character sequences, 129
 - grapheme clusters, 135
 - jamo, 761
 - categories, 129
 - conjoining Hangul jamo, 736
 - jamo short name, 761
 - johab encoding, 50–51
 - jongseong, 391, 762
 - jungseong, 391, 762
 - precomposed Hangul syllable, 781
 - syllable breaks, 130
- Hangul Compatibility Jamo block, 392, 494
- Hangul Jamo block, 390–391
- Hangul Syllables Area, 98–99, 392–393, 753
- Hangzhou numerals, 433, 753
- hanja, 349, 753
- hankaku, 393, 753
- Hanunóo script, 344, 779
- hanzi, 349, 753
- hard coding, string comparison and, 616
- hard sign (Cyrillic), 245–246, 753
- hardware, video, displaying characters, 63
- headstrokes
 - Bengali, 312
 - Devanagari, 301–302
 - Gujarati, 317
 - Gurmukhi, 314
 - Indic, 753
 - Kannada, 325
 - Oriya, 318
 - Telugu, 323
- Hebrew alphabet, 276–279
 - cantillation mark, 278, 729
 - dagesh, 277, 738
 - pointed Hebrew, 781
 - points, 277, 781
 - rafe, 277, 784
 - sheva, 277, 789
 - shin dot, 277, 789
 - sin dot, 277, 789
- Hebrew block, 279
- hex digit property, 150, 164, 753
- high surrogate, 192, 753
- higher-level protocols, 11–12, 754
- Hindi numerals, 285, 754
- hint, 674, 754
- Hiragana, 379–382, 754, 763
 - palatalized syllables, 380
- Hiragana block, 385–386
- historical alphabets, 191, 417
 - Deseret, 422–423
 - Gothic, 421–422

- historical alphabets, *continued*
 Ogham, 419–420
 Old Italic, 420
 Runic, 418–419
- history
 Arabic alphabet, 280
 ASCII, 36–38
 FIELDATA and, 32–33
 Baudot code, 28–29
 bit codes, state and, 31
 Brahmi scripts, 297–298
 Cherokee alphabet, 413–414
 combining character sequences, 111–113
 encoding space, 186–187
 Ethiopic alphabet, 411–412
 FIELDATA code, 32–33
 Greek alphabet, 237
 Gujarati script, 316–317
 Gurmukhi script, 314
 Han characters, 349–357
 Hangul, 387
 Hebrew alphabet, 276
 Indic scripts, 297–298
 International Telegraphy Alphabet #2
 (ITA2), 30
 Japanese scripts, 378–379
 Latin alphabet, 217
 Mongolian alphabet, 407
 Morse code, 26–27
 Murray's code, 29–31
 punched cards, 33–35
 Syriac alphabet, 290–291
 Tabulating Machine Company, 33
 telegraph, 25–27
 Tibetan script, 340–341
 Western Union code, 30
 Yi syllabary, 403
- hit testing, 754
- HKSCS (Hong Kong Supplemental Character Set), 754
- Hollerith, Herman, 33, 754
- Hollerith code, 33–34, 754
- horizontal bar, 445
- HTML (Hypertext Markup Language), 11, 705–706, 754
- HTTP (Hypertext Transfer Protocol), 705–706, 754
- hyphen property, 164
 PropList.txt, 150, 754
- hyphenation, 754
- hyphens
 line breaking and, 665
 non-breaking, 445, 773
 soft hyphen, 445, 789–790
 line breaking and, 665
- HZ encoding scheme, 50, 754
- I**
- I8 encoding form, 200–201
- IAB (Internet Architecture Board), 755
- IANA (Internet Assigned Numbers Authority), 710, 755
- IBM, 34
- ICU (International Components for Unicode), 547, 558, 587, 716, 755
- IDC. *See* ideographic description character identifiers, 755
 guidelines, 712–713
- ideograph, 755. *See also* Han character ideographic description character, 372–376, 755
- ideographic description sequence, 372–376, 755
- Ideographic property, 150, 755
- Ideographic Rapporteur Group, 54, 368, 755
- ideographic space, 441
- ideographic variation indicator, 372, 755
- ideographs, 349–350, 352
- IDS. *See* ideographic description sequence
- IDS_Binary_Operator property, 150, 166, 756
- IDS_Tertiary_Operator property, 150, 166, 756
- IEC (International Electrotechnical Commission), 756
- IETF (Internet Engineering Task Force), 708, 756
- ignorable characters, 756
 sorting and, 603
 UCA and, 617
 weighting and, 624–625

- illegal values (planes), 100–103
- implementation levels, 230
- inches symbol (′), 485
- independent forms, 283, 756
- independent vowels, 756
 - Devanagari, 303
 - Gurmukhi, 315
- indexing, W3C and, 702
- Index.txt, 142, 756
- Indian Script Code for Information Interchange. *See* ISCII
- Indic scripts, 297–300, 756
 - Bengali, 312–314
 - bottom-joining vowel, 728
 - Devanagari, 300–312
 - double danda, 742
 - Gujarati, 316–318
 - Gurmukhi, 314–316
 - headstroke, 301, 753
 - Kannada, 324–326
 - Khmer, 335–338
 - Lao, 333–335
 - layout, 667
 - left-joining vowels, 764
 - Malayalam, 326–328
 - Myanmar, 338–340
 - nukta, 306, 775
 - Oriya, 318–319
 - right-joining vowels, 786
 - Sinhala, 328–330
 - sorting, 611–613
 - Tamil, 320–323
 - Telugu, 323–324
 - Thai, 331–333
 - Tibetan, 340–344
 - visarga, 306, 805
- Internet Assigned Numbers Authority, 710, 758
- information technology
 - Morse code, 26–27
 - overview, 3–4
 - telegraph, 25–27
- informative elements, 756
- inherent directionality, 262–265, 757
- inherent vowel sounds, 302, 757
- inhibit Arabic form shaping character, 476
- inhibit symmetric swapping character, 475
- initial forms, 283, 757
- initial-quote punctuation, 757
 - general category property, 161
- input
 - arrow keys, 692–695
 - text, accepting, 691–692
- input methods, 691, 757
 - East Asian languages, 692
- insertion caret, 693, 757
- insertion points, 693, 757
- instruction, 674, 754, 757
- interact typographically, 757–758
- interlinear annotation, 476–477, 758
 - Han characters, 400–402
- International Business Machines (IBM), 34
- International Components for Unicode, 547, 558, 587, 758
- International Electrotechnical Commission, 758
- International Organization for Standardization (ISO), 34, 758
- International Phonetic Alphabet (IPA), 226–228, 758
- International Telegraphy Alphabet #2 (ITA2), 30, 758
- internationalization, 6–7, 758
 - problems with, 12–13
- Internet, Unicode and, 699–700
 - HTML, 705–706
 - HTTP, 705–706
 - mail, 708–712
 - URLs, 706–708
 - Usenet, 708–712
 - W3C character model, 700–704
 - XML, 704–705
- Internet Architecture Board, 758
- Internet-draft, 759
- Internet Engineering Task Force, 708, 758–759
- intersection of character sets, testing for, 510–512

- inversion lists, 504–506, 759
 - inversion maps, 512–514
 - set operations, 506–512
 - inversion maps, 512–514, 759
 - inverting case, language-sensitive
 - comparisons and, 629
 - invisible formatting character, 759
 - invisible math operator, 488–489, 759
 - iota subscript, 240
 - IPA (International Phonetic Alphabet), 226–228, 759
 - IRG (Ideographic Rapporteur Group), 54, 368, 759
 - IRV (International Reference Version), 759
 - ISCII (Indian Script Code for Information Interchange), 41, 759
 - Devanagari block, 307
 - ISO 646, 36, 760
 - ISO 2022, 37, 760
 - ISO 6429, 38, 760
 - ISO 8859, 38–40, 760
 - ISO 10646, 52–57, 760
 - code chart comments, 155, 760
 - encoding spaces, 187
 - ISO 8859-1, 38–39, 760–761, 764. *See also* Latin-1
 - ISO-2022-CN, 49, 761
 - ISO-2022-CN-EXT, 49, 761
 - ISO-2022-JP, 49, 761
 - ISO-2022-KR, 49, 761
 - ISO/IEC JTC1/SC2/WG2, 52, 58
 - ISO (International Organization for Standardization), 34, 760
 - ASCII and, 36
 - isolated forms, 283, 761
 - issnar, 313, 761
 - ITA2 (International Telegraphy Alphabet #2), 30, 761
- J**
- J source, 367, 761
 - jamo (Hangul), 388, 761
 - categories, 129
 - Jamo short name, 154, 761
 - Jamo.txt, 142, 761
 - Japanese, 378–385
 - encoding standards, 46
 - Gakushu Kanji, 45, 749
 - geta mark, 371, 750
 - hankaku, 393, 753
 - hiragana, 379–382, 754
 - J source, 367, 761
 - Jinmei-yo Kanji, 45, 761
 - Joyo Kanji, 45, 762
 - kanji, 349, 763
 - katakana, 382–384, 763
 - romaji, 787
 - Ruby (interlinear annotation), 400–402
 - tate-chu-yoko, 399, 795
 - zenkaku, 393, 808
 - Japanese Industrial Standards Commission (JISC), 46, 761
 - jargon, 5–6
 - Java, 713–714
 - implementing SCSU, 550–557
 - intersection function, 510
 - union function, 508
 - Javascript, 715
 - Jinmei-yo Kanji, 45, 761
 - JIS C 6226, 46, 762
 - JIS-Roman, 46, 762
 - JIS X 0201, 41, 762
 - JIS X 0208, 46, 762
 - JIS X 0212, 46, 762
 - JIS X 0213, 46, 762
 - JISC (Japanese Industrial Standards Commission), 46, 762
 - Johab encoding scheme, 50, 391, 762
 - Join_Control property, 149, 762
 - joiners, 762
 - cursive, 286–288
 - Devanagari, 309
 - jongseong, 129, 391, 762
 - Joyo Kanji, 45, 762
 - JScript, 715
 - JTC1 (Joint Technical Committee #1), 52, 762
 - jungseong, 129, 391

- justification, 762
 AAT, 680–681
 kashida justification, Arabic, 763
 line breaking and, 665
- K**
- K source, 367, 762
- Kana, 378–385
 dakuten, 380, 738
 handakuten, 380, 753
 hiragana, 379–382, 754
 katakana, 382–384, 763
 maru, 380, 769
 nigori, 380, 772
- Kanbun block, 386–387
- kangxi radical, 763
- Kangxi Radicals block, 369–370
- Kangxi Zidian, 357, 763
- kanji, 349, 763. *See also* Japanese
- Kannada script, 324–326
- Kannada block, 326
- kashidas, 282, 763
 justification, 763
 line breaking and, 665–666
- Katakana, 379, 382–385, 763
 choon, 732
 half-width characters, 394
- Katakana block, 386
- Katakana Phonetic Extensions block, 386
- kerning, 763
 AAT and, 680
- kerning pairs, font technology and, 677
- key closure, 526, 763
- khan, 337, 763
- Khmer script, 335–338
 bantoc, 337, 725
 bariyoosan, 337, 725
 coeng, 338, 733
 khan, 337, 763
 muusikatoan, 336, 770
 nikahit, 337, 772
 reahmuk, 337, 784
 series-shifter, 336, 788
 triisap, 336, 797
- Khmer block, 338
- khomut, 332, 763
- khutsuri, 251, 763
- killers, 763
- King Sejong, 387
- Komi, 249
- Koranic annotation marks, 286
- Korean. *See also* Hangul
 Hangul, 50–51, 128–129, 387–390, 753
 hanja, 349, 753
 johab encoding, 50, 391, 762
 K source, 367, 762
 language-sensitive comparisons and, 611
 variant Han characters, 361–362
- koronis, 239, 763
- KS C 5601, 763
- KS X 1001, 50, 763
- kunddaliya, 329, 764
- L**
- L.2 (National Committee for Information
 Technology Standards), 58, 764
- labialization, 764
- lam-alef ligatures, 284, 764
- language-sensitive comparisons
 inverting case, 629
 mapping tables, 613–617
 normalization, 629
 optimization, 628–629
 phases, 614–617
 reordering scripts, 630
 sort keys, 626–627
 strings, 595, 630–632
 overview, 596–600
- UCA and, 617–619
- Unicode normalization, 609–611
- language-specific grapheme clusters,
 contracting character sequences and,
 605
- language tagging characters, 479–481
- languages. *See also* alphabets
 information technology and, 3–4
 jargon, 5–6
 spoken, number of, 5

- Lao block, 335
- Lao script, 333–335
 boundary analysis and, 662
 sorting and, 611–613
 word breaks, 656–657
- last resort font, 685, 764
- Latin-1 encoding, 764
 characters, 219–220
 sorting, 597–599
- Latin alphabet. *See also* Western alphabets
 suitability for computer systems, 214
- Latin Extended A block, 220–222
- Latin Extended Additional block, 225–226
- Latin Extended B block, 222–225
- layout properties, 171
 Arabic contextual shaping, 173–174
 bidirectional layout, 172
 East Asian Width, 174–175
 line-breaking, 175–176
 mirroring, 173
- leaders, dot, 455
- left-joining letter (Arabic/Syriac), 289, 678, 764
- left-joining vowels, Indic script, 764
- left-to-right embedding, 474, 764
- left-to-right marks, 267–269, 474, 764
- left-to-right override, 474, 764
- legacy encoding systems, 764
 compatibility with, 73
 EBCDIC, 200–201
- letter-mark combination, 765
- letter number (general category property), 160, 765
- letter shapes (simplified printed Arabic), 283
- Letterlike Symbols block, 125, 485, 493
- letters, 765
 directionality, 262
 general category properties, 156–159
 properties, 166–167
- level buffer, line layout, 668
- level separators, 765
 sort key length, 636–637
- levels
 character encoding, 42–44
 comparison, 600
 implementation, 230
- lexical comparison. *See* binary comparison
- LF (line feed), 29, 457, 765
- ligatures, 13, 765–766
 Arabic Presentation Form A block, 290
 Armenian, 250–251
 lam-alef, 284
 Lao, 334
 Malayalam, 327
 Mongolian, 408
 Oriya, 318
 rendering, 65–66
 Sinhala, 329
 Tamil, 321–322
- line break
 properties, 766
- line breaking, 652–653, 664–666, 766
 bidirectional text, 272
 boundary analysis, 654
 breaking characters, 654–655
 combining characters, 655
 glue characters, 655
 glyph mapping and, 665
 hyphens and, 665
 justification and, 665
 kashidas and, 665–666
 line lengths, 653
 line starts, 653
 properties, 175–176, 653–657
 soft hyphen and, 665
- line layout, 666–672, 766
 bi-di, 667
 buffers, 668
 category buffer, 668
 character buffer, 668
 combining character sequences, 666
 Indic scripts and, 667
 level buffer, 668
 LRE, 669
 LRO, 669

- neutral characters, 670
 - RLE, 669
 - RLO, 669
 - weakly typed characters, 670
 - line separator, 457–459, 766
 - general category property, 162
 - line starts array, 652, 766
 - LineBreak.txt, 142, 766
 - line breaking properties, 654
 - linked lists, 766
 - LI18NUNIX (Linux Internationalization Initiative), 719, 765
 - Linux Standards Base, 719, 766
 - little-endian computer architectures, 193, 766
 - Ll (lowercase letter), 157, 766
 - Lm (modifier letter), 158, 767
 - Lo (other letter), 158–159, 767
 - localization, 767
 - logical characters. *See* graphemes
 - logical order, 68–70, 767
 - logical-order exception property, 151, 166, 767
 - logical order memory storage, 259–260
 - logical selection, 767
 - text-editing software, 274–275
 - logographs, 350, 767
 - long s, 221, 767
 - loops, 674, 767
 - low surrogates, 186, 768
 - lowercase letters, 768
 - general category property, 157
 - lowercase property, 165
 - LRE (left-to-right embedding), 270–271, 474, 669, 768
 - LRM (left-to-right mark), 267–269, 474, 768
 - LRO (left-to-right override), 269–270, 474, 669, 768
 - LS (line separator), 457–459, 768
 - LSB (Linux Standards Base), 719, 768
 - Lt (titlecase letter), 157–158, 768
 - LTRS, 28, 768
 - Lu (uppercase letter), 157, 768
 - LZW (Lempel-Ziv-Welch), 205, 768
- M**
- Macedonian, 247
 - MacOS, 718
 - macrons, 768
 - madda, 281, 768
 - mail, Internet, 708–712
 - mailing lists, Unicode Consortium, 59
 - maintenance (standards), 57–59
 - maiya mok, 332, 768
 - major version numbers, 90, 768
 - Malayalam script, 326–328
 - Malayalam block, 328
 - man'yogana, 378–379
 - mapping characters
 - compact arrays, 514–519
 - exception tables, tries as, 527–530
 - inversion mapping, 512–514, 759
 - main lookup tables, tries as, 530–533
 - multiple character key exception tables, 523–527
 - single characters to multiple values, exception tables, 521–523
 - tables, 675
 - transliteration, 589–593
 - two-level compact arrays, 519–520
 - between Unicode and other standards, obtaining mappings, 140, 579–585
 - W3C technologies, 700
 - to weights, 600–602
 - mapping tables, string comparison, 613–617
 - marks, 769
 - general category properties, 159–160
 - maru, 380, 769
 - math property, 164
 - math symbols, 485–489
 - Mathematical Alphanumeric Symbols block, 489–490
 - mathematical operators, 487
 - mathematical symbols, 769
 - general category property, 162
 - properties, 170–171
 - MathML (Mathematical Markup Language), 704–705, 769

- matra, 302, 769
- Mc (combining spacing mark), 159
- Me (enclosing mark), 159–160
- measurement unit symbols, 484
- medial form, 283, 769
- medium mathematical space, 442
- memory locations
 - bidirectional scripts, issues with, 256–261
 - endian-ness, 192–195
- metamorphosis table, 679, 770
- MICR (Magnetic Ink Character Recognition), 494, 769
- Middle Eastern alphabets
 - Arabic, 280–286
 - Arabic Presentation Forms A block, 290
 - Arabic Presentation Forms B block, 288–289
 - characteristics, 255–256
 - Hebrew, 276–279
 - Syriac alphabet, 290–294
 - Thaana, 294–296
- MIME (Multipurpose Internet Mail Extensions), 709–712, 769
 - Base64, 710–712
 - quoted-printable, 710–712
- minor version numbers, 90, 769
- minus sign, 445
- minutes symbol (′), 485
- mirroring, 69, 769
 - characters, 271
 - properties, 173
- miscellaneous, general category properties, 162–163
- Miscellaneous Symbols block, 494
- Miscellaneous Technical block, 493
- missing glyphs, 769
- mixed-direction text, positioning, 69
- mkhedruli, 251, 769
- Mn (non-spacing mark), 159
- modeling glyphs, 62–66
- Modern Chinese Writing, 349
- modifier letters, 452, 770
 - general category property, 158
 - modifier symbols, 770
 - general category property, 162
- Mongolian alphabet, 247, 407–409
 - compared to Arabic, 407
- Mongolian block, 409–411
- Mongolian Todo soft hyphen, 446, 463
- monospaced fonts, 770
- monosyllabic language, 770
- monotonic Greek, 239, 770
 - 240–241
- Moore, J. Strother, 640
- Morse, Samuel, 26, 770
- Morse code, 26–27, 770
- mort table, 679, 770
- multibyte encodings, converting from/to, 579
- multilevel comparisons, 600–602, 770
- multiple click selections, 696–697
- Murray, Donald, 29, 770
- Murray’s code, 29–31
- musical notation
 - articulation mark, 491, 723
 - augmentation dot, 491, 724
 - beam, 492, 726
 - slur, 492, 790
 - symbols, 490–492
 - tie, 492, 796
- muusikatoan, 336, 770
- Myanmar script, 338–340
- Myanmar block, 340
- N**
- Nagari script. *See* Devanagari script
- names
 - algorithmically derived, 153–154
 - control characters, 154–155
 - standard character, 152
- NamesList.html file, 142, 770
- NamesList.txt, 142, 771
- narrow no-break space, 443, 464
- narrow property, 395, 771
- nasalization, 771
- National Committee on Information Technology Standards (NCITS), 58, 771

- national digit shapes character, 476, 771
- National Phonetic Alphabet, 377
- national-use characters, 36, 771
- Native American languages, 223
- NBSP (non-breaking space), 443, 771
- NCITS (National Committee on Information Technology Standards), 58, 771
- NCR (numeric character reference), 704, 771
- Nd (decimal-digit number), 160, 771
- NEL (new line), 771, 773
- Nestorian, Syriac, 291, 772
- neutral characters, line layout, 670
- neutral directionality, 265–266, 772
- new-line function, 457–459, 772
- newsgroups (Usenet), 708–712
- NFC (Normalized Form C), 132–133, 179–181, 557–558, 772
- NFC_MAYBE property, 182, 772
- NFC_NO property, 182, 772
- NFD (Normalized Form D), 132, 558, 772
- NFD_NO property, 182, 772
- NFKC (Normalized Form KC), 133–134, 558, 772
- NFKC_MAYBE property, 182, 772
- NFKC_NO property, 182, 772
- NFKD (Normalized Form KD), 132, 558, 772
- NFKD_NO property, 182, 772
- nigori, 380, 772
- nikahit, 337, 772
- Nl (letter number), 160, 773
- NL (new line), 457, 773
- NLF (new-line function), 457–459, 773
- no-break space, 443, 464
- No (other number), 160, 773
- nominal digit shapes character, 476, 773
- non-breaking characters, 126–127, 773
- non-breaking hyphen, 445, 464, 773
- non-breaking spaces, 443–444, 773
- non-ignorable characters, weighting and, 624
- non-joiners, 465–471, 774
 - and Arabic, 286–288
 - and Devanagari, 309–310
- non-joining letters (Arabic and Syriac), 289, 678, 774
- non-spacing marks, 113, 774
 - general category property, 159
 - rules governing, 114–117
- non-starter decomposition, 180–181, 774
- nonce form, 363, 487, 773
- noncharacter code points, 100–103, 481–482, 773
- Noncharacter_Code_Point property, 150, 165, 773
- noncognate rule, 364, 773
- noncombining diacritical marks, 235–237
- noninflecting language, 773
- nonstarter decompositions, 180–181, 774
- norm of a matrix symbol, 486
- normalization, 78, 557–559, 774
 - avoiding, language-sensitive comparisons and, 629
 - canonical composition, 567–568
 - composition exclusion, 571–573
 - sequence of Hangul Jamo, 570–571
 - single pair of characters, 568–570
 - on strings, 573–575
 - canonical decomposition, 559–565
 - canonical reordering, 563–564
 - compatibility decomposition, 565–567
 - converting to/from Unicode, 583
 - forms, 558
 - language-sensitive comparisons, 609–611
 - optimizing, 575–576
 - testing, 576–577
 - uniform early normalization, 133, 802
 - W3C, 701–702
 - W3C normalization, 806
- normalization forms, 774
- normalization properties, 176–183
- normalization test file, 181
- NormalizationTest.txt (UNIDATA directory), 142, 774
- Normalized Form C, 132–133, 179–181, 557–558, 774
- Normalized Form D, 132, 558, 775
- Normalized Form KC, 133–134, 558, 775
- Normalized Form KD, 132, 558, 775

- normalized forms, 132–134
 support for, 105
- normative, 775
- not sign symbol, 486
- notation
 additive, 430, 433, 722
 additive-multiplicative, 429, 722
- nukta, 775
 Devanagari, 306
 Gujarati, 317
- NULL character, 29–30, 775
- Number Forms block, 127, 494
- numbers, 426, 775
 alphabetic numerals, 427–428
 Arabic, 285
 Bengali, 313
 bi-di algorithm, 266–267
 character representation, 14–18
 Cherokee, 414
 circled, 435
 currency symbols, 483–484
 Ethiopic, 413, 433–434
 fractions, 435–436
 general category properties, 160
 Gothic, 421
 Greek, 238
 Gujarati, 317
 Gurmukhi, 316
 Han characters as, 430–433
 Hebrew, 278–279
 Kannada, 326
 Khmer, 338
 Korean, 390
 Lao, 334
 Malayalam, 328
 Mongolian, 409
 Myanmar, 340
 numeric presentation forms, 435–436
 Oriya, 319
 parentheses, 435
 periods, 435
 positional notation, 426–427
 Roman, 429–430
 Runic, 418
 Sinhala, 329
 subscripts, 435
 superscripts, 435
 symbols, 482–490
 Tamil, 322, 433
 Telugu, 324
 Thaana, 296
 Thai, 332
 Tibetan, 343, 434
 variable shapes, 436
- number properties, 170–171
- numerals, 775
 accounting numerals, 431–432, 721
 additive-multiplicative, 429, 722
 additive notation, 430, 433, 722
 Hangzhou, 433, 753
 hindi, Arabic, 754
 strings, sorting, 608
 Suzhou, 433, 753
- numeric character references, 704, 776
- numeric code point, string comparison, 596
- numeric presentation forms, 435–436
- numeric punctuation, 483
- numeric values, 171, 776
- nuskhuri, 251, 776
- nyis shad, 342, 776
- O**
- object replacement character, 477–478, 776
- OCR (Optical Character Recognition), 494,
 776
- off-curve point, 776
- Ogham alphabet, 419–420
 forfeda, 748
- ogonek, 776
- Old Italic alphabet, 420
- on-curve points, 674, 776
- onomatopoeia, 382, 776
- opening punctuation, 776
 general category property, 160
 word breaks, 656
- OpenType, 776
 glyph-substitution table, 682
 glyphs

- compound, 683
 - selection and position, 682–684
 - operating systems, Unicode and
 - MacOS, 718
 - UNIX, 718–719
 - Windows, 717
 - optical alignment, 776–777
 - AAT, 681
 - Optical Character Recognition block, 494
 - optimizing
 - arrays, 531–533
 - normalization, 575–576
 - Oriya alphabet, 318–319
 - Oriya block, 319
 - orthographic syllables, 301, 777
 - grapheme clusters, 136
 - other letter (general category property), 158–159, 777
 - other number (general category property), 160, 777
 - other punctuation (general category property), 161, 777
 - other symbol (general category property), 162, 777
 - Other_Alphabetic property, 150, 777
 - Other_Default_Ignorable_Code_Point property, 151, 777
 - Other_Grapheme_Extend property, 150, 777–778
 - Other_Lowercase property, 150, 778
 - Other_Math property, 150, 778
 - Other_Uppercase property, 150, 778
 - out-of-band information, 778
 - outline fonts, 673–674, 778
 - oxia, 238, 778
- P**
- page separators, 459–460
 - pair tables, 778
 - boundary analysis, implementing, 657–659
 - paired punctuation, 454–455, 778
 - paiyannoi, 332, 778
 - palatalization, 245–246, 778
 - palatalized syllables, 380
 - palochka, 248–249
 - pamudpod, 345, 779
 - paragraph breaks, bidirectional text, 272
 - paragraph marks, 11
 - paragraph separators, 457–459, 779
 - general category property, 162
 - parentheses, 454–455
 - numbers, 435
 - Pc (general category property), 161, 779
 - Pd (general category property), 161, 779
 - PDUTR #25, 89, 489, 779
 - PDUTR #28, 89, 91, 779
 - PDUTR (Proposed Draft Unicode Technical Report), 86, 779
 - Pe (general category property), 161, 779
 - per thousand symbol (‰), 485
 - percent symbol (%), 484
 - periods, numbers, 435
 - perispomeni, 238, 779
 - Perl, 715–716
 - Pf (general category property), 161, 779
 - Philippine scripts, 344–345
 - phnaek muan, 456
 - Pi (general category property), 161, 779
 - pictographs, 350–351, 779
 - Pinyin system, 223, 779
 - place-value notation. *See* positional notation
 - plain text, 10–11, 780
 - planes, 53, 93–95, 780
 - illegal values, 100–103
 - Po (general category property), 161, 780
 - pointed Hebrew, 781
 - points (Hebrew alphabet), categories of, 277–278
 - polarity-switching, gap storage and, 690
 - polytonic Greek, 239, 241–242, 781
 - pop directional formatting character, 475, 781
 - positional notation, 426–427, 781
 - positioning
 - characters, 68–70
 - combining character sequences, displaying, 116
 - post-composition version characters, 181, 781

- PostScript, 781
 - fonts, 674
- pound symbol (#), 484
- precomposed characters, 118–120, 781
 - canonical accent ordering, 120–123
- precomposed Hangul syllables, 781
 - canonical decomposition, 130–131
- presentation forms, 79, 494–496, 781
- primary key, 633, 782
- primary-level difference, 782
- primary source standards, 364, 782
- primary weight value, 601, 782
- printing telegraph, 28
- Private Use Area (PUA), 99, 782
- private-use characters, 782
 - general category property, 163
 - W3C and, 701
- Private Use Planes, 95, 782–783
- programming, summary of common
 - operations, 500
- programming languages
 - C, illegal values (planes) and, 101–102
 - Unicode and
 - C/C++, 714–715
 - ICU, 716
 - Java, 713–714
 - Javascript, 715
 - JScript, 715
 - Perl, 715–716
 - Unicode identifier guidelines, 712–713
 - Visual Basic, 715
- programs. *See* software
- properties
 - alphabetic, 164–165
 - ASCII hex digit, 164
 - bi-di control, 164
 - CJK_Unified_Ideograph, 166
 - combining class, 179
 - composition exclusion list, 179–181
 - dash, 164
 - decomposition, 177
 - decomposition type, 177–178
 - default-ignorable code points, 165
 - deprecated, 166
 - derived, 143–145, 740
 - derived normalization, 182–183
 - diacritic, 165
 - digits, 170–171
 - extender, 165
 - general category
 - letters, 156–159
 - marks, 159–160
 - miscellaneous, 162–163
 - numbers, 160
 - punctuation, 160–161
 - separators, 162
 - symbols, 161–162
 - grapheme clusters, 183
 - Han characters, 183–184
 - hex digit, 164, 753
 - hyphen, 164
 - ideographic, 165, 755
 - IDS_Binary_Operator, 166, 756
 - IDS_Tertiary_Operator, 166, 756
 - Join_Control, 164, 762
 - letters, 166–167
 - line breaking, 653–657, 766
 - Logical_Order_Exception, 166, 767
 - lowercase, 165
 - math, 164
 - mathematical symbols, 170–171
 - noncharacter code points, 165
 - normalization, 176–183
 - numerals, 170–171
 - PropList.txt, 149–151
 - Quotation_Mark, 164, 784
 - radical, 166
 - Soft_Dotted, 166, 790
 - standard character names, 152
 - Terminal_Punctuation, 164, 795
 - text rendering layout, 171–176
 - Unicode 1.0 name, 152, 799
 - uppercase, 165
 - White_space, 164, 806–807
 - wide, 807
- PropertyAliases.txt, 142, 783
- PropertyValueAliases.txt, 142, 783
- PropList.html, 142, 783

- PropList.txt, 149–151, 783
 proposals, submitting to Unicode
 Consortium, 59
 Proposed Draft Unicode Technical Report
 (PDUTR), 86, 783
 prosgegrammeni, 240, 783
 protocols, higher-level, 11, 754
 PS (Paragraph Separator), 458, 783
 ps (starting punctuation), 160, 783
 psili, 239, 783
 PUA (Private Use Area), 99, 783
 punched cards, 33–35
 punctuation, 436–437, 784
 apostrophe, 451–452
 Arabic, 286
 Canadian Aboriginal Syllables, 415–417
 CJK Symbols and Punctuation block, 440
 connector punctuation, 736
 dashes, 444–446
 Devanagari, 306
 Ethiopic, 413
 general category properties, 160–161
 General Punctuation block, 437–440
 Georgian, 253
 Greek, 240
 Hebrew, 278
 hyphens, 444–446
 Katakana, 383–384
 Khmer, 337
 modifier letters, 452
 Mongolian, 408–409
 numeric, 483
 paired, 454–455
 Philippine, 345
 positioning, 69
 quotation marks, 446–450
 Runic, 418
 script-specific, 437
 single quotes, 451–452
 Sinhala, 329
 spaces, 440–444
 Syriac, 292
 Syriac block, 293
 Thaana, 296
 Thai, 332
 Tibetan, 342
 vertical text layout and, 397
 word breaks and, 656
 punctuation space, 442
 Punjabi. *See* Gurmukhi script
- Q**
 quotation marks, 446–450
 Japanese vertical text, 397
 property, 164
 Quotation_Mark property, 784
 PropList.txt, 150
 quoted-printable, 784
 MIME and, 710–712
- R**
 radical property, 151, 166
 radicals, 356–357, 404, 784
 rafe, 277, 784
 ReadMe.txt, 143, 784
 reahmuk, 337, 784
 rebus, 353, 784
 referencing, Unicode versions, 91
 regular expressions, 646–649, 784
 rendering, 651, 784
 bidirectional scripts, 256–261
 canonical accent ordering, 122–123
 combining marks, problems with, 230–234
 contextual shaping, 64–65
 Devanagari, 304–306
 headstrokes, issues with, 302
 diacritical marks, 68
 grapheme clusters, 134, 136
 layout properties, 171–176
 line breaking, 652–666
 special-purpose technology, 684
 reordering
 Indic glyphs, 68, 308
 scripts, language-sensitive comparisons, 630
 sorting and, 611–613
 repertoire, Abstract character repertoire, 42,
 785
 repetition marks, 382, 785

- repha, 304, 785
 replacement characters, 478–479, 785
 representative glyphs, 83, 785
 reserved values, 785
 resolved directionality, 785
 RFC 822, 786
 RFC (Request for Comments), 708, 786
 rhotacization, 235–236, 786
 rhotic hook, 235–236
 rich text, 10–11, 747, 786
 right-joining letters, Arabic, 289, 678, 786
 right-joining vowels, Indic script, 786
 right-to-left embedding, 474, 786
 right-to-left marks, 267–269, 474, 786
 right-to-left override, 474, 786
 RLE (right-to-left embedding), 474, 786
 Line layout and, 669
 RLM (right-to-left mark), 267–269, 474, 787
 RLO (right-to-left override), 474, 787
 line layout and, 669
 romaji, Japanese, 787
 Roman numerals, 429–430
 Romanization, 787
 rough-breathing mark, Greek, 239, 787
 round-trip compatibility, 73, 787
 ruby, 400–402, 787
 run arrays, 688, 787
 run-length encoding, 787
 sort keys and, 637–638
 run positions, gap storage and, 688–689
 Runic alphabet, 418–419
 Russian alphabet, 245–247. *See also* Cyrillic
 alphabet
- S**
- SAM (Syriac abbreviation mark), 293–294, 787
 SC2, 52, 787
 Sc (currency symbol), 162, 787
 scalar values, 80, 800
 script-specific excluded characters, 180
 scripts, 787
 Arabic, 280–290
 Armenian, 249–251
 Bengali, 312–314
 Bopomofo, 376–378
 Canadian Aboriginal Syllables, 415–417
 character assignment, 156
 Cherokee, 413–415
 Coptic, 242–243
 Cyrillic, 243–249
 Deseret, 422–423
 Devanagari, 300–312
 Ethiopic, 411–413
 Georgian, 251–253
 Gothic, 421–422
 Greek, 237–242
 Gujarati, 316–318
 Gurmukhi, 314–316
 Han, 349–370
 Hangul, 387–393
 Hebrew, 276–279
 Hiragana, 385–386
 Indic, 297–300
 IPA (International Phonetic Alphabet),
 226–228
 Japanese, 378–385
 Kanbun, 386–387
 Kannada, 324–326
 Katakana, 386
 Khmer, 335–338
 Lao, 333–335
 Latin, 216–226
 Malayalam, 326–328
 Mongolian, 407–411
 Myanmar, 338–340
 Ogham, 419–420
 Old Italic, 420
 Oriya, 318–319
 Philippine, 344–345
 Runic, 418–419
 Russian, 245–247
 Serbian, 247
 Sinhala, 328–330
 Syriac, 290–294
 Tamil, 320–323
 Telugu, 323–324
 Thaana, 294–296
 Thai, 331–333

- Tibetan, 340–344
- Unified Canadian Aboriginal Syllabics, 417
- Yi, 402–404
- Scripts.txt, 143, 787
- SCSU (Standard Compression Scheme for Unicode), 204–207, 546–547, 788
 - algorithms, 555–557
 - implementing, 549–557
 - modes, 548–549
- search routines
 - combining character sequences, 117–118
 - grapheme clusters, 135
- searching, 595, 638–640
 - Boyer-Moore algorithm, 640–644
 - language-sensitive comparison, 595
 - whole word searches, 645–646
- second-series consonants, 336, 788
- secondary keys, 788
 - sorting and, 632–635
- secondary-level difference, 788
- secondary weight value, 601, 788
 - sort key length, 637
- seconds symbol ("), 485
- security, code point value errors, 107
- segment separators, 459–460
- selections
 - discontiguous, 695–696
 - multiple click, 696–697
- separators, 788
 - general category properties, 162
- sequences of characters, contracting, sorting, 605–606
- Sequoyah, 413–414
- Serbian alphabet, 247
 - mapping to Croatian, 223–224
- serialization format, 43, 185, 788
- series-shifter, 336, 788
- Serto, 291, 788
- shad, 342, 788
- shadda, 281, 789
- sheva, 277–278, 789
- Shift-JIS encoding system, 47–48, 789
- shift-trimmed, 789
 - weighting and, 625–626
- shifted, weights, 789
- weighting and, 625
- shin dot, Hebrew, 277, 789
- shortest-sequence rule, 107
- Shou wen jie zi, 356
- SHY, soft hyphen, 445, 462, 789
- simplified Arabic printing, 283–285, 288–289
- Simplified Chinese, 789
 - compared to Traditional Chinese, 360–361
- sin dot, Hebrew, 277, 789
- Sinhala, al-lakuna, 329, 722
- single-byte encodings, converting from/to, 578–579
- single-byte mode (SCSU), 548–549
- single quotes, 451–452
- singleton decomposition, 77, 127–128, 789
- Sinhala script, 328–330
 - kunddaliya, 764
- Sinhala block, 330
- SIP (Supplementary Ideographic Plane), 94, 789
- six-per-em space, 442
- Sk (modifier symbol), 162, 790
- slur, music, 492, 790
- Sm (mathematical symbol), 162, 790
- Small Form Variants block, 126, 495
- small forms characters, 126
- smooth-breathing mark, Greek, 239, 790
- SMP (Supplementary Multilingual Plane), 94, 790
- So (other symbol), 162, 790
- soft hyphen, 445, 462, 790
 - line breaking and, 665
 - Mongolian, 410
- soft signs, 790
 - Russian alphabet, 245–246
- Soft_Dotted property, 151, 166, 790
- case mapping and, 586–587
- software
 - internationalization, 6–7
 - problems with, 12–13
 - sample code, obtaining, 140
 - text-editing, bidirectional text and, 272–275

- sort element, 607
- sort keys, 607, 635–636, 790
 - computing values, 626–627
 - length
 - level separators and, 636–637
 - run-length encoding and, 637–638
 - secondary weight, 637
 - tertiary weight, 637
 - zero-terminated strings, 637
- sorting, 595
 - ASCII compared to EBCDIC, 35
 - ASCII order, 596–597
 - binary order and, 599
 - character representation and, 15–16
 - characters, expanding, 606
 - collation element, 607
 - collation strength, 632–635
 - context-sensitive weighting, 606–607
 - contracting character sequences, 605–606
 - French accents and, 604
 - German, 599
 - grapheme clusters, 134
 - ignorable characters, 603
 - Indic scripts, 611–613
 - language-sensitive comparison, 595
 - Latin-1 encoding, 597–599
 - multilevel comparisons, 600–602
 - reordering, 611–613
 - secondary keys and, 632–635
 - stable sorts, 632
 - surnames, 608
 - Swedish and, 599
 - tailorings and, 618
 - titles, 608
 - unstable sorts, 632
 - weight value and, 600–602
- source separation rule, 364, 790
- sources of characters, 71–72
- space separators, 791
 - general category property, 162
- spaces, 440–444
 - neutral directionality, 265–266
 - non-breaking, 443–444, 773
 - word breaks and, 656
- Spacing Modifier Letters block, 235–237, 452–453
 - Bopomofo, 377
- special characters, 456
- SpecialCasing.txt, 143, 167–169, 586, 791
- split carets, 274, 791
- split vowels, 791
 - Bengali, 313
 - Gujarati, 318
- square form characters, 125–126
- SSP (Supplementary Special-Purpose Plane), 95, 791
- St. Cyril, 243–244
- stability policies, 91
- stable sorts, 632
- standard annexes, 87
- standard character names, 152
- Standard Compression Scheme for Unicode.
 - See* SCSU
- StandardizedVariants.html, 76, 110, 143, 472, 791
- standards
 - character appearance, 71
 - character mappings, obtaining, 140
 - drawing characters, 61
 - ECMA-35, 37
 - ISO
 - 646, 36
 - 2022, 37–38
 - 8859, 38–40
 - ISO 10646, 52–57
 - layers and, 81–82
 - maintenance, 57–59
 - submitting proposals, 59
 - technical reports, 86
 - Unicode
 - conformity with, 104–110
 - length of, 14
 - UnicodeData.txt file, 145–148
 - updating, Unicode compared to ISO 10646, 57
- starting punctuation, 791
 - general category, 160
 - state, bit codes and, 31

- state machines, 791
 boundary analysis implementation,
 659–662
- state tables, 791
- stateful encodings, 792
 converting from/to, 579–580
- stem, 792
- Devanagari, 302
- musical notation, 491
- STIX (Scientific and Technical Information
 Exchange), 486, 792
- storage (computer memory)
 bidirectional scripts, issues with, 256–261
 character encoding forms, 185–186
 gap storage, 686
- storage format, 43, 185, 792
 detecting Unicode formats, 208–209
- string comparisons. *See also* language-
 sensitive comparisons
 hard coding and, 616
 mapping tables, 613–617
 optimization, 628–629
 UCA and, 617–619
- strings, case mapping, 587–588
- strong directionality, 267, 792
- style runs, 688, 792
 polarity switching, 690–691
- styled text, 10–11, 792
- SUB (substitute), 478, 582, 792
- subjoined consonant, Tibetan, 343–344,
 792–793
- subscripting, 124–125
- subscripts, numbers, 435
- sukun, 295, 793
- superscripting, 124–125
- Superscripts and Subscripts block, 494
- supervisory code, 32, 793
- Supplementary Ideographic Plane (SIP), 94,
 793
- Supplementary Multilingual Plane (SMP), 94,
 793
- supplementary-plane characters, canonical
 decomposition, 560–562
- Supplementary Planes, 99–100, 793
- Supplementary Special-Purpose Plane (SSP),
 95, 793
- support, requirements for, 104–106
- surnames, sorting on, 608
- surrogate mechanism, 56, 93, 793
- surrogates, 189–192, 793
- Surrogates Area (BMP), 99, 793
- surrogates (general category property),
 163
- Survey of Character Encodings, 37
- suzhou numerals, 433, 753, 793
- SVG (Scalable Vector Graphics), 704–705,
 793
- Swedish, sorting and, 599
- swung dash, 446
- syllabaries, 722, 794
- Canadian aboriginal, 415–417
- Cherokee, 411–415
- Ethiopic, 411–413
- Japanese, 378–387
- Yi, 402–404
- syllable breaks, Hangul, 130
- syllable clusters. *See* orthographic syllables
- syllables
 Hangul, 389
 Tibetan, 342
- symbols, 794
 Braille, 492–493
 currency, 483–484
 general category properties, 161–162
 math, 485–489
 measurement units, 484
 musical notation, 490–492
 numbers, 482–490
- Symbols Area (BMP), 97, 794
- symmetric swapping, 475, 794. *See also*
 mirroring
- syntax, XML, 705
- Syriac
 alphabet, 290–293
 block, 293–294
 Garshuni, 291, 749
- Syriac Abbreviation Mark, 293–294,
 794

T

- T sources, 794
- Tabulating Machine Company, 33
- tag characters, 479–491, 794
- Tagalog alphabet, 344
- Tagbanwa alphabet, 344
- tailorings, 628, 794–795
 - sorting and, 618
- Taiwan, CNS 11643, 733
- Taiwanese, Big5 standard, 50
- Tamil script, 320–323
 - numbers, 433
- Tamil block, 323
- tate-chu-yoko, Japanese, 399, 795
- tatweels, 282, 795
- TCVN 5712, Vietnamese, 41, 795
- Technical Committee L2, 58
- technical reports, 86, 795
 - old, obtaining, 140
 - Unicode Standard Annexes, 87
- Unicode Technical Standards, 88
- teh marbuta, 281, 795
- telegraph, 25–27, 795
 - printing, 28
- teletypewriter, 28, 795
- Telugu script, 323–324
- Telugu block, 324
- Terminal_Punctuation property, 150, 164, 795
- terminology, 41–45
- ternary trees, 530, 795
- tertiary-level difference, weights, 796
- tertiary weight, 795
- tertiary weight value, sort key length, 637
- testing
 - characters for membership in a class, 501–504
 - inversion lists, 504–506
 - set operations, 506–512
 - normalization, 576–577
- T_EX, 489, 796
- text
 - conformity requirements
 - drawing on output devices, 108–110
 - interpreting, 107
 - outputting, 106–107
 - passing through, 108
 - editing, 651
 - arrow keys, 692–695
 - optimization, 685–690
 - editing software, grapheme clusters and, 134
 - input
 - accepting, 691–692
 - input method, 691
 - rendering software, grapheme clusters and, 134
 - rich, 10–11, 786
 - styled, 10–11, 792
- text blocks, dividing into lines, 652–653
- text-editing applications, bidirectional
 - applications, 272–275
- Thaana alphabet, 294–296
 - alifu, 295, 722
 - fili, 295, 747
 - sukun, 295, 793
- Thaana block, 296
- Thai script, 331–332
 - angkhankhu, 332, 723
 - boundary analysis and, 662
 - fongman, 332, 748
 - khomut, 332, 763
 - maiyamok, 332, 768
 - paiyannoi, 332, 778
 - sorting and, 611–613
 - UCA, 618
 - thanthakhat, 332, 796
 - TIS 620, 41, 796
 - word breaks, 656–657
- Thai block, 333
- thanthakhat, 332, 796
- thin space, 442
- three-per-em space, 441
- Tibetan script, 340–343
 - numbers, 434
 - shad, 342, 788
 - tsheg, 342, 797
 - tsheg bstar, 464
- Tibetan block, 343–344
- tie, music, 492, 796

- tilde, 446, 796
tippi, 316, 796
TIS 620, 41, 796
titlecase letters, 796
 general category property, 157–158
titles, sorting and, 608
tone letters, 236–237, 796
tone marks, 796
 Bopomofo, 377
 Lao, 333
 Myanmar, 340
 Thai, 332
tonos, 239, 796
top-joining vowels, 796
tracking, 797
 AAT, 680–681
Traditional Chinese, 797
 compared to Simplified Chinese,
 360–361
transfer encoding syntax, 44–45, 80, 797
transformation format support, 105
transformations
 between Unicode encoding forms,
 536–537
 UTF-16 and UTF-32, 538–540
transliteration, 589–593, 797
trema, 239, 797
tries, 797
 exception tables, 527–530
 main lookup table, 530–533
triisap, 336, 797
TrueType, 674–675, 797
tsheg, 342, 797
two-level compact arrays, mapping
 characters, 519–520
typeface, 797
typographer's ellipsis, 455, 798
- U**
U source, 367, 798
UAX. *See* Unicode Standard Annexes
UCA (Unicode Collation Algorithm),
617–619, 798
 default sort order, 619–623
UCAS (Unified Canadian Aboriginal
 Syllabics), 417, 798
UCS-2, 187, 798
 compared to UTF-16, 191
UCS-4, 187, 799
umlaut, 799
 compared to diaeresis, 218
unassigned code points (general category
 property), 163, 799
Unicode. *See also* alphabets, blocks
 16-bit character encoding, advantages of, 9
 character encoding levels, 42–45
 compared to other standards, 61
 converting between encoding systems, 9–10
 glyphs, 13
 identifier guidelines, 712–713
 internationalization, 12–13
 Internet and, 699–700
 HTML, 705–706
 HTTP, 705–706
 mail, 708–712
 URLs, 706–708
 Usenet, 708–712
 W3C character model, 700–704
 XML, 704–705
ISO 10646 and, 54–57
need for, 7–6
operating systems and
 MacOS, 718
 UNIX, 718–719
 Windows, 717
plain text, 10–11
programming languages
 C/C++, 714–715
 ICU, 716
 Java, 713–714
 Javascript, 715
 JScript, 715
 Perl, 715–716
 Visual Basic, 715
stability policies, 91
standard
 code charts, 83–84
 conformity with, 104–110

- Unicode, standard, *continued*
 length of, 14
 technical reports, 86
 UnicodeData.txt file, 145–148
 standard maintenance, 57–59
The Unicode Standard Version 3.0, 18–19
 versions, 90–91
 Web site, 19
- Unicode 1.0 name property, 152, 799
- Unicode Character Database. *See* Character Database
- Unicode Collation Algorithm, 617–619, 799
- Unicode Consortium, 57–59, 799
- Unicode mailing list, 59, 799
- Unicode normalization form, 132–134, 800
- Unicode scalar value, 80, 800
- The Unicode Standard, Version 3.0*, 18–19
- Unicode Standard Annexes (UAX), 86, 800
- #9 (bi-di algorithm), 87, 262, 798
- #11, (East Asian width) 87, 394, 798
- #13 (newline guidelines), 87, 459, 798
- #14 (line breaking properties), 87, 654, 798
- #15 (normalization forms), 87, 567, 798
- #19 (UTF-32), 87, 188, 798
- #27 (Unicode 3.1), 87, 91, 798
- #28 (Unicode 3.2), 89, 779
- Unicode Technical Committee (UTC), 58–59, 800
- Unicode Technical Reports (UTR), 86, 800
- #16 (UTF_EBCDIC), 88, 200, 804
- #17 (character encoding model), 42, 88, 804
- #18 (regular expression guidelines), 88, 647, 804
- #20 (Unicode and XML), 88, 703, 804
- #21 (case mapping), 88, 585, 804
- #22 (CharMapML), 88, 578, 804
- #24 (script names), 88, 804
- #25 (mathematics), 89, 489, 779
- #26 (CESU-8), 89, 199, 743
- Unicode Technical Standards (UTS), 86, 801
- #6 (SCSU), 546, 804
- #10 (UCA), 617, 804
- Unicode Transformation Formats (UTFs), 185
- Unicode Web site, 801
- UnicodeCharacterDatabase.html, 143, 801
- UnicodeData.txt, 143, 145–148, 801
- Unicore mailing list, 59, 801
- UNIDATA directory, 141–145
- unification, 71–75, 801
 Han characters, 364–366
- Unified Canadian Aboriginal Syllabics block, 417
- Unified Han repertoire, 366, 801
- Unified Repertoire and Ordering, 367, 802
- Unified_Ideograph property, 151, 802
- uniform early normalization, 133, 802
- unihan, 366, 802
- Unihan.txt, 143, 183–184, 802
- union of character sets, testing for, 507–510
- Uniscribe, 717, 802
- UNIVAC, FIELDATA code, 32
- UNIX, 718–719
 extended UNIX code, 747
- unstable sorts, 632
- update version number, 91, 802
- updating, standards, Unicode compared to
 ISO 10646, 57
- uppercase letters, 802
 general category property, 157
- uppercase property, 165
- URIs (Uniform Resource Identifiers), 706, 802
- URLs (Uniform Resource Locators), 706–708, 802
- URNs (Uniform Resource Names), 706, 802
- URO (Unified Repertoire and Ordering), 367, 802
- US Army, FIELDATA code, 32–33
- Usenet (newsgroups), 708–712
- user characters. *See* grapheme clusters
- user-defined grapheme clusters, 136
- UTC (Unicode Technical Committee), 58–59, 803
- UTF-7, 201–202, 803
- UTF-8, 195–199, 803
 compression schemes, 202–204
 transforming to UTF-16, 546
 transforming to UTF-32, 540–545

- UTF-8-EBCDIC, 200–201, 803
 W3C technologies and, 701
- UTF-16, 56, 803
 encoding form, 189–192
 endian-ness and, 193–194
 transforming to UTF-8, 546
 transforming to UTF-32, 538–540
 UTF-16BE and UTF-16LE, 193–194
 W3C technologies and, 701
- UTF-32, 803
 encoding form, 188–189
 transforming to UTF-8, 540–545
 transforming to UTF-16, 538–540
 UTF-32E and UTF-32LE, 195
- UTF-EBCDIC, 200–201, 803
 (UTFs) Unicode Transformation Formats,
 185
- UTR. *See* Unicode Technical Reports
- UTS. *See* Unicode Technical Standards
- V**
- V source, 368, 804
- Vail, Alfred, 26
- varia, 238, 804
- variation selector characters, 76, 804
- Variation Selectors block, 471–472
- vendor-defined encoding schemes, mappings,
 obtaining, 140
- version numbers, 91, 768
- versions, 90–91
 support for, 104–105
- Vertical Extension A, 804
- Vertical Extension B, 804
- vertical form characters, 126
- vertical text layout, East Asian scripts, 396–400
- video hardware, displaying characters, 63
- Vietnamese, 225–226
 chu nom, 358, 732
 TCVN 5712, 41, 795
- Vietnamese Standard Code for Information
 Interchange (VSCII), 41
- virama, 752, 763, 805
 Bengali, 313
 Devanagari, 303
 Gurmukhi, 315
 Hanunóo, 345
 Lao, 334
 Malayalam, 327
 Sinhala, 329
 Tamil, 320
 Telugu, 323
 Thai, 332
- virtual fonts, 684–685, 805
- visarga, 805
 Devanagari, 306
 Gujarati, 317
- VISCII (Vietnamese Standard Code for
 Information Interchange), 41, 805
- Visual Basic, 715
- visual order, 68–70, 805
 memory storage, 259–260
- visual ordering, 611–613
- visual selection, 805
 text-editing software, 274–275
- vocabularies, sizes of, 359–360
- vowel bearers, 805
- vowel marks, 805
- vowel points, 805
 Hebrew alphabet, 277–278
- vowel signs, 805
- vowels
 Arbic, 280
 Canadian Aboriginal Syllables, 415–417
 Devanagari, 302
 Ethiopic, 412
 Gujarati, 317
 Gurmukhi, 315
 Hangul, 388
 Hebrew, 277–278
 Hiragana, 382
 Katakana, 382
 Malayalam, 326
 Myanmar, 339
 Philippine alphabets, 345
 split vowels, 791
 Syriac, 292
 Tamil, 320
 Thaana, 295

vowels, *continued*

Thai, 331
top-joining, 796

VT (vertical tab), 458, 805
vulgar fractions, 435, 805

W

wave dash, 446
W3C character model, 700–704, 806
 escaping mechanism, 701
 indexing and, 702
 normalization, 701–702
 Unicode private-use characters, 701
 UTF-8, 701
 UTF-16, 701
W3C (World Wide Web Consortium), 700,
 805, 807
wchar_t data type, 13, 101, 714, 806
weak directionality, 267, 792, 806
weakly typed characters, line layout and,
 670
Web sites
 Character Database current version, 140
 software internationalization information,
 13
 Unicode
 derived files, 143
 UNIDATA directory, 141–145
 Unicode Consortium, 84
 conversion information, 578
 Unicode standard, 19
weights, 806
 alternate weighting, 624–626, 722
 collation element and, 607
 context-sensitive weighting, sorting and,
 606–607, 736–737
 ignorable characters, 624–625
 mapping characters to, 600–602
 non-ignorable characters, 624
 pairs of letters, sorting and, 605
 primary weight value, 601
 secondary, 788
 secondary, sort key length, 637
 secondary weight value, 601

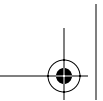
shift-trimmed, 625–626
shifted, 625, 789
sort keys, 607
tertiary, 795
tertiary, sort key length, 637
tertiary-level difference, 796
UCA and, 617
values
 increasing repertoire of available,
 627–628
 reducing size, 627

Western alphabets

Armenian, 249–251
characteristics, 214–216
Coptic, 242–243
Cyrillic, 243–249
Georgian, 251–253
Greek, 237–242
Latin, 216–226
Latin-1 characters, 219–220
Macedonian, 247
Russian, 245–247
Serbian, 247
Western Union code, 30
WG2, 52, 58, 806
White_space property, 149, 164, 806–807
whole word searches, 645–646
wide property, 394, 807
Windows, 717
WJ (word joiner), 195, 444, 463, 807
word breaks, 655
 punctuation and, 656
 spaces and, 656
 ZWSP and, 655
word joiner, 195, 444, 463, 807
word wrapping, 461–464, 807
World Wide Web Consortium. *See* W3C
writing, development of, 350–351
writing systems, *see* scripts
WRU (Who are you?), 807

X

X3.4, 32, 808
x-height, 808



XCCS (Xerox Coded Character Set), 53, 808
XHTML (Extensible Hypertext Markup Language), 11, 704–705, 808
XML (Extensible Markup Language), 11, 704–705, 808
 CharMapML, 731
XSL (Extensible Stylesheet Language), 704–705, 808
Xu Shen, 356, 808

Y

Yi alphabet, 402–403
Yi Area, 98, 808
Yi Radicals block, 404
Yi Syllables block, 404
ypogegrammeni, 240, 808

Z

Zapf Dingbats, 496
zenkaku, 393, 808
zero-terminated strings, sort keys and, 637
zero width joiner, 76, 465–471, 762, 808
zero width non-breaking space, 194–195, 443, 464, 809
zero width non-joiner, 465–471, 774, 809
zero width space, 442, 461, 809
 word breaks, 655
Zhuang, 223
Zhuyin Zimu, 377
Zl (line separator), 162, 809
zone row, 34, 809
Zp (paragraph separator), 162, 809
Zs (space separator), 162, 809

