Characters

Goldsmiths Computing

Motivation

In order to represent natural language, we need to be able to divide it up and represent individual components of text.

Definitions

grapheme cluster roughly, a letter

grapheme smallest meaningful unit in writing in a given language

symbol individual member of an alphabet

code point numeric value assigned to some kind of text unit

character highly context-dependent meaning: could be any of the above

Properties

numeric does the character represent some kind of number? 0, 3, X

lowercase is the character lowercase? a, z

uppercase is the character uppercase? A, Z, Dz

whitespace is the character whitespace?

Character repertoires

ASCII

128 code points

- 10 digits
- 26 lowercase letters
- 26 uppercase letters
- 1 whitespace
- 32 punctuation
- 33 control-codes

Characters in common use in USA

examples 5, e, Z, &, $

Character repertoires

Latin-1

256 code point superset of ASCII: includes everything there and:

- 32 lowercase letters
- 30 uppercase letters
- 1 whitespace
- 33 punctuation
- 32 control-codes

Adds characters useful in Western European languages

examples é, Ç, ÷, £

(but not €)

Character repertoires

Unicode

1114112 code points

- code points [0,1114111]
- (some code points do not correspond directly to characters)

Aims to standardise all human languages and text (*e.g.* Greek, Cyrillic, Arabic, Hebrew, Hangul, Ethiopic, Mongolian, Mathematical operators, Braille, CJK, mediaeval Latin)

examples Θ, Щ, ,א ,☒ ☒, ☒, ☒, ☒, ☒, ☒, ☒

(Klingon and Tengwar out of scope)

Combining characters

- e-acute: U+00E9, é
- a-acute: U+00E1, á
- z-acute: U+017A, ź
- v-acute: U+0076 U+0301, v́

Some characters (grapheme clusters) have multiple representations:

- o-acute: U+00F3 ó or U+006F 0+0301 ó

Work

1. Reading
    - Unicode FAQ: Basic Questions
    - Marcus Kuhn, UTF-8 and Unicode FAQ