L2/09-180

Character Frequency by Unicode Version

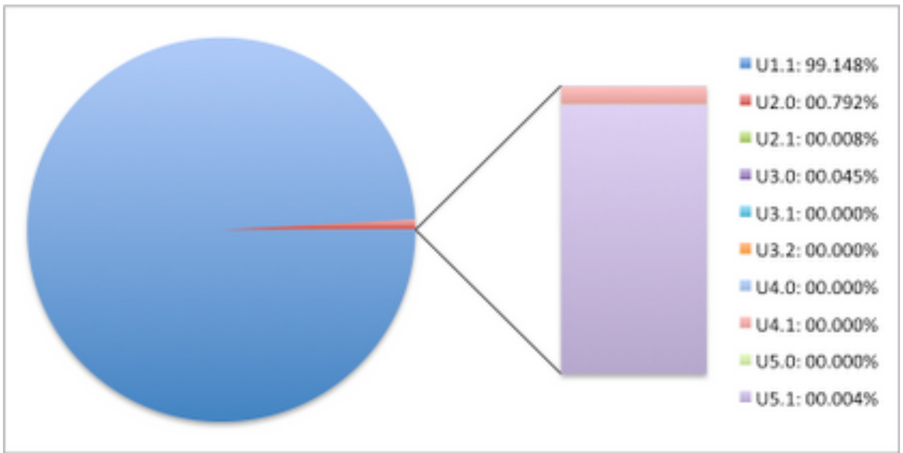
Last updated: May 2, 2009 11:58 AM

This page: http://www.macchiato.com/unicode/character-frequency-by-unicode-version

The image below shows frequency of usage of Unicode characters on the web based on the Unicode version where each character first appeared. One must not read too much into this data; some factors to keep in mind are:

- certain characters occur with much higher frequency in specific languages; if character is low in frequency it doesn't mean it isn't important for some modern orthography.
- there is a certain amount of noise in the data. Below a certain threshold the counts are unreliable because they are more likely to be from garbaged files.
- data is from a sample of the web (a question for the reader: how many pages are there on the web?)
- the characters are not by usage (views); that is, an occurrence of 'é' on Le Monde.fr counts the same as one on Jeau Bleau's Facebook page, even though the former is viewed vastly more often.
- it only include public web pages; in particular it does not include the vast amount of character data in emails and other sources.

For dates of Unicode versions, see http://www.unicode.org/history/publicationdates.html



Below are the numbers, including the ten most frequent characters.

Value	Count	Relative Frequency	Most Frequent Chars
1.1.0.0	25,244	99.148%	e, a, i, o, t, n, r, s, l, d, ...
2.0.0.0	13,400	0.792%	이, 기, 다, 지, 의, 가, 는, 하, 사, 에, ...
2.1.0.0	2	0.008%	€,
3.0.0.0	10,083	0.045%	့, း, ာ, ေ, က, ိ, ပ, ထ, ု, င, ...
3.1.0.0	43,398	0%	𐌹, 𐍃, 𐌰, 𐌴, 𐌽, 𐌲, 𐌺, 𐍄, 𐌳, [illegible], ...
3.2.0.0	904	0%	ϙ, U+2062, ₱, ﷼, ǀ, ǀ, ǀ, ǀ, ⁎, (, ...
4.0.0.0	1,145	0%	[illegible], ३॰, [illegible], ب, ب, ب, ᴍ, ⁚, ...
4.1.0.0	1,108	0%	○, ʕ, [illegible], [illegible], [illegible], [illegible], [illegible], [illegible], [illegible], [illegible], ...
5.0.0.0	1,367	0%	ɕ, [illegible], [illegible], ı, א, [illegible], [illegible], ɘ, [illegible], [illegible], ...
5.1.0.0	1,621	0.004%	̀, [illegible], [illegible], [illegible], ʎ, [illegible], [illegible], ꟈ, ...

By Age | By Script/Category

On the second sheet are numbers by script. If there is no explicit script, the general category is given.

Character Frequency by Script/Category

Value	Count	Relative Frequency	Most Frequent Chars
Latn - Latin	1016	73.771%	e, a, i, o, t, n, r, s, l, d, ...
Nd - Decimal_Number	20	4.883%	0, 1, 2, 3, 5, 9, 4, 8, 6, 7, ...
Cyrl - Cyrillic	403	4.592%	о, а, и, е, н, т, р, с, в, л, ...
Po - Other_Punctuation	103	3.791%	., ,, :, /, U+0027, U+0022, !, ?, , , ۔ , ...
Hani - Han	70362	3.701%	的, 人, 一, 中, 日, 年, 大, 有, 用, 新, ...
Arab - Arabic	264	2.442%	ب, د, ن, ت, و, ر, ي, م, ل, ا, ...
Hira - Hiragana	88	0.92%	の, い, に, し, で, て, す, る, な, た, ...
Hang - Hangul	11412	0.789%	이, 기, 다, 지, 의, 가, 는, 하, 사, 에, ...
Pd - Dash_Punctuation	11	0.768%	-, ⸗, —, ー, —, -, ～, -, ｜, －, ...
Kana - Katakana	108	0.641%	ン, ト, ス, イ, ッ, ル, ク, リ, ラ, ア, ...
Thai - Thai	85	0.611%	า, น, ร, อ, ◌, ก, เ, ม, ง, ◌, ...
Pe - Close_Punctuation	47	0.395%	),], ）, 」, 】, }, 》, 』, ］, 」, ...
Ps - Open_Punctuation	48	0.393%	(, [, （, 「, 【, {, 《, „, 『, ,, ...
Sm - Math_Symbol	899	0.392%	|, >, =, +, <, ~, ～, ｜, →, ×, ...
Grek - Greek	452	0.306%	α, ο, ι, τ, ε, ν, σ, ρ, κ, η, ...
Hebr - Hebrew	88	0.297%	ש, א, ב, מ, ת, ר, ל, ה, ו, ', ...
Deva - Devanagari	99	0.168%	ा, ◌, र, क, ◌, न, ि, स, त, म, ...
Armn - Armenian	84	0.125%	ա, ն, ր, ո, ե, ւ, հ, մ, կ, ս, ...
Taml - Tamil	72	0.118%	, ◌, க, த, ◌, ப, ம, ர, ட, ற, ...
Lm - Modifier_Letter	41	0.114%	⸗, -, ◌, –, ◌, ◌, ◌, ◌, ◌, ◌, ...
Pc - Connector_Punctuation	4	0.112%	_, ⸗, ⸗, ---, ⁀, ---, |, ｛, ‿, ‿
Geor - Georgian	119	0.086%	ა, ი, ე, რ, ს, ო, მ, ბ, ლ, ნ, ...
So - Other_Symbol	1811	0.073%	©, °, , ►, ®, ★, ■, ●, ™, ╞, ...
Pf - Final_Punctuation	10	0.072%	», ', ", ›, ', ◌, ◌, ◌, ◌, ...
Beng - Bengali	88	0.064%	া, র, ে, ি, ন, ◌, ক, ব, ম, ত, ...
Telu - Telugu	93	0.047%	◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ...
Mlym - Malayalam	95	0.042%	◌, ി, ക, ന, ാ, ര, ത, ു, യ, ല, ...
Laoo - Lao	62	0.041%	า, ນ, ', ◌, ກ, ◌, ◌, ◌, ◌, ◌, ...
Knda - Kannada	84	0.039%	◌, ◌, ರ, ಾ, ಎ, ನ, ದ, ಕ, ಗ, ತ, ...
Sc - Currency_Symbol	27	0.039%	$, €, £, ¤, ¥, ¥, ¢, ₪, ฿, $, ...
Mymr - Myanmar	156	0.03%	◌, း, ာ, ေ, ◌, ◌, ◌, ◌, ◌, ...
Pi - Initial_Punctuation	12	0.029%	“, «, ‘, ‹, ‟, ', ‛, ‟, ⊦, `, ...
Gujr - Gujarati	83	0.026%	ા, ર, ન, ◌, ◌, ◌, ચ, ક, ન, વ, ...
Mn - Nonspacing_Mark	483	0.017%	◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ...
Guru - Gurmukhi	73	0.016%	ਾ, ਰ, ◌, ਿ, ੀ, ਨ, ਕ, ਦ, ਹ, ਤ, ...
Sk - Modifier_Symbol	66	0.015%	^, `, ´, ¨, ◌, ◌, ◌, ◌, ◌, ◌, ...
Thaa - Thaana	50	0.005%	◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ...
Khmr - Khmer	146	0.004%	◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ...
Tibt - Tibetan	181	0.004%	◌, ས, ག, ད, ◌, ◌, ◌, ར, ང, ◌, ...
Sinh - Sinhala	80	0.004%	◌, න, ම, ා, ශ, ක, ර, ම, ත, ...
Cf - Format	136	0.004%	U+200E, U+200B, U+FEFF, U+200F, U+202C, U+202A, U+202B, U+202E, ◌, U+202D, ...
Ethi - Ethiopic	461	0.004%	ኀ, ት, ር, ም, የ, ክ, ስ, ው, በ, ል, ...
No - Other_Number	114	0.003%	², ½, ³, ¾, ¼, ¹, ①, ②, ③, ④, ...
Orya - Oriya	82	0.001%	।, ଉ, ◌, ଜ, ୬, ନ, ବ, ସ, ତ, ...
Cans - Canadian_Aboriginal	630	0%	ᐧ, ᓂ, ᖽᐧ, ᑎ, ᓇ, ᐢ, ᓄ, ᐊ, ᐃ, ᐅ, ...
Bopo - Bopomofo	65	0%	ㄌ, ㄧ, ㄚ, ㄟ, ㄓ, ㄩ, ㄨ, ㄌ, ㄋ, ㄥ, ...
Syrc - Syriac	77	0%	◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ◌, ...
Cher - Cherokee	85	0%	Ꭻ, D, Ꮾ, Θ, Ꭰ, Ꭹ, T, Ꮎ, Ᏼ, Ꮮ, ...
...			

Sheet1