What has case distinction but is neither uppercase nor lowercase?

devblogs.microsoft.com/oldnewthing/20241031-00

October 31, 2024

If you go exploring the Unicode Standard, you may be surprised to find that there are some characters that have case distinction yet are themselves neither uppercase nor lowercase.

Oooooh, spooky.

In other words, it is a character *c* with the properties that

- toUpper(*c*) ≠ toLower(*c*), yet
- *c* ≠ toUpper(*c*) and *c* ≠ toLower(*c*).

Congratulations, you found the mysterious third case: Title case.

There are some Unicode characters that occupy a single code point but represent two graphical symbols packed together. For example, the Unicode character dz (U+01F1 LATIN SMALL LETTER DZ), looks like two Unicode characters placed next to each other: dz (U+0064 LATIN SMALL LETTER D followed by U+007A LATIN SMALL LETTER Z).

These diagraphs are characters in the alphabets of some languages, most notably Hungarian. In those languages, the diagraph is considered a separate letter of the alphabet. For example, the first ten letters of the Hungarian alphabet are[1]

a	á	b	c	cs	d	dz	dzs	e	é

These digraphs (and one trigraph) have three forms.

Form	**Result**
Uppercase	DZ
Title case	Dz
Lowercase	dz

Unicode includes four diagraphs in its encoding.

Uppercase	**Title case**	**Lowercase**
DŽ	Dž	dž
LJ	Lj	lj
NJ	Nj	nj
DZ	Dz	dz

But wait, we have a Unicode code point for the dz digraph, but we don't have one for the cs digraph or the dzs trigraph. What's so special about dz?

These digraphs owe their existence in Unicode not to Hungarian but to Serbo-Croatian. Serbo-Croatian is written in both Latin script (Croatian) and Cyrillic script (Serbian), and these digraphs permit one-to-one transliteration between them.[1]

Just another situation where the world is more complicated than you think. You thought you understood uppercase and lowercase, but there's another case in between that you didn't know about.

Bonus chatter: The fact that dz is treated as a single letter in Hungarian means that if you search for "mad", it should not match "madzag" (which means "string") because the "dz" in "madzag" is a single letter and not a "d" followed by a "z", no more than "lav" should match

“law” just because the first part of the letter “w” looks like a “v”. Another surprising result if you mistakenly use a literal substring search rather than a locale-sensitive one. We’ll look at locale-sensitive substrings searches next time.

[1] I got this information from the Unicode Standard, Version 15.0, Chapter 7: “Europe I”, Section 7.1: “Latin”, subsection “Latin Extended-B: U+0180-U+024F”, sub-subsection “Croatian Digraphs Matching Serbian Cyrillic Letters.”