L2/23-062

Add Simple_Case_Folding mappings for three existing characters

Markus Scherer, 2023-mar-02

Proposal

In CaseFolding.txt, add the following three Simple_Case_Folding (scf) mappings:

```
1FD3; S; 0390; # GREEK SMALL LETTER IOTA WITH DIALYTIKA AND OXIA
1FE3; S; 03B0; # GREEK SMALL LETTER UPSILON WITH DIALYTIKA AND OXIA
FB05; S; FB06; # LATIN SMALL LIGATURE LONG S T
```

Rationale

Case foldings are intended for case-insensitive matching. The Simple_Case_Folding is used by some implementations that are defined to always map one code point to one code point, such as in ECMAScript (JavaScript) regular expressions.

For example, we get the following matches: s=S=ſ , ß=ẞ, k=K=K (Kelvin)

However, it is surprising that the following do not match: ΐ ≠ ΐ́, ΰ ≠ ΰ́, ﬅ ≠ ﬆ

In addition, one way of implementing a Simple_Case_Folding transitive closure is to apply a full Case_Folding closure and retain only the simple (1:1) mappings. This works, except for these six characters which have to be exceptionally removed. (In 2022, it was noticed that the Chrome browser implementation needed to make these exceptions: https://crbug.com/v8/13377)

By adding the three proposed scf mappings, we achieve the expected behavior and simplify such implementations.

Note that the Case Folding Stability does not guarantee stability of case foldings of characters which NFKC maps to something else.

Details

Characters mentioned below:

- Basic Latin — Lowercase Latin alphabet
  - s U+0073 LATIN SMALL LETTER S
  - t U+0074 LATIN SMALL LETTER T
- Combining Diacritical Marks — Ordinary diacritics
  - U+0301 COMBINING ACUTE ACCENT
  - U+0308 COMBINING DIAERESIS
- Greek And Coptic — Letter
  - ΐ U+0390 GREEK SMALL LETTER IOTA WITH DIALYTIKA AND TONOS
  - ΰ U+03B0 GREEK SMALL LETTER UPSILON WITH DIALYTIKA AND TONOS
  - ι U+03B9 GREEK SMALL LETTER IOTA
  - υ U+03C5 GREEK SMALL LETTER UPSILON
- Greek Extended — Precomposed polytonic Greek
  - ῒ U+1FD3 GREEK SMALL LETTER IOTA WITH DIALYTIKA AND OXIA
  - ῢ U+1FE3 GREEK SMALL LETTER UPSILON WITH DIALYTIKA AND OXIA
- Alphabetic Presentation Forms — Latin ligatures
  - ﬅ U+FB05 LATIN SMALL LIGATURE LONG S T
  - ﬆ U+FB06 LATIN SMALL LIGATURE ST

Normalization results and case mappings/foldings:

char	cp	NFD	NFKC	scf	cf	proposed scf
ΐ	0390	03B9 0308 0301	0390	-	03B9 0308 0301	-
ΰ	03B0	03C5 0308 0301	03B0	-	03C5 0308 0301	-
ῒ	1FD3	03B9 0308 0301	0390	-	03B9 0308 0301	0390
ῢ	1FE3	03C5 0308 0301	03B0	-	03C5 0308 0301	03B0
ﬅ	FB05	FB05	0073 0074	-	0073 0074	FB06
ﬆ	FB06	FB06	0073 0074	-	0073 0074	-

None of these characters have simple lowercase, titlecase, or uppercase mappings.