February 26, 1998

# Example Tailorings of Collations

Ken Whistler

The following examples illustrate the tailoring rules described in Unicode Technical Report #10.

These examples add a little vendor-specific syntax on top of the basic tailoring operators, using a simple .ini-style presentation.

Examples are chosen to show the typical kinds of tailorings required to deal with cultural-specific orderings for particular languages.

```
*****************************************************************************

; english.sdf
;
; Sort Definition File tailoring for English (and French).
;
; Note that French accent handling is dealt with by a parameter to
; the collation API, rather than by tailoring any individual character
; orders.

[file format]
     version = 1.0.0

[expansion]
     00C6 = 0041 0045 ; Sort AE as equivalent to A + E
     00E6 = 0061 0065 ; Sort ae as equivalent to a + e

*****************************************************************************

; german.sdf
;
; Sort Definition File tailoring for German.
;

[file format]
     version = 1.0.0

[reduction]
     0075 0065 = 00FC ; Sort u+e as equivalent to u-umlaut
     0055 0045 = 00DC ; Sort U+E as equivalent to U-umlaut
     0055 0065 = 00DC ; Sort U+e as equivalent to U-umlaut

     006F 0065 = 00F6 ; Sort o+e as equivalent to o-umlaut
     004F 0045 = 00D6 ; Sort O+E as equivalent to O-umlaut
     004F 0065 = 00D6 ; Sort O+e as equivalent to O-umlaut

     0061 0065 = 00E4 ; Sort a+e as equivalent to a-umlaut
     0041 0045 = 00C4 ; Sort A+E as equivalent to A-umlaut
     0041 0065 = 00C4 ; Sort A+e as equivalent to A-umlaut
```

```
;*****************************************************************************

; swedish.sdf
;
; Sort Definition File tailoring for Swedish.
;

[file format]
      version = 1.0.0

[individual order]
      00C5 >|  005A ; Sort A-ring after Z
      00E5 >|  007A ; Sort a-ring after z
      00C4 >|  00C5 ; Sort A-diaeresis after A-ring
      00E4 >|  00E5 ; Sort a-diaeresis after a-ring
      00D6 >|  00C4 ; Sort O-diaeresis after A-diaeresis
      00F6 >|  00E4 ; Sort o-diaeresis after a-diaeresis
      00DC =   0059 ; Sort U-diaeresis as equivalent to Y (for German names)
      00FC =   0079 ; Sort u-diaeresis as equivalent to y
      0057 >>| 0056 ; Sort W as secondary weight difference after V
      0077 >>| 0076 ; Sort w as secondary weight difference after v

[primary order]
      00C6 |=  00C4 ; Sort ae (and all accented forms) as equivalent to
                    ; a-diaeresis (for Danish names).
      00D8 |=  00D6 ; Sort o-slash (and all accented forms) as equivalent
                    ; to o-diaeresis (for Danish and Norwegian names).

[reduction]
      0041 0041 = 00C5 ; Sort AA as equivalent to A-ring (for Danish names)
      0061 0041 = 00C5 ; Sort Aa as equivalent to A-ring
      0061 0061 = 00E5 ; Sort aa as equivalent to a-ring

;*****************************************************************************

; danish.sdf
;
; Sort Definition File tailoring for Danish.
;

[file format]
      version = 1.0.0

[primary order]
      00C6 |>  005A ; Sort ae (and all accented forms) after z.
      00D8 |>  00C6 ; Sort o-slash (and all accented forms) after ae.

[individual order]
      00C5 >| 00D8 ; Sort A-ring after O-slash
      00E5 >| 00F8 ; Sort a-ring after o-slash

[reduction]
      0041 0041 = 00C5 ; Sort AA as equivalent to A-ring
      0061 0041 = 00C5 ; Sort Aa as equivalent to A-ring
      0061 0061 = 00E5 ; Sort aa as equivalent to a-ring
```

```
**************************************************************************

; czech.sdf
;
; Sort Definition File tailoring for Czech.
;

[file format]
      version = 1.0.0

[individual order]
; Note that for Czech, d-hacek, n-hacek, and t-hacek sort
; with d, n, and t, respectively, with secondary differences
; for the haceks. Those characters do not require tailoring
; from the default values.
; c-hacek, r-hacek, s-hacek, and z-hacek, on the other hand,
; take primary weight distinctions that require tailoring.
      010C >| 0043 ; Sort C-hacek after C
      010D >| 0063 ; Sort c-hacek after c
      0158 >| 0052 ; Sort R-hacek after R
      0159 >| 0072 ; Sort r-hacek after r
      0160 >| 0053 ; Sort S-hacek after S
      0161 >| 0073 ; Sort s-hacek after s
      017D >| 005A ; Sort Z-hacek after Z
      017E >| 007A ; Sort z-hacek after z

[reduction]
; For Czech, "ch" is a primary letter of the alphabet,
; always sorting after h. It represents the IPA value [x].
      0043 0048 >| 0048 ; Sort CH after H
      0043 0068 >| 0048 ; Sort Ch after H
      0063 0068 >| 0068 ; Sort ch after h
```