# Collation

Sharon Correll

# Collation basics

- Wikipedia: “the assembly of written material into a standard order”
- Unicode: UTR #10
- Multiple levels of significant differences
- Unicode defines the DUCET
  - Default Unicode Collation Element Table
  - Associates each character with a multi-level sort key
- Collation sequences are defined using a syntax that starts with the DUCET and extends/modifies it.

# Collation levels

- Latin script
  - (1) Base characters: a < b < c < d < e < f …
  - (2) Diacritics: a << á << à << ä << ã
  - (3) Case: lowercase before uppercase
    - a <<< A < b <<< B < c <<< C …
  - (4) Punctuation
    - E.g., base ball < baseball < base-ball < Baseball < baseballs < baseball's
  - (5) Identical characters

# Collation levels: Latin

0061 ; [.1C47.0020.0002] # LATIN SMALL LETTER A

0041 ; [.1C47.0020.0008] # LATIN CAPITAL LETTER A

0062 ; [.1C60.0020.0002] # LATIN SMALL LETTER B

0042 ; [.1C60.0020.0008] # LATIN CAPITAL LETTER B

0063 ; [.1C7A.0020.0002] # LATIN SMALL LETTER C

0043 ; [.1C7A.0020.0008] # LATIN CAPITAL LETTER C

0064 ; [.1C8F.0020.0002] # LATIN SMALL LETTER D

0044 ; [.1C8F.0020.0008] # LATIN CAPITAL LETTER D

0065 ; [.1CAA.0020.0002] # LATIN SMALL LETTER E

0045 ; [.1CAA.0020.0008] # LATIN CAPITAL LETTER E

# Collation levels: Latin

0061 ; [.1C47.0020.0002] # LATIN SMALL LETTER A

0041 ; [.1C47.0020.0008] # LATIN CAPITAL LETTER A

0062 ; [.1C60.0020.0002] # LATIN SMALL LETTER B

0042 ; [.1C60.0020.0008] # LATIN CAPITAL LETTER B

0063 ; [.1C7A.0020.0002] # LATIN SMALL LETTER C

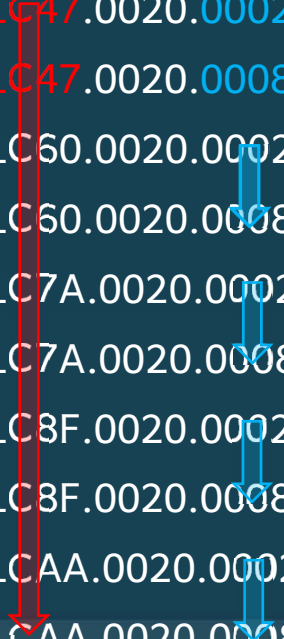
0043 ; [.1C7A.0020.0008] # LATIN CAPITAL LETTER C

0064 ; [.1C8F.0020.0002] # LATIN SMALL LETTER D

0044 ; [.1C8F.0020.0008] # LATIN CAPITAL LETTER D

0065 ; [.1CAA.0020.0002] # LATIN SMALL LETTER E

0045 ; [.1CAA.0020.0008] # LATIN CAPITAL LETTER E



# Collation levels: Latin

0061 ; [.1C47.0020.0002] # LATIN SMALL LETTER A

1D68A ; [.1C47.0020.0005] # MATHEMATICAL MONOSPACE SMALL A

24D0 ; [.1C47.0020.0006] # CIRCLED LATIN SMALL LETTER A

0041 ; [.1C47.0020.0008] # LATIN CAPITAL LETTER A

FF21 ; [.1C47.0020.0009] # FULLWIDTH LATIN CAPITAL LETTER A

24B6 ; [.1C47.0020.000C] # CIRCLED LATIN CAPITAL LETTER A

2090 ; [.1C47.0020.0015] # LATIN SUBSCRIPT SMALL LETTER A

0062 ; [.1C60.0020.0002] # LATIN SMALL LETTER B

0042 ; [.1C60.0020.0008] # LATIN CAPITAL LETTER B

0063 ; [.1C7A.0020.0002] # LATIN SMALL LETTER C

# Collation levels: Latin

0061 ; [.1C47.0020.0002] # LATIN SMALL LETTER A

00E1 ; [.1C47.0020.0002][.0000.0024.0002] # LATIN SMALL LETTER A WITH ACUTE

00E0 ; [.1C47.0020.0002][.0000.0025.0002] # LATIN SMALL LETTER A WITH GRAVE

00E3 ; [.1C47.0020.0002][.0000.002D.0002] # LATIN SMALL LETTER A WITH TILDE

24D0 ; [.1C47.0020.0006] # CIRCLED LATIN SMALL LETTER A

0041 ; [.1C47.0020.0008] # LATIN CAPITAL LETTER A

0062 ; [.1C60.0020.0002] # LATIN SMALL LETTER B

0042 ; [.1C60.0020.0008] # LATIN CAPITAL LETTER B

212C ; [.1C60.0020.000B] # SCRIPT CAPITAL B

0063 ; [.1C7A.0020.0002] # LATIN SMALL LETTER C

# Other collation options

- Treat two characters as one

  09CB ; [.26FC.0020.0002] # BENGALI VOWEL SIGN O

  09C7 09BE ; [.26FC.0020.0002] # BENGALI VOWEL SIGN O

- Treat one character as two

  004F ; [.1DDD.0020.0008] # LATIN CAPITAL LETTER O

  0152 ; [.1DDD.0020.000A][.0000.0110.0004][.1CAA.0020.000A]
  # LATIN CAPITAL LIGATURE OE

- Ignore characters – Arabic honorifics, number sign
  - Use [.0000.0000.0000]
- Backwards accent ordering (needed for French)
- Sort longer words first

# Collation levels

- Arabic script
  - (1) Base characters:
    - ا < ب < ﺑ < پ < ﭘ < ﭙ < ت < ث < ...
  - (2) Vowel marks:
    - ً << ٌ << ٍ << َ << ُ << ِ << ْ << ّ << ٰ << ٓ
  - (3) Variations, ligatures
    - Not used extensively

# Collation levels

- Arabic script
  - Level (3)
    - Variations
      - hamza <<< high hamza, alef <<< low alef
      - waw <<< small waw, yeh <<< small yeh

```
0627  ; [.230B.0020.0002] # ARABIC LETTER ALEF
08AD  ; [.230B.0020.0004] # ARABIC LETTER LOW ALEF
```

    - Ligatures
      - sad <<< “sallallahou alayhe wasallam”

```
0635  ; [.2364.0020.0002] # ARABIC LETTER SAD
FDFA  ; [.2364.0020.001A][.239C.0020.001A][.23C5.0020.001A]...
              # ARABIC LIGATURE SALLALLAHOU ALAYHE WASALLAM
```

# Customizing collations

- Syntax to extend/modify DUCET sort keys
- Possible needs:
  - Language-specific
  - Punctuation
  - Merged tailorings: define how to sort several languages together
  - Numbers: treat them as alphabetical or numeric
- Can be included in LDML files, as a block of text

```
<collation type="standard">
        <cr><![CDATA[
                &B<t<<<T<s<<<S<e<<<E
                &C<k<<<K<x<<<X<i<<<I
                &D<q<<<Q<r<<<R
                &G<o<<<O
                &W<h<<<H
        ]]></cr>
</collation>
```

# Customizing collations

- &
  - Reset – the following character is assigned its default place from the DUCET
- <
  - Primary level ordering: &a < b < c < x < d
- <<
  - Secondary level ordering: &á << å << à
- <<<
  - Tertiary level ordering: &a <<< A
- =
  - Identical: &s < sh = ∫ < t

# Customized collation specifications

- Base character: primary level

  &پ < ق < پ

- Vowel marks: secondary level

  &ِ << ً << ٰ << ّ

  - Ambiguity! What about curly kasra and kasra with dot below?

    &ِ << ً << ٰ << ّ << ٖ << ٍ

- Honorific marks: tertiary level

  &ؐ <<< ؑ <<< ّ

# Optional exercise

- Latin and Arabic exercises
- DUCET subset plus tailoring rules
- Solutions are given at the end