Janusz S. Bień

Informal notes on Unicode and MUFI in Emacs

January 3, 2021

1. Introduction

To make a long story short, we distinguish here three groups of characters

— unassigned

— nonstandard (unofficial)

— standard (official)

We will focus here on a set of nonstandard characters provided by the recommendation of Medieval Unicode Font Initiative.

2. The Unicode Character Database

The Unicode Character Database includes in particular `UnicodeData.txt` defining 12 character properties. It is a CVS file, for the time being we are interested only in the first two fields:

1. the code point
2. the name

The code point is obvious.

Not all Unicode characters have names, the unassigned characters do not have them for obvious reasons. Some characters (e.g. kanji) have names generated algorithmically. Hence the name field can be empty, can contain just a string (the name) or a string enclosed in angle brackets, i.e. either a reference to a name generation rule or just a comment. For example

```
0000;<control>;Cc;0;BN;;;;;N;NULL;;;;
...
0020;SPACE;Zs;0;WS;;;;;N;;;;;
...
20000;<CJK Ideograph Extension B, First>;Lo;0;L;;;;;N;;;;;
2A6DD;<CJK Ideograph Extension B, Last>;Lo;0;L;;;;;N;;;;;
```

3. The Unicode Character Database in Emacs

Emacs source contains the file `admin/notes/unicode` entitled *Importing a new Unicode Standard version into Emacs*, which describes the process. Some useful informations are also provided by Eli Zaretskii in his comments to my feature request (`https://debbugs.gnu.org/cgi/bugreport.cgi?bug=32599`).

Eight files from UCD are copied to `admin/unidata` and further processed. We will discuss here only the treatment of `UnicodeData.txt`.

The directory contains already various tools and a Makefile. Additionaly the file `https://www.unicode.org/copyright.html` is also copied to the directory.

The Makefile contains target `unidata.txt` which creates the file in the following format:

```
(#x0000 "<control>" "Cc" "0" "BN" "" "" "" "" "N" "NULL" "" "" "" "")
...
(#x0020 "SPACE" "Zs" "0" "WS" "" "" "" "" "N" "" "" "" "" "")
...
(#x20000 "<CJK Ideograph Extension B, First>" "Lo" "0" "L" "" "" "" "" "N" "" "" "" "" "")
(#x2A6DD "<CJK Ideograph Extension B, Last>" "Lo" "0" "L" "" "" "" "" "N" "" "" "" "" "")
```

For every property at least one separate file is created in the directory `lisp/international` The copyright clause from `https://www.unicode.org/copyright.html` is added to every file.

The list of the file names is contained in `unidata-gen.el` in the definition of the constant `unidata-file-alist`. The list is extracted in Makefile, prefixed by the appropriate path and assigned to `unidir` which is used as a target. The files are created by calling emacs, which loads `unidata-gen.el` and runs the `unidata-gen-file` with the appropriate parameters. In the source all the parameters are defined as optional, but the first one seems obligatory. They are e.g.

1. `../../lisp/international/uni-name.el`, i.e. the output file,

2. `.`, i.e. data directory containing the UCD files,
3. `unicode.txt`, i.e. the input file (actually this parameter is omitted and this is the default value).

Generating some files require also other input files besides `unidata.txt`.

The generated files contain calls to `define-char-code-property` defined in `mule-cmds.el` which store the properties in a compact way in appropriate char-tables (from the manual: *A char-table is a one-dimensional array of elements of any type, indexed by character codes. Char-tables have certain extra features to make them more useful for many jobs that involve assigning information to character codes—for example, a char-table can have a parent to inherit from, a default value, and a small number of extra slots to use for special purposes. A char-table can also specify a single value for a whole character set.*)

The files are loaded when needed by `charprop.el` which in turn is preloaded by `loadup.el`.

For more efficient name searching a hash table `ucs-names` is created by `ucs-names` defined in `mule-cmds.el` (*ucs* stands for *Universal Character Set*, which in principle is the name of the ISO/IEC 10646 standard, but in practice is often used to refer also to Unicode).

4. Private Use Areas

There are several conventions to use PUA, but their discussion is outside the scope of this note.

4.1. Medieval Unicode Font Initiative

The official site of MUFI is `https://mufi.info`. Unfortunately it does provide only names and codepoints.

The files partially mimicking the Unicode Character Database has been prepared by Rebecca G. Bettencourt and made available in particular in the repository:

`https://github.com/kreativekorp/charset`.

The file `unicodedata.txt` from the repository has been converted to `mufidata` analogous to `unidata` mentioned above, but all the names has been give `[MUFI 4.0]` postfix to distinguish them from the official characters (this convention is already used for some time with `https://bitbucket.org/jsbien/unihistext`). The conversion was done by a script containing the `sed` invocation extracted from `Makefile.`

The function unidata-gen-file and the file unidata-gen.el has been replaced respectively by the function mufidata-gen-file and the file mufidata-gen.el; the changes concerned only the output file names; the copyright notice is not yet updated. A script mufi-gen-file was prepared with the appropriate calls to the function.

Loading mufi-name.el was sufficent for describe-char to find the name of a MUFI character. However for insert-character to work a quick-and-dirty modification of ucs-names was necessary: the list of ranges should include the MUFI range and the hash table is to be extended and not created from the scratch; the modified function is placed in the mufi-mule-cmds.el file.

To facilitate testing the mufitest.txt file was prepared. For viewing the file I recommend the JuniusX font (https://github.com/psb1558/Junicode-New/).

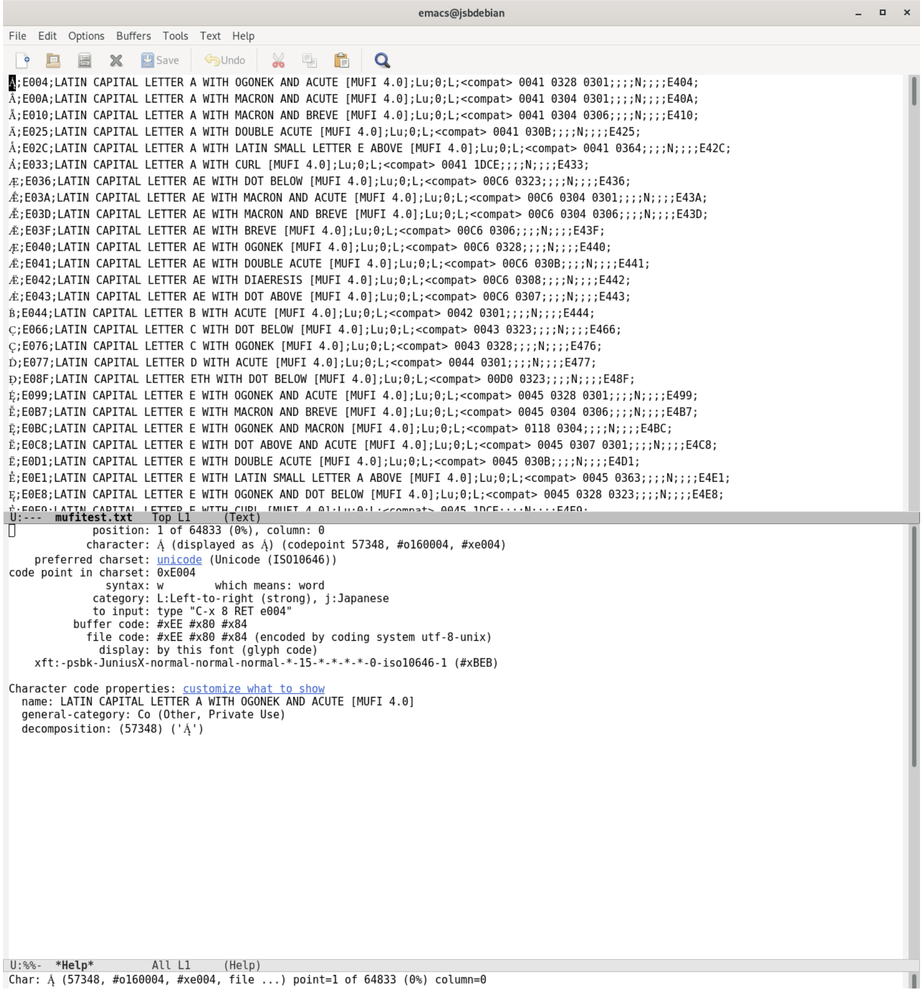
All the files mentioned in this section are available at https://github.com/jsbien/unicode4polish/ in the directory Emacs_MUFI.

4.2. Concluding remarks

There is still some work to be done, in particular other properties of the MUFI characters should be also accounted for in Emacs, for example forward-word should recognize a MUFI letter as a letter.

Another direction of work is to generalize the approach to various non-MUFI PUA character sets, e.g. those already described at https://github.com/kreativekorp/charset and http://www.kreativekorp.com/charset/PUADATA/.

The present note is available at https://github.com/jsbien/unicode4polish/Emacs-MUFI/doc but this can change in the future.

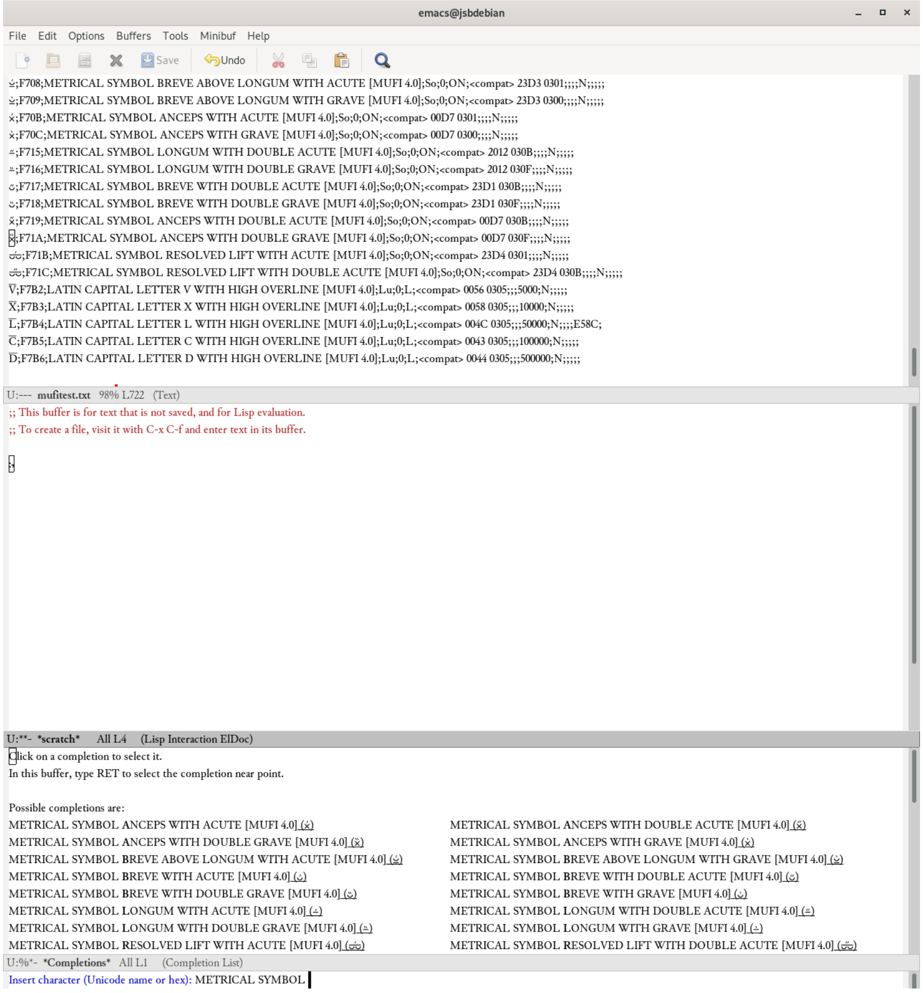


```
Ą́;E004;LATIN CAPITAL LETTER A WITH OGONEK AND ACUTE [MUFI 4.0];Lu;0;L;<compat> 0041 0328 0301;;;;N;;;;E404;
Ā́;E00A;LATIN CAPITAL LETTER A WITH MACRON AND ACUTE [MUFI 4.0];Lu;0;L;<compat> 0041 0304 0301;;;;N;;;;E40A;
Ā̆;E010;LATIN CAPITAL LETTER A WITH MACRON AND BREVE [MUFI 4.0];Lu;0;L;<compat> 0041 0304 0306;;;;N;;;;E410;
A̋;E025;LATIN CAPITAL LETTER A WITH DOUBLE ACUTE [MUFI 4.0];Lu;0;L;<compat> 0041 030B;;;;N;;;;E425;
Aͤ;E02C;LATIN CAPITAL LETTER A WITH LATIN SMALL LETTER E ABOVE [MUFI 4.0];Lu;0;L;<compat> 0041 0364;;;;N;;;;E42C;
A᷎;E033;LATIN CAPITAL LETTER A WITH CURL [MUFI 4.0];Lu;0;L;<compat> 0041 1DCE;;;;N;;;;E433;
Æ̣;E036;LATIN CAPITAL LETTER AE WITH DOT BELOW [MUFI 4.0];Lu;0;L;<compat> 00C6 0323;;;;N;;;;E436;
Ǣ́;E03A;LATIN CAPITAL LETTER AE WITH MACRON AND ACUTE [MUFI 4.0];Lu;0;L;<compat> 00C6 0304 0301;;;;N;;;;E43A;
Ǣ̆;E03D;LATIN CAPITAL LETTER AE WITH MACRON AND BREVE [MUFI 4.0];Lu;0;L;<compat> 00C6 0304 0306;;;;N;;;;E43D;
Æ̆;E03F;LATIN CAPITAL LETTER AE WITH BREVE [MUFI 4.0];Lu;0;L;<compat> 00C6 0306;;;;N;;;;E43F;
Æ̨;E040;LATIN CAPITAL LETTER AE WITH OGONEK [MUFI 4.0];Lu;0;L;<compat> 00C6 0328;;;;N;;;;E440;
Æ̋;E041;LATIN CAPITAL LETTER AE WITH DOUBLE ACUTE [MUFI 4.0];Lu;0;L;<compat> 00C6 030B;;;;N;;;;E441;
Æ̈;E042;LATIN CAPITAL LETTER AE WITH DIAERESIS [MUFI 4.0];Lu;0;L;<compat> 00C6 0308;;;;N;;;;E442;
Æ̇;E043;LATIN CAPITAL LETTER AE WITH DOT ABOVE [MUFI 4.0];Lu;0;L;<compat> 00C6 0307;;;;N;;;;E443;
B́;E044;LATIN CAPITAL LETTER B WITH ACUTE [MUFI 4.0];Lu;0;L;<compat> 0042 0301;;;;N;;;;E444;
C̣;E066;LATIN CAPITAL LETTER C WITH DOT BELOW [MUFI 4.0];Lu;0;L;<compat> 0043 0323;;;;N;;;;E466;
C̨;E076;LATIN CAPITAL LETTER C WITH OGONEK [MUFI 4.0];Lu;0;L;<compat> 0043 0328;;;;N;;;;E476;
D́;E077;LATIN CAPITAL LETTER D WITH ACUTE [MUFI 4.0];Lu;0;L;<compat> 0044 0301;;;;N;;;;E477;
Ð̣;E08F;LATIN CAPITAL LETTER ETH WITH DOT BELOW [MUFI 4.0];Lu;0;L;<compat> 00D0 0323;;;;N;;;;E48F;
Ę́;E099;LATIN CAPITAL LETTER E WITH OGONEK AND ACUTE [MUFI 4.0];Lu;0;L;<compat> 0045 0328 0301;;;;N;;;;E499;
Ē̆;E0B7;LATIN CAPITAL LETTER E WITH MACRON AND BREVE [MUFI 4.0];Lu;0;L;<compat> 0045 0304 0306;;;;N;;;;E4B7;
Ę̄;E0BC;LATIN CAPITAL LETTER E WITH OGONEK AND MACRON [MUFI 4.0];Lu;0;L;<compat> 0118 0304;;;;N;;;;E4BC;
Ė́;E0C8;LATIN CAPITAL LETTER E WITH DOT ABOVE AND ACUTE [MUFI 4.0];Lu;0;L;<compat> 0045 0307 0301;;;;N;;;;E4C8;
E̋;E0D1;LATIN CAPITAL LETTER E WITH DOUBLE ACUTE [MUFI 4.0];Lu;0;L;<compat> 0045 030B;;;;N;;;;E4D1;
Eͣ;E0E1;LATIN CAPITAL LETTER E WITH LATIN SMALL LETTER A ABOVE [MUFI 4.0];Lu;0;L;<compat> 0045 0363;;;;N;;;;E4E1;
Ę̣;E0E8;LATIN CAPITAL LETTER E WITH OGONEK AND DOT BELOW [MUFI 4.0];Lu;0;L;<compat> 0045 0328 0323;;;;N;;;;E4E8;
E᷎;E0E9;LATIN CAPITAL LETTER E WITH CURL [MUFI 4.0];Lu;0;L;<compat> 0045 1DCE;;;;N;;;;E4E9;
U:---  mufitest.txt   Top L1     (Text)
             position: 1 of 64833 (0%), column: 0
            character: Ą́ (displayed as Ą́) (codepoint 57348, #o160004, #xe004)
    preferred charset: unicode (Unicode (ISO10646))
code point in charset: 0xE004
               syntax: w        which means: word
             category: L:Left-to-right (strong), j:Japanese
             to input: type "C-x 8 RET e004"
          buffer code: #xEE #x80 #x84
            file code: #xEE #x80 #x84 (encoded by coding system utf-8-unix)
              display: by this font (glyph code)
    xft:-psbk-JuniusX-normal-normal-normal-*-15-*-*-*-*-0-iso10646-1 (#xBEB)

Character code properties: customize what to show
  name: LATIN CAPITAL LETTER A WITH OGONEK AND ACUTE [MUFI 4.0]
  general-category: Co (Other, Private Use)
  decomposition: (57348) ('Ą́')

U:%%-  *Help*         All L1     (Help)
Char: Ą́ (57348, #o160004, #xe004, file ...) point=1 of 64833 (0%) column=0
```

Figure 1. `C-U C-x =`



```
⏓́;F708;METRICAL SYMBOL BREVE ABOVE LONGUM WITH ACUTE [MUFI 4.0];So;0;ON;<compat> 23D3 0301;;;;N;;;;;
⏓̀;F709;METRICAL SYMBOL BREVE ABOVE LONGUM WITH GRAVE [MUFI 4.0];So;0;ON;<compat> 23D3 0300;;;;N;;;;;
×́;F70B;METRICAL SYMBOL ANCEPS WITH ACUTE [MUFI 4.0];So;0;ON;<compat> 00D7 0301;;;;N;;;;;
×̀;F70C;METRICAL SYMBOL ANCEPS WITH GRAVE [MUFI 4.0];So;0;ON;<compat> 00D7 0300;;;;N;;;;;
‒̋;F715;METRICAL SYMBOL LONGUM WITH DOUBLE ACUTE [MUFI 4.0];So;0;ON;<compat> 2012 030B;;;;N;;;;;
‒̏;F716;METRICAL SYMBOL LONGUM WITH DOUBLE GRAVE [MUFI 4.0];So;0;ON;<compat> 2012 030F;;;;N;;;;;
⏑̋;F717;METRICAL SYMBOL BREVE WITH DOUBLE ACUTE [MUFI 4.0];So;0;ON;<compat> 23D1 030B;;;;N;;;;;
⏑̏;F718;METRICAL SYMBOL BREVE WITH DOUBLE GRAVE [MUFI 4.0];So;0;ON;<compat> 23D1 030F;;;;N;;;;;
×̋;F719;METRICAL SYMBOL ANCEPS WITH DOUBLE ACUTE [MUFI 4.0];So;0;ON;<compat> 00D7 030B;;;;N;;;;;
×̏;F71A;METRICAL SYMBOL ANCEPS WITH DOUBLE GRAVE [MUFI 4.0];So;0;ON;<compat> 00D7 030F;;;;N;;;;;
⏔́;F71B;METRICAL SYMBOL RESOLVED LIFT WITH ACUTE [MUFI 4.0];So;0;ON;<compat> 23D4 0301;;;;N;;;;;
⏔̋;F71C;METRICAL SYMBOL RESOLVED LIFT WITH DOUBLE ACUTE [MUFI 4.0];So;0;ON;<compat> 23D4 030B;;;;N;;;;;
V̅;F7B2;LATIN CAPITAL LETTER V WITH HIGH OVERLINE [MUFI 4.0];Lu;0;L;<compat> 0056 0305;;;5000;N;;;;;
X̅;F7B3;LATIN CAPITAL LETTER X WITH HIGH OVERLINE [MUFI 4.0];Lu;0;L;<compat> 0058 0305;;;10000;N;;;;;
L̅;F7B4;LATIN CAPITAL LETTER L WITH HIGH OVERLINE [MUFI 4.0];Lu;0;L;<compat> 004C 0305;;;50000;N;;;;E58C;
C̅;F7B5;LATIN CAPITAL LETTER C WITH HIGH OVERLINE [MUFI 4.0];Lu;0;L;<compat> 0043 0305;;;100000;N;;;;;
D̅;F7B6;LATIN CAPITAL LETTER D WITH HIGH OVERLINE [MUFI 4.0];Lu;0;L;<compat> 0044 0305;;;500000;N;;;;;

U:---  mufitest.txt   98% L722   (Text)
;; This buffer is for text that is not saved, and for Lisp evaluation.
;; To create a file, visit it with C-x C-f and enter text in its buffer.

⏑

U:**-  *scratch*    All L4    (Lisp Interaction ElDoc)
Click on a completion to select it.
In this buffer, type RET to select the completion near point.

Possible completions are:
METRICAL SYMBOL ANCEPS WITH ACUTE [MUFI 4.0] (×́)                    METRICAL SYMBOL ANCEPS WITH DOUBLE ACUTE [MUFI 4.0] (×̋)
METRICAL SYMBOL ANCEPS WITH DOUBLE GRAVE [MUFI 4.0] (×̏)             METRICAL SYMBOL ANCEPS WITH GRAVE [MUFI 4.0] (×̀)
METRICAL SYMBOL BREVE ABOVE LONGUM WITH ACUTE [MUFI 4.0] (⏓́)        METRICAL SYMBOL BREVE ABOVE LONGUM WITH GRAVE [MUFI 4.0] (⏓̀)
METRICAL SYMBOL BREVE WITH ACUTE [MUFI 4.0] (⏑́)                     METRICAL SYMBOL BREVE WITH DOUBLE ACUTE [MUFI 4.0] (⏑̋)
METRICAL SYMBOL BREVE WITH DOUBLE GRAVE [MUFI 4.0] (⏑̏)              METRICAL SYMBOL BREVE WITH GRAVE [MUFI 4.0] (⏑̀)
METRICAL SYMBOL LONGUM WITH ACUTE [MUFI 4.0] (‒́)                    METRICAL SYMBOL LONGUM WITH DOUBLE ACUTE [MUFI 4.0] (‒̋)
METRICAL SYMBOL LONGUM WITH DOUBLE GRAVE [MUFI 4.0] (‒̏)             METRICAL SYMBOL LONGUM WITH GRAVE [MUFI 4.0] (‒̀)
METRICAL SYMBOL RESOLVED LIFT WITH ACUTE [MUFI 4.0] (⏔́)             METRICAL SYMBOL RESOLVED LIFT WITH DOUBLE ACUTE [MUFI 4.0] (⏔̋)
U:%*-  *Completions*   All L1    (Completion List)
Insert character (Unicode name or hex): METRICAL SYMBOL
```

Figure 2. `C-x 8 RET`