Providing some UTF-8 support via `inputenc`

Frank Mittelbach Chris Rowley*

? 2015/06/27 v1.1n UTF-8 support for inputenc? printed July 7, 2015

Contents

*Borrowing heavily from code by David Carlisle and tables by Sebastian Rahtz; some table and code cleanup by Javier Bezos

1 Introduction

[The whole section is rather unfinished ... just like the code, sorry!]

1.1 Background and general stuff

For many reasons what this package provides is a long way from any type of 'Unicode compliance'.

In stark contrast to 8-bit character sets, with 16 or more bits it can easily be very inefficient to support the full range.[1] Moreover, useful support of character input by a typesetting system overwhelmingly means finding an acceptable visual representation of a sequence of characters and this, for LaTeX, means having available a suitably encoded 8-bit font.

Unfortunately it is not possible to predict exactly what valid UTF-8 octet sequences will appear in a particular file so it is best to make all the unsupported but valid sequences produce a reasonably clear and noticeable error message.

There are two directions from which to approach the question of what to load. One is to specify the ranges of Unicode characters that will result in some sensible typesetting; this requires the provider to ensure that suitable fonts are loaded and that these input characters generate the correct typesetting via the encodings of those fonts. The other is to inspect the font encodings to be used and use these to define which input Unicode characters should be supported.

For Western European languages, at least, going in either direction leads to many straightforward decisions and a few that are more subjective. In both cases some of the specifications are TeX specific whilst most are independent of the particular typesetting software in use.

As we have argued elsewhere, LaTeX needs to refer to characters via 'seven-bit-text' names and, so far, these have been chosen by reference to historical sources such as Plain TeX or Adobe encoding descriptions. It is unclear whether this ad hoc naming structure should simply be extended or whether it would be useful to supplement it with standardised internal Unicode character names such as one or more of the following:[2]

```
\ltxutwochar <4 hex digits>

\ltxuchar {<hex digits>}
  B H U R R R

\ltxueightchartwo   <2 utf8 octets as 8-bit char tokens>
\ltxueightcharthree <3 utf8 octets ...>
\ltxueightcharfour  <4 utf8 octets ...>
```

1.2 More specific stuff

In addition to setting up the mechanism for reading UTF-8 characters and specifying the LaTeX-level support available, this package contains support for some

[1] In fact, LaTeX's current 8-bit support does not go so far as to make all 8-bit characters into valid input.

[2] Burkhard und Holger Mittelbach spielen mit mir! Sie haben etwas hier geschrieben.

default historically expected TEX-related characters and some example 'Unicode definition files' for standard font encodings.

1.3 Notes

This package does not support Unicode combining characters as TEX is not really equipped to make this possible.

No attempt is made to be useful beyond Latin, and maybe Cyrillic, for European languages (as of now).

1.4 Basic operation of the code

The `inputenc` package makes the upper 8-bit characters active and assigns to all of them an error message. It then waits for the input encoding files to change this set-up. Similarly, whenever `\inputencoding` is encountered in a document, first the upper 8-bit characters are set back to produce an error and then the definitions for the new input encoding are loaded, changing some of the previous settings.

The 8-bit input encodings currently supported by `inputenc` all use declarations such as `\DeclareInputText` and the like to map an 8-bit number to some LATEX internal form, e.g. to `\"a`.

The situation when supporting UTF-8 as the input encoding is different, however. Here we only have to set up the actions of those 8-bit numbers that can be the first octet in a UTF-8 representation of a Unicode character. But we cannot simply set this to some internal LATEX form since the Unicode character consists of more than one octet; instead we have to define this starting octet to parse the right number of further octets that together form the UTF-8 representation of some Unicode character.

Therefore when switching to `utf8` within the `inputenc` framework the characters with numbers (hex) from `"C2` to `"DF` are defined to parse for a second octet following, the characters from `"E0` to `"EF` are defined to parse for two more octets and finally the characters from `"F0` to `"F3` are defined to parse for three additional octets. These additional octets are always in the range `"80` to `"B9`.

Thus, when such a character is encountered in the document (so long as expansion is not prohibited) a defined number of additional octets (8-bit characters) are read and from them a unique control sequence name is immediately constructed.

This control sequence is either defined (good) or undefined (likely); in the latter case the user gets an error message saying that this UTF-8 sequence (or, better, Unicode character) is not supported.

If the control sequence is set up to do something useful then it will expand to a LATEX internal form: e.g. for the utf8 sequence of two octets `"C3 "A4` we get `\"a` as the internal form which then, depending on the font encoding, eventually resolves to the single glyph 'latin-a-umlaut' or to the composite glyph 'latin-a with an umlaut accent'.

These mappings from (UTF-8 encoded) Unicode characters to LATEX internal forms are made indirectly. The code below provides a declaration `\DeclareUnicodeCharacter` which maps Unicode numbers (as hexadecimal) to LATEX internal forms.

This mapping needs to be set up only once so it is done at `\begin{document}` by looking at the list of font encodings that are loaded by the document and providing mappings related to those font encodings whenever these are available.

Thus at most only those Unicode characters that can be represented by the glyphs available in these encodings will be defined.

Technically this is done by loading one file per encoding, if available, that is supposed to provide the necessary mapping information.

2 Coding

2.1 Housekeeping

The usual introductory bits and pieces:

```
⟨utf8⟩\ProvidesFile{utf8.def}
⟨test⟩\ProvidesFile{utf8-test.tex}
⟨+lcy⟩ \ProvidesFile{lcyenc.dfu}
⟨+ly1⟩ \ProvidesFile{ly1enc.dfu}
⟨+oms⟩ \ProvidesFile{omsenc.dfu}
⟨+ot1⟩ \ProvidesFile{ot1enc.dfu}
⟨+ot2⟩ \ProvidesFile{ot2enc.dfu}
⟨+t1⟩  \ProvidesFile{t1enc.dfu}
⟨+t2a⟩ \ProvidesFile{t2aenc.dfu}
⟨+t2b⟩ \ProvidesFile{t2benc.dfu}
⟨+t2c⟩ \ProvidesFile{t2cenc.dfu}
⟨+ts1⟩ \ProvidesFile{ts1enc.dfu}
⟨+x2⟩  \ProvidesFile{x2enc.dfu}
⟨+all⟩ \ProvidesFile{utf8enc.dfu}
   [2015/06/27 v1.1n UTF-8 support for inputenc]
⟨*utf8⟩
\makeatletter
```

We restore the `\catcode` of space (which is set to ignore in `inputenc`) while reading `.def` files. Otherwise we would need to explicitly use `\space` all over the place in error and log messages.

```
\catcode`\ \saved@space@catcode
```

2.2 Parsing UTF-8 input

`\UTFviii@two@octets`
`\UTFviii@three@octets`
`\UTFviii@four@octets`

A UTF-8 char (that is not actually a 7-bit char, i.e. a single octet) is parsed as follows: each starting octet is an active TEX character token; each of these is defined below to be a macro with one to three arguments nominally (depending on the starting octet). It calls one of `\UTFviii@two@octets`, `\UTFviii@three@octets`, or `\UTFviii@four@octets` which then actually picks up the argument(s).

From the arguments a control sequence with a name of the form `u8:#1#2...` is constructed where the `#i` ($i > 1$) are the arguments and `#1` is the starting octet (as a TEX character token). Since some or even all of these characters are active (when inputenc is loaded) we need to use `\string` when building the csname.

The csname thus constructed can of course be undefined but to avoid producing an unhelpful low-level undefined command error we pass it to `\UTFviii@defined` which is responsible for producing a more sensible error message (not yet done!!). If, however, it is defined we simply execute the thing (which should then expand to an encoding specific internal LATEX form).

```
\def\UTFviii@two@octets#1#2{\expandafter
    \UTFviii@defined\csname u8:#1\string#2\endcsname}
```

```
\def\UTFviii@three@octets#1#2#3{\expandafter
    \UTFviii@defined\csname u8:#1\string#2\string#3\endcsname}
\def\UTFviii@four@octets#1#2#3#4{\expandafter
    \UTFviii@defined\csname u8:#1\string#2\string#3\string#4\endcsname}
```

`\UTFviii@defined` This tests whether its argument is different from `\relax`: it either calls for a sensible error message (not done), or it gets the `\fi` out of the way (in case the command has arguments) and executes it.

```
\def\UTFviii@defined#1{%
  \ifx#1\relax
```

The endline character has a special definition within the inputenc package (it is gobbling spaces). For this reason we can't produce multiline strings without some precaution.

```
      \PackageError{inputenc}{Unicode\space char\space \string#1\space
                              not\space set\space up\space
                              for\space use\space with\space LaTeX}\@eha
  \else\expandafter
    #1%
  \fi
}
```

`\UTFviii@loop` This wonderful bit of code from Dr Carlisle defines the starting octets to call `\UTFviii@two@octets` etc as appropriate. The starting octet itself is passed directly as the first argument, the others are picked up later en route.

The `\UTFviii@loop` loops through the numbers starting at `\count@` and ending at `\@tempcnta` − 1, each time executing the code in `\UTFviii@tmp`.

All this is done in a group so that temporary catcode changes etc. vanish after everything is set up.

It may be a good idea to add code to deal with 'illegal utf8 octets': at present these will be handled by whatever code was in use for 8-bit input before this code is executed.

```
\begingroup
\catcode`\~13
\catcode`\"12
\def\UTFviii@loop{%
  \uccode`\~\count@
  \uppercase\expandafter{\UTFviii@tmp}%
  \advance\count@\@ne
  \ifnum\count@<\@tempcnta
  \expandafter\UTFviii@loop
  \fi}
```

Setting up 2-byte UTF-8:

```
    \count@"C2
    \@tempcnta"E0
    \def\UTFviii@tmp{\xdef~{\noexpand\UTFviii@two@octets\string~}}
\UTFviii@loop
```

Setting up 3-byte UTF-8:

```
    \count@"E0
    \@tempcnta"F0
    \def\UTFviii@tmp{\xdef~{\noexpand\UTFviii@three@octets\string~}}
\UTFviii@loop
```

Setting up 4-byte UTF-8:

```
    \count@"F0
    \@tempcnta"F4
    \def\UTFviii@tmp{\xdef~{\noexpand\UTFviii@four@octets\string~}}
\UTFviii@loop
\endgroup
```

For this case we must disable the warning generated by `inputenc` if it doesn't see any new `\DeclareInputText` commands.

```
\@inpenc@test
```

If this file (`utf8.def`) is not being read while setting up `inputenc`, i.e. in the preamble, but when `\inputencoding` is called somewhere within the document, we do not need to input the specific Unicode mappings again. We therefore stop reading the file at this point.

```
\ifx\@begindocumenthook\@undefined
  \makeatother
```

The `\fi` must be on the same line as `\endinput` or else it will never be seen!

```
  \endinput \fi
```

2.3 Mapping Unicode codes to LaTeX internal forms

`\DeclareUnicodeCharacter`

The `\DeclareUnicodeCharacter` declaration defines a mapping from a Unicode character code point to a LaTeX internal form. The first argument is the Unicode number as hexadecimal digits and the second is the actual LaTeX internal form.

We start by making sure that some characters have the right `\catcode` when they are used in the definitions below.

```
\begingroup
\catcode`\"=12
\catcode`\<=12
\catcode`\.=12
\catcode`\,=12
\catcode`\;=12
\catcode`\!=12
\catcode`\~=13
```

```
\gdef\DeclareUnicodeCharacter#1#2{%
  \count@"#1\relax
  \wlog{ \space\space defining Unicode char U+#1 (decimal \the\count@)}%
  \begingroup
```

Next we do the parsing of the number stored in `\count@` and assign the result to `\UTFviii@tmp`. Actually all this could be done in-line, the macro `\parse@XML@charref` is only there to extend this code to parsing Unicode numbers in other contexts one day (perhaps).

```
    \parse@XML@charref
```

Here is an example of what is happening, for the pair `"C2 "A3` (which is the utf8 represenation for the character £). After `\parse@XML@charref` we have, stored in `\UTFviii@tmp`, a single command with two character tokens as arguments:

> [t_{C2} and t_{A3} are the characters corresponding to these two octets]
> `\UTFviii@two@octets` $t_{\mathrm{C2}}t_{\mathrm{A3}}$

what we actually need to produce is a definition of the form

\def\u8:$t_{C2}t_{A3}$ {*LaTeX internal form*}.

So here we temporarily redefine the prefix commands \UTFviii@two@octets, etc. to generate the csname that we wish to define¿ the \strings are added in case these tokens are still active.

```
\def\UTFviii@two@octets##1##2{\csname u8:##1\string##2\endcsname}%
\def\UTFviii@three@octets##1##2##3{\csname u8:##1%
                                   \string##2\string##3\endcsname}%
\def\UTFviii@four@octets##1##2##3##4{\csname u8:##1%
                            \string##2\string##3\string##4\endcsname}%
```

Now we simply:-) need to use the right number of \expandafters to finally construct the definition: expanding \UTFviii@tmp once to get its contents, a second time to replace the prefix command by its \csname expansion, and a third time to turn the expansion into a csname after which the \gdef finally gets applied. We add an irrelevant \IeC and braces around the definition, in order to avoid any space after the command being gobbled up when the text is written out to an auxiliary file (see inputenc for further details

```
  \expandafter\expandafter\expandafter
  \expandafter\expandafter\expandafter
  \expandafter
   \gdef\UTFviii@tmp{\IeC{#2}}%
 \endgroup
}
```

\parse@XML@charref This macro parses a Unicode number (decimal) and returns its UTF-8 representation as a sequence of non-active TEX character tokens. In the original code it had two arguments delimited by ; here, however, we supply the Unicode number implicitly.

```
\gdef\parse@XML@charref{%
```

We need to keep a few things local, mainly the \uccode's that are set up below. However, the group originally used here is actually unnecessary since we call this macro only within another group; but it will be important to restore the group if this macro gets used for other purposes.

```
%  \begingroup
```

The original code from David supported the convention that a Unicode slot number could be given either as a decimal or as a hexadecimal (by starting with x). We do not do this so this code is also removed. This could be reactivated if one wants to support document commands that accept Unicode numbers (but then the first case needs to be changed from an error message back to something more useful again).

```
%  \uppercase{\count@\if x\noexpand#1"\else#1\fi#2}\relax
```

As \count@ already contains the right value we make \parse@XML@charref work without arguments.

```
\ifnum\count@<"A0\relax
  \PackageError{inputenc}{Cannot\space define\space Unicode\space
                          char\space value\space <\space 00A0}\@eha
```

Do not ask us to provide an explanation for the code below, it is borrowed straight from `xmltex` by David and we trust him totally (and we are too lazy to reread the Unicode book to see if this is the correct algorithm).[3]

```
  \else\ifnum\count@<"800\relax
     \parse@UTFviii@a,%
     \parse@UTFviii@b C\UTFviii@two@octets.,%
  \else\ifnum\count@<"10000\relax
     \parse@UTFviii@a;%
     \parse@UTFviii@a,%
     \parse@UTFviii@b E\UTFviii@three@octets.{,;}%
   \else
     \parse@UTFviii@a;%
     \parse@UTFviii@a,%
     \parse@UTFviii@a!%
     \parse@UTFviii@b F\UTFviii@four@octets.{!,;}%
    \fi
    \fi
  \fi
%  \endgroup
}
```

`\parse@UTFviii@a` ... so somebody else can document this part :-) ... David?:-))))!

```
\gdef\parse@UTFviii@a#1{%
     \@tempcnta\count@
     \divide\count@ 64
     \@tempcntb\count@
     \multiply\count@ 64
     \advance\@tempcnta-\count@
     \advance\@tempcnta 128
     \uccode`#1\@tempcnta
     \count@\@tempcntb}
```

`\parse@UTFviii@b` ... same here

```
\gdef\parse@UTFviii@b#1#2#3#4{%
     \advance\count@ "#10\relax
     \uccode`#3\count@
     \uppercase{\gdef\UTFviii@tmp{#2#3#4}}}
\endgroup

\@onlypreamble\DeclareUnicodeCharacter
```

These are preamble only as long as we don't support Unicode charrefs in documents.

```
\@onlypreamble\parse@XML@charref
\@onlypreamble\parse@UTFviii@a
\@onlypreamble\parse@UTFviii@b
```

2.4 Loading Unicode mappings at begin document

The original plan was to set up the UTF-8 support at `\begin{document}`; but then any text characters used in the preamble (as people do even though advised

[3] We were hoping to also find in his work the TEX code for going the other way: from UTF-8 octets to Unicode slot number, but no luck!

against it) would fail in one way or the other. So the implementation was changed and the Unicode definition files for already defined encodings are loaded here.

We loop through all defined font encodings (stored in `\cdp@list`) and for each load a file *name*`enc.dfu` if it exist. That file is then supposed to contain `\DeclareUnicodeCharacter` declarations.

```
\begingroup
  \def\cdp@elt#1#2#3#4{%
    \wlog{Now handling font encoding #1 ...}%
    \lowercase{%
        \InputIfFileExists{#1enc.dfu}}%
           {\wlog{... processing UTF-8 mapping file for font %
                     encoding #1}%
```

The previous line is written to the log with the newline char being ignored (thus not producing a space). Therefore either everything has to be on a single input line or some special care must be taken. From this point on we ignore spaces again, i.e., while we are reading the `.dfu` file. The `\endgroup` below will restore it again.

```
            \catcode`\ 9\relax}%
          {\wlog{... no UTF-8 mapping file for font encoding #1}}%
  }
  \cdp@list
\endgroup
```

However, we don't know if there are font encodings still to be loaded (either with `fontenc` or directly with `\input` by some some package). Font encoding files are loaded only if the corresponding encoding has not been loaded yet, and they always begin with `\DeclareFontEncoding`. We now redefine the internal kernel version of the latter to load the Unicode file if available.

```
\def\DeclareFontEncoding@#1#2#3{%
  \expandafter
  \ifx\csname T@#1\endcsname\relax
    \def\cdp@elt{\noexpand\cdp@elt}%
    \xdef\cdp@list{\cdp@list\cdp@elt{#1}%
                   {\default@family}{\default@series}%
                   {\default@shape}}%
    \expandafter\let\csname#1-cmd\endcsname\@changed@cmd
    \begingroup
      \wlog{Now handling font encoding #1 ...}%
      \lowercase{%
        \InputIfFileExists{#1enc.dfu}}%
           {\wlog{... processing UTF-8 mapping file for font %
                     encoding #1}}%
           {\wlog{... no UTF-8 mapping file for font encoding #1}}%
    \endgroup
  \else
     \@font@info{Redeclaring font encoding #1}%
  \fi
  \global\@namedef{T@#1}{#2}%
  \global\@namedef{M@#1}{\default@M#3}%
  \xdef\LastDeclaredEncoding{#1}%
  }
```

⟨/utf8⟩

3 Mapping characters — based on font (glyph) encodings

This section is a first attempt to provide Unicode definitions for characters whose standard glyphs are currently provided by the standard LaTeX font-encodings `T1`, `OT1`, etc. They are by no means completed and need checking.

For example, one should check the already existing input encodings for glyphs that may in fact be available and required, e.g. `latin4` has a number of glyphs with the `\=` accent. Since the `T1` encoding does not provide such glyphs, these characters are not listed below (yet).

The list below was generated by looking at the current LaTeX font encoding files, e.g., `t1enc.def` and using the work by Sebastian Rahtz (in `ucharacters.sty`) with a few modifications. In combinations such as `\^\i` the preferred form is that and not `\^i`.

This list has been built from several sources, obviously including the Unicode Standard itself. These sources include Passive TeX by Sebastian Rahtz, the `unicode` package by Dominique P. G. Unruh (mainly for Latin encodings) and `text4ht` by Eitan Gurari (for Cyrillic ones).

Note that it strictly follows the Mittelbach principles for input character encodings: thus it offers no support for using utf8 representations of math symbols such as × or ÷ (in math mode).

3.1 About the table itself

In addition to generating individual files, the table below is, at present, a one-one (we think) partial relationship between the (ill-defined) set of LICRs and the Unicode slots "0080 to "FFFF. At present these entries are used only to define a collection of partial mappings from Unicode slots to LICRs; each of these mappings becomes full if we add an exception value ('not defined') to the set of LICRs.

It is probably not essential for the relationship in the full table to be one-one; this raises questions such as: the exact role of LICRs; the formal relationships on the set of LICRs; the (non-mathematical) relationship between LICRs and Unicode (which has its own somewhat fuzzy equivalences); and ultimately what a character is and what a character representation and/or name is.

Viewed this a way, the result has, perhaps puzzling, just two (we think) gaps in the second 128 'Unicode slots' (00A0 and 00AD): neither of these is really a character, of course.

It is unclear the extent to which entries in this table should resemble the closely related ones in the 8-bit `inputenc` files. The Unicode standard claims that the first 256 slots 'are' ASCII and Latin-1.

Of course, TeX itself typically does not treat even many perfectly 'normal text' 7-bit slots as text characters, so it is unclear whether LaTeX should even attempt to deal in any consistent way with those Unicode slots that are not definitive text characters.

3.2 The mapping table

Note that the first argument must be a 4-hex-digit number greater than `00BF`.

There are few notes about inconsistencies etc at the end of the table.

```
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{00A1}{\textexclamdown}
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{00A2}{\textcent}
⟨all, ts1, t1, ot1, ly1⟩\DeclareUnicodeCharacter{00A3}{\textsterling}
⟨all, x2, ts1, t2c, t2b, t2a, ly1, lcy⟩\DeclareUnicodeCharacter{00A4}{\textcurrency}
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{00A5}{\textyen}
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{00A6}{\textbrokenbar}
⟨all, x2, ts1, t2c, t2b, t2a, oms, ly1⟩\DeclareUnicodeCharacter{00A7}{\textsection}
⟨all, ts1⟩\DeclareUnicodeCharacter{00A8}{\textasciidieresis}
⟨all, ts1, utf8⟩\DeclareUnicodeCharacter{00A9}{\textcopyright}
⟨all, ts1, ly1, utf8⟩\DeclareUnicodeCharacter{00AA}{\textordfeminine}
⟨*all, x2, t2c, t2b, t2a, t1, ot2, ly1, lcy⟩
\DeclareUnicodeCharacter{00AB}{\guillemotleft}
⟨/all, x2, t2c, t2b, t2a, t1, ot2, ly1, lcy⟩
⟨all, ts1⟩\DeclareUnicodeCharacter{00AC}{\textlnot}
⟨all, ts1, ly1, utf8⟩\DeclareUnicodeCharacter{00AE}{\textregistered}
⟨all, ts1⟩\DeclareUnicodeCharacter{00AF}{\textasciimacron}
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{00B0}{\textdegree}
⟨all, ts1⟩\DeclareUnicodeCharacter{00B1}{\textpm}
⟨all, ts1⟩\DeclareUnicodeCharacter{00B2}{\texttwosuperior}
⟨all, ts1⟩\DeclareUnicodeCharacter{00B3}{\textthreesuperior}
⟨all, ts1⟩\DeclareUnicodeCharacter{00B4}{\textasciiacute}
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{00B5}{\textmu} % micro sign
⟨all, ts1, oms, ly1⟩\DeclareUnicodeCharacter{00B6}{\textparagraph}
⟨all, oms, ts1, ly1⟩\DeclareUnicodeCharacter{00B7}{\textperiodcentered}
⟨all, ot1⟩\DeclareUnicodeCharacter{00B8}{\c\ }
⟨all, ts1⟩\DeclareUnicodeCharacter{00B9}{\textonesuperior}
⟨all, ts1, ly1, utf8⟩\DeclareUnicodeCharacter{00BA}{\textordmasculine}
⟨*all, x2, t2c, t2b, t2a, t1, ot2, ly1, lcy⟩
\DeclareUnicodeCharacter{00BB}{\guillemotright}
⟨/all, x2, t2c, t2b, t2a, t1, ot2, ly1, lcy⟩
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{00BC}{\textonequarter}
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{00BD}{\textonehalf}
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{00BE}{\textthreequarters}
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{00BF}{\textquestiondown}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00C0}{\@tabacckludge`A}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00C1}{\@tabacckludge'A}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00C2}{\^A}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00C3}{\~A}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00C4}{\"A}
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{00C5}{\r A}
⟨all, t1, ot1, ly1, lcy⟩\DeclareUnicodeCharacter{00C6}{\AE}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00C7}{\c C}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00C8}{\@tabacckludge`E}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00C9}{\@tabacckludge'E}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00CA}{\^E}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00CB}{\"E}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00CC}{\@tabacckludge`I}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00CD}{\@tabacckludge'I}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00CE}{\^I}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00CF}{\"I}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00D0}{\DH}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00D1}{\~N}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00D2}{\@tabacckludge`O}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00D3}{\@tabacckludge'O}
```

```
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00D4}{\^O}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00D5}{\~O}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00D6}{\"O}
⟨all, ts1⟩\DeclareUnicodeCharacter{00D7}{\texttimes}
⟨all, t1, ot1, ly1, lcy⟩\DeclareUnicodeCharacter{00D8}{\O}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00D9}{\@tabacckludge‘U}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00DA}{\@tabacckludge’U}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00DB}{\^U}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00DC}{\"U}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00DD}{\@tabacckludge’Y}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00DE}{\TH}
⟨all, t1, ot1, ly1, lcy⟩\DeclareUnicodeCharacter{00DF}{\ss}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E0}{\@tabacckludge‘a}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E1}{\@tabacckludge’a}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E2}{\^a}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E3}{\~a}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E4}{\"a}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E5}{\r a}
⟨all, t1, ot1, ly1, lcy⟩\DeclareUnicodeCharacter{00E6}{\ae}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E7}{\c c}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E8}{\@tabacckludge‘e}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00E9}{\@tabacckludge’e}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00EA}{\^e}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00EB}{\"e}
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{00EC}{\@tabacckludge‘\i}
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{00ED}{\@tabacckludge’\i}
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{00EE}{\^\i}
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{00EF}{\"\i}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00F0}{\dh}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00F1}{\~n}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00F2}{\@tabacckludge‘o}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00F3}{\@tabacckludge’o}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00F4}{\^o}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00F5}{\~o}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00F6}{\"o}
⟨all, ts1⟩\DeclareUnicodeCharacter{00F7}{\textdiv}
⟨all, t1, ot1, ly1, lcy⟩\DeclareUnicodeCharacter{00F8}{\o}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00F9}{\@tabacckludge‘u}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00FA}{\@tabacckludge’u}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00FB}{\^u}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00FC}{\"u}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00FD}{\@tabacckludge’y}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00FE}{\th}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{00FF}{\"y}
⟨all, t1⟩\DeclareUnicodeCharacter{0102}{\u A}
⟨all, t1⟩\DeclareUnicodeCharacter{0103}{\u a}
⟨all, t1⟩\DeclareUnicodeCharacter{0104}{\k A}
⟨all, t1⟩\DeclareUnicodeCharacter{0105}{\k a}
⟨all, t1⟩\DeclareUnicodeCharacter{0106}{\@tabacckludge’C}
⟨all, t1⟩\DeclareUnicodeCharacter{0107}{\@tabacckludge’c}
⟨all, t1⟩\DeclareUnicodeCharacter{010C}{\v C}
⟨all, t1⟩\DeclareUnicodeCharacter{010D}{\v c}
⟨all, t1⟩\DeclareUnicodeCharacter{010E}{\v D}
⟨all, t1⟩\DeclareUnicodeCharacter{010F}{\v d}
```

```
⟨all, t1⟩\DeclareUnicodeCharacter{0110}{\DJ}
⟨all, t1⟩\DeclareUnicodeCharacter{0111}{\dj}
⟨all, t1⟩\DeclareUnicodeCharacter{0118}{\k E}
⟨all, t1⟩\DeclareUnicodeCharacter{0119}{\k e}
⟨all, t1⟩\DeclareUnicodeCharacter{011A}{\v E}
⟨all, t1⟩\DeclareUnicodeCharacter{011B}{\v e}
⟨all, t1⟩\DeclareUnicodeCharacter{011E}{\u G}
⟨all, t1⟩\DeclareUnicodeCharacter{011F}{\u g}
⟨all, t1⟩\DeclareUnicodeCharacter{0130}{\.I}
⟨all, t2c, t2b, t2a, t1, ot2, ot1, ly1, lcy⟩\DeclareUnicodeCharacter{0131}{\i}
⟨all, t1⟩\DeclareUnicodeCharacter{0132}{\IJ}
⟨all, t1⟩\DeclareUnicodeCharacter{0133}{\ij}
⟨all, t1⟩\DeclareUnicodeCharacter{0139}{\@tabacckludge'L}
⟨all, t1⟩\DeclareUnicodeCharacter{013A}{\@tabacckludge'l}
⟨all, t1⟩\DeclareUnicodeCharacter{013D}{\v L}
⟨all, t1⟩\DeclareUnicodeCharacter{013E}{\v l}
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{0141}{\L}
⟨all, t1, ot1, ly1⟩\DeclareUnicodeCharacter{0142}{\l}
⟨all, t1⟩\DeclareUnicodeCharacter{0143}{\@tabacckludge'N}
⟨all, t1⟩\DeclareUnicodeCharacter{0144}{\@tabacckludge'n}
⟨all, t1⟩\DeclareUnicodeCharacter{0147}{\v N}
⟨all, t1⟩\DeclareUnicodeCharacter{0148}{\v n}
⟨all, t1⟩\DeclareUnicodeCharacter{014A}{\NG}
⟨all, t1⟩\DeclareUnicodeCharacter{014B}{\ng}
⟨all, t1⟩\DeclareUnicodeCharacter{0150}{\H O}
⟨all, t1⟩\DeclareUnicodeCharacter{0151}{\H o}
⟨all, t1, ot1, ly1, lcy⟩\DeclareUnicodeCharacter{0152}{\OE}
⟨all, t1, ot1, ly1, lcy⟩\DeclareUnicodeCharacter{0153}{\oe}
⟨all, t1⟩\DeclareUnicodeCharacter{0154}{\@tabacckludge'R}
⟨all, t1⟩\DeclareUnicodeCharacter{0155}{\@tabacckludge'r}
⟨all, t1⟩\DeclareUnicodeCharacter{0158}{\v R}
⟨all, t1⟩\DeclareUnicodeCharacter{0159}{\v r}
⟨all, t1⟩\DeclareUnicodeCharacter{015A}{\@tabacckludge'S}
⟨all, t1⟩\DeclareUnicodeCharacter{015B}{\@tabacckludge's}
⟨all, t1⟩\DeclareUnicodeCharacter{015E}{\c S}
⟨all, t1⟩\DeclareUnicodeCharacter{015F}{\c s}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{0160}{\v S}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{0161}{\v s}
⟨all, t1⟩\DeclareUnicodeCharacter{0162}{\c T}
⟨all, t1⟩\DeclareUnicodeCharacter{0163}{\c t}
⟨all, t1⟩\DeclareUnicodeCharacter{0164}{\v T}
⟨all, t1⟩\DeclareUnicodeCharacter{0165}{\v t}
⟨all, t1⟩\DeclareUnicodeCharacter{016E}{\r U}
⟨all, t1⟩\DeclareUnicodeCharacter{016F}{\r u}
⟨all, t1⟩\DeclareUnicodeCharacter{0170}{\H U}
⟨all, t1⟩\DeclareUnicodeCharacter{0171}{\H u}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{0178}{\"Y}
⟨all, t1⟩\DeclareUnicodeCharacter{0179}{\@tabacckludge'Z}
⟨all, t1⟩\DeclareUnicodeCharacter{017A}{\@tabacckludge'z}
⟨all, t1⟩\DeclareUnicodeCharacter{017B}{\.Z}
⟨all, t1⟩\DeclareUnicodeCharacter{017C}{\.z}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{017D}{\v Z}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{017E}{\v z}
⟨all, ts1, ly1⟩\DeclareUnicodeCharacter{0192}{\textflorin}
```

```
⟨all, ly1, utf8⟩\DeclareUnicodeCharacter{02C6}{\textasciicircum}
⟨all, ts1⟩\DeclareUnicodeCharacter{02C7}{\textasciicaron}
⟨all, ly1, utf8⟩\DeclareUnicodeCharacter{02DC}{\textasciitilde}
⟨all, ts1⟩\DeclareUnicodeCharacter{02D8}{\textasciibreve}
⟨all, ts1⟩\DeclareUnicodeCharacter{02DD}{\textacutedbl}
```

The Cyrillic code points have been recently checked (2007) and extended and corrected by Matthias Noe (a9931078@unet.univie.ac.at) — thanks.

```
⟨*all, x2, t2c, t2b, t2a, ot2, lcy⟩
\DeclareUnicodeCharacter{0400}{\@tabacckludge‘\CYRE}
⟨/all, x2, t2c, t2b, t2a, ot2, lcy⟩
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{0401}{\CYRYO}
⟨all, x2, t2a, ot2⟩\DeclareUnicodeCharacter{0402}{\CYRDJE}
⟨*all, x2, t2c, t2b, t2a, ot2, lcy⟩
\DeclareUnicodeCharacter{0403}{\@tabacckludge’\CYRG}
⟨/all, x2, t2c, t2b, t2a, ot2, lcy⟩
⟨all, x2, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{0404}{\CYRIE}
⟨all, x2, t2c, t2b, t2a, ot2⟩\DeclareUnicodeCharacter{0405}{\CYRDZE}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{0406}{\CYRII}
⟨all, x2, t2a, lcy⟩\DeclareUnicodeCharacter{0407}{\CYRYI}
⟨all, x2, t2c, t2b, t2a, ot2⟩\DeclareUnicodeCharacter{0408}{\CYRJE}
⟨all, x2, t2b, t2a, ot2⟩\DeclareUnicodeCharacter{0409}{\CYRLJE}
⟨all, x2, t2b, t2a, ot2⟩\DeclareUnicodeCharacter{040A}{\CYRNJE}
⟨all, x2, t2a, ot2⟩\DeclareUnicodeCharacter{040B}{\CYRTSHE}
⟨*all, x2, t2c, t2b, t2a, ot2, lcy⟩
\DeclareUnicodeCharacter{040C}{\@tabacckludge’\CYRK}
\DeclareUnicodeCharacter{040D}{\@tabacckludge‘\CYRI}
⟨/all, x2, t2c, t2b, t2a, ot2, lcy⟩
⟨all, x2, t2b, t2a, lcy⟩\DeclareUnicodeCharacter{040E}{\CYRUSHRT}
⟨all, x2, t2c, t2a, ot2⟩\DeclareUnicodeCharacter{040F}{\CYRDZHE}
⟨*all, x2, t2c, t2b, t2a, ot2, lcy⟩
\DeclareUnicodeCharacter{0410}{\CYRA}
\DeclareUnicodeCharacter{0411}{\CYRB}
\DeclareUnicodeCharacter{0412}{\CYRV}
\DeclareUnicodeCharacter{0413}{\CYRG}
\DeclareUnicodeCharacter{0414}{\CYRD}
\DeclareUnicodeCharacter{0415}{\CYRE}
\DeclareUnicodeCharacter{0416}{\CYRZH}
\DeclareUnicodeCharacter{0417}{\CYRZ}
\DeclareUnicodeCharacter{0418}{\CYRI}
\DeclareUnicodeCharacter{0419}{\CYRISHRT}
\DeclareUnicodeCharacter{041A}{\CYRK}
\DeclareUnicodeCharacter{041B}{\CYRL}
\DeclareUnicodeCharacter{041C}{\CYRM}
\DeclareUnicodeCharacter{041D}{\CYRN}
\DeclareUnicodeCharacter{041E}{\CYRO}
\DeclareUnicodeCharacter{041F}{\CYRP}
\DeclareUnicodeCharacter{0420}{\CYRR}
\DeclareUnicodeCharacter{0421}{\CYRS}
\DeclareUnicodeCharacter{0422}{\CYRT}
\DeclareUnicodeCharacter{0423}{\CYRU}
\DeclareUnicodeCharacter{0424}{\CYRF}
\DeclareUnicodeCharacter{0425}{\CYRH}
\DeclareUnicodeCharacter{0426}{\CYRC}
```

```
\DeclareUnicodeCharacter{0427}{\CYRCH}
\DeclareUnicodeCharacter{0428}{\CYRSH}
\DeclareUnicodeCharacter{0429}{\CYRSHCH}
\DeclareUnicodeCharacter{042A}{\CYRHRDSN}
\DeclareUnicodeCharacter{042B}{\CYRERY}
\DeclareUnicodeCharacter{042C}{\CYRSFTSN}
\DeclareUnicodeCharacter{042D}{\CYREREV}
\DeclareUnicodeCharacter{042E}{\CYRYU}
\DeclareUnicodeCharacter{042F}{\CYRYA}
\DeclareUnicodeCharacter{0430}{\cyra}
\DeclareUnicodeCharacter{0431}{\cyrb}
\DeclareUnicodeCharacter{0432}{\cyrv}
\DeclareUnicodeCharacter{0433}{\cyrg}
\DeclareUnicodeCharacter{0434}{\cyrd}
\DeclareUnicodeCharacter{0435}{\cyre}
\DeclareUnicodeCharacter{0436}{\cyrzh}
\DeclareUnicodeCharacter{0437}{\cyrz}
\DeclareUnicodeCharacter{0438}{\cyri}
\DeclareUnicodeCharacter{0439}{\cyrishrt}
\DeclareUnicodeCharacter{043A}{\cyrk}
\DeclareUnicodeCharacter{043B}{\cyrl}
\DeclareUnicodeCharacter{043C}{\cyrm}
\DeclareUnicodeCharacter{043D}{\cyrn}
\DeclareUnicodeCharacter{043E}{\cyro}
\DeclareUnicodeCharacter{043F}{\cyrp}
\DeclareUnicodeCharacter{0440}{\cyrr}
\DeclareUnicodeCharacter{0441}{\cyrs}
\DeclareUnicodeCharacter{0442}{\cyrt}
\DeclareUnicodeCharacter{0443}{\cyru}
\DeclareUnicodeCharacter{0444}{\cyrf}
\DeclareUnicodeCharacter{0445}{\cyrh}
\DeclareUnicodeCharacter{0446}{\cyrc}
\DeclareUnicodeCharacter{0447}{\cyrch}
\DeclareUnicodeCharacter{0448}{\cyrsh}
\DeclareUnicodeCharacter{0449}{\cyrshch}
\DeclareUnicodeCharacter{044A}{\cyrhrdsn}
\DeclareUnicodeCharacter{044B}{\cyrery}
\DeclareUnicodeCharacter{044C}{\cyrsftsn}
\DeclareUnicodeCharacter{044D}{\cyrerev}
\DeclareUnicodeCharacter{044E}{\cyryu}
\DeclareUnicodeCharacter{044F}{\cyrya}
\DeclareUnicodeCharacter{0450}{\@tabacckludge`\cyre}
\DeclareUnicodeCharacter{0451}{\cyryo}
⟨/all, x2, t2c, t2b, t2a, ot2, lcy⟩
⟨all, x2, t2a, ot2⟩\DeclareUnicodeCharacter{0452}{\cyrdje}
⟨*all, x2, t2c, t2b, t2a, ot2, lcy⟩
\DeclareUnicodeCharacter{0453}{\@tabacckludge'\cyrg}
⟨/all, x2, t2c, t2b, t2a, ot2, lcy⟩
⟨all, x2, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{0454}{\cyrie}
⟨all, x2, t2c, t2b, t2a, ot2⟩\DeclareUnicodeCharacter{0455}{\cyrdze}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{0456}{\cyrii}
⟨all, x2, t2a, lcy⟩\DeclareUnicodeCharacter{0457}{\cyryi}
⟨all, x2, t2c, t2b, t2a, ot2⟩\DeclareUnicodeCharacter{0458}{\cyrje}
⟨all, x2, t2b, t2a, ot2⟩\DeclareUnicodeCharacter{0459}{\cyrlje}
```

```
⟨all, x2, t2b, t2a, ot2⟩\DeclareUnicodeCharacter{045A}{\cyrnje}
⟨all, x2, t2a, ot2⟩\DeclareUnicodeCharacter{045B}{\cyrtshe}
⟨*all, x2, t2c, t2b, t2a, ot2, lcy⟩
\DeclareUnicodeCharacter{045C}{\@tabacckludge'\cyrk}
\DeclareUnicodeCharacter{045D}{\@tabacckludge`\cyri}
⟨/all, x2, t2c, t2b, t2a, ot2, lcy⟩
⟨all, x2, t2b, t2a, lcy⟩\DeclareUnicodeCharacter{045E}{\cyrushrt}
⟨all, x2, t2c, t2a, ot2⟩\DeclareUnicodeCharacter{045F}{\cyrdzhe}
⟨all, x2, ot2⟩\DeclareUnicodeCharacter{0462}{\CYRYAT}
⟨all, x2, ot2⟩\DeclareUnicodeCharacter{0463}{\cyryat}
⟨all, x2⟩\DeclareUnicodeCharacter{046A}{\CYRBYUS}
⟨all, x2⟩\DeclareUnicodeCharacter{046B}{\cyrbyus}
```

The next two declarations are questionable, the encoding definition should probably contain `\CYROTLD` and `\cyrotld`. Or alternatively, if the characters in the X2 encodings are really meant to represent the historical characters in Ux0472 and Ux0473 (they look like them) then they would need to change instead.

However, their looks are probably a font designers decision and the next two mappings are wrong or rather the names in OT2 should change for consistency.

On the other hand the names `\CYROTLD` are somewhat questionabled as the Unicode standard only describes "Cyrillic barred O" while `TLD` refers to a tilde (which is more less what the "Cyrillic FITA looks according to the Unicode book).

```
⟨all, ot2⟩\DeclareUnicodeCharacter{0472}{\CYRFITA}
⟨all, ot2⟩\DeclareUnicodeCharacter{0473}{\cyrfita}

⟨all, x2, ot2⟩\DeclareUnicodeCharacter{0474}{\CYRIZH}
⟨all, x2, ot2⟩\DeclareUnicodeCharacter{0475}{\cyrizh}
```

While the double grave accent seems to exist in X2, T2A, T2B and T2C encoding, the letter izhitsa exists only in X2 and OT2. Therefore, izhitsa with double grave seems to be possible only using X2.

```
⟨all, x2⟩\DeclareUnicodeCharacter{0476}{\C\CYRIZH}
⟨all, x2⟩\DeclareUnicodeCharacter{0477}{\C\cyrizh}

⟨all, t2c⟩\DeclareUnicodeCharacter{048C}{\CYRSEMISFTSN}
⟨all, t2c⟩\DeclareUnicodeCharacter{048D}{\cyrsemisftsn}
⟨all, t2c⟩\DeclareUnicodeCharacter{048E}{\CYRRTICK}
⟨all, t2c⟩\DeclareUnicodeCharacter{048F}{\cyrrtick}
⟨all, x2, t2a, lcy⟩\DeclareUnicodeCharacter{0490}{\CYRGUP}
⟨all, x2, t2a, lcy⟩\DeclareUnicodeCharacter{0491}{\cyrgup}
⟨all, x2, t2b, t2a⟩\DeclareUnicodeCharacter{0492}{\CYRGHCRS}
⟨all, x2, t2b, t2a⟩\DeclareUnicodeCharacter{0493}{\cyrghcrs}
⟨all, x2, t2c, t2b⟩\DeclareUnicodeCharacter{0494}{\CYRGHK}
⟨all, x2, t2c, t2b⟩\DeclareUnicodeCharacter{0495}{\cyrghk}
⟨all, x2, t2b, t2a⟩\DeclareUnicodeCharacter{0496}{\CYRZHDSC}
⟨all, x2, t2b, t2a⟩\DeclareUnicodeCharacter{0497}{\cyrzhdsc}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{0498}{\CYRZDSC}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{0499}{\cyrzdsc}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{049A}{\CYRKDSC}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{049B}{\cyrkdsc}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{049C}{\CYRKVCRS}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{049D}{\cyrkvcrs}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{049E}{\CYRKHCRS}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{049F}{\cyrkhcrs}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04A0}{\CYRKBEAK}
```

```
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04A1}{\cyrkbeak}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04A2}{\CYRNDSC}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04A3}{\cyrndsc}
⟨all, x2, t2b, t2a⟩\DeclareUnicodeCharacter{04A4}{\CYRNG}
⟨all, x2, t2b, t2a⟩\DeclareUnicodeCharacter{04A5}{\cyrng}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04A6}{\CYRPHK}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04A7}{\cyrphk}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04A8}{\CYRABHHA}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04A9}{\cyrabhha}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04AA}{\CYRSDSC}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04AB}{\cyrsdsc}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04AC}{\CYRTDSC}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04AD}{\cyrtdsc}
⟨all, x2, t2b, t2a⟩\DeclareUnicodeCharacter{04AE}{\CYRY}
⟨all, x2, t2b, t2a⟩\DeclareUnicodeCharacter{04AF}{\cyry}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04B0}{\CYRYHCRS}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04B1}{\cyryhcrs}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04B2}{\CYRHDSC}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04B3}{\cyrhdsc}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04B4}{\CYRTETSE}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04B5}{\cyrtetse}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04B6}{\CYRCHRDSC}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04B7}{\cyrchrdsc}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04B8}{\CYRCHVCRS}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04B9}{\cyrchvcrs}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04BA}{\CYRSHHA}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04BB}{\cyrshha}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04BC}{\CYRABHCH}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04BD}{\cyrabhch}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04BE}{\CYRABHCHDSC}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04BF}{\cyrabhchdsc}
```

The character `\CYRpalochka` is not defined by OT2 and LCY. However it is looking identical to `\CYRII` and the Unicode standard explicitly refers to that (and to Latin I). So perhaps those encodings could get an alias? On the other hand, why are there two distinct slots in the T2 encodings even though they are so pressed for space? Perhaps they don't always look alike.

```
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04C0}{\CYRpalochka}

⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04C1}{\U\CYRZH}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04C2}{\U\cyrzh}
⟨all, x2, t2b⟩\DeclareUnicodeCharacter{04C3}{\CYRKHK}
⟨all, x2, t2b⟩\DeclareUnicodeCharacter{04C4}{\cyrkhk}
```

According to the Unicode standard Ux04C5 should be an L with "tail" not with descender (which also exists as Ux04A2) but it looks as if the char names do not make this distinction). Should they?

```
⟨all, x2, t2c, t2b⟩\DeclareUnicodeCharacter{04C5}{\CYRLDSC}
⟨all, x2, t2c, t2b⟩\DeclareUnicodeCharacter{04C6}{\cyrldsc}

⟨all, x2, t2c, t2b⟩\DeclareUnicodeCharacter{04C7}{\CYRNHK}
⟨all, x2, t2c, t2b⟩\DeclareUnicodeCharacter{04C8}{\cyrnhk}
⟨all, x2, t2b⟩\DeclareUnicodeCharacter{04CB}{\CYRCHLDSC}
⟨all, x2, t2b⟩\DeclareUnicodeCharacter{04CC}{\cyrchldsc}
```

According to the Unicode standard Ux04CD should be an M with "tail" not with descender. However this time there is no M with descender in the Unicode standard.

```
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04CD}{\CYRMDSC}
⟨all, x2, t2c⟩\DeclareUnicodeCharacter{04CE}{\cyrmdsc}

⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04D0}{\U\CYRA}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04D1}{\U\cyra}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04D2}{\"\CYRA}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04D3}{\"\cyra}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04D4}{\CYRAE}
⟨all, x2, t2a⟩\DeclareUnicodeCharacter{04D5}{\cyrae}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04D6}{\U\CYRE}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04D7}{\U\cyre}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04D8}{\CYRSCHWA}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04D9}{\cyrschwa}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04DA}{\"\CYRSCHWA}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04DB}{\"\cyrschwa}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04DC}{\"\CYRZH}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04DD}{\"\cyrzh}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04DE}{\"\CYRZ}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04DF}{\"\cyrz}
⟨all, x2, t2c, t2b⟩\DeclareUnicodeCharacter{04E0}{\CYRABHDZE}
⟨all, x2, t2c, t2b⟩\DeclareUnicodeCharacter{04E1}{\cyrabhdze}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04E2}{\=\CYRI}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04E3}{\=\cyri}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04E4}{\"\CYRI}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04E5}{\"\cyri}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04E6}{\"\CYRO}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04E7}{\"\cyro}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04E8}{\CYROTLD}
⟨all, x2, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{04E9}{\cyrotld}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04EC}{\"\CYREREV}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04ED}{\"\cyrerev}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04EE}{\=\CYRU}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04EF}{\=\cyru}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04F0}{\"\CYRU}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04F1}{\"\cyru}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04F2}{\H\CYRU}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04F3}{\H\cyru}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04F4}{\"\CYRCH}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04F5}{\"\cyrch}
⟨all, x2, t2b⟩\DeclareUnicodeCharacter{04F6}{\CYRGDSC}
⟨all, x2, t2b⟩\DeclareUnicodeCharacter{04F7}{\cyrgdsc}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04F8}{\"\CYRERY}
⟨all, x2, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{04F9}{\"\cyrery}
⟨all, t2b⟩\DeclareUnicodeCharacter{04FA}{\CYRGDSCHCRS}
⟨all, t2b⟩\DeclareUnicodeCharacter{04FB}{\cyrgdschcrs}
⟨all, x2, t2b⟩\DeclareUnicodeCharacter{04FC}{\CYRHHK}
⟨all, x2, t2b⟩\DeclareUnicodeCharacter{04FD}{\cyrhhk}
⟨all, t2b⟩\DeclareUnicodeCharacter{04FE}{\CYRHHCRS}
⟨all, t2b⟩\DeclareUnicodeCharacter{04FF}{\cyrhhcrs}
⟨all, ts1⟩\DeclareUnicodeCharacter{0E3F}{\textbaht}
⟨all, x2, t2c, t2b, t2a, t1, utf8⟩\DeclareUnicodeCharacter{200C}{\textcompwordmark}
```

```
⟨*all, x2, t2c, t2b, t2a, t1, ot2, ot1, ly1, lcy⟩
\DeclareUnicodeCharacter{2013}{\textendash}
\DeclareUnicodeCharacter{2014}{\textemdash}
⟨/all, x2, t2c, t2b, t2a, t1, ot2, ot1, ly1, lcy⟩
⟨all, ts1⟩\DeclareUnicodeCharacter{2016}{\textbardbl}
⟨*all, x2, t2c, t2b, t2a, t1, ot2, ot1, lcy⟩
\DeclareUnicodeCharacter{2018}{\textquoteleft}
\DeclareUnicodeCharacter{2019}{\textquoteright}
⟨/all, x2, t2c, t2b, t2a, t1, ot2, ot1, lcy⟩
⟨all, t1⟩\DeclareUnicodeCharacter{201A}{\quotesinglbase}
⟨*all, x2, t2c, t2b, t2a, t1, ot2, ot1, ly1, lcy⟩
\DeclareUnicodeCharacter{201C}{\textquotedblleft}
\DeclareUnicodeCharacter{201D}{\textquotedblright}
⟨/all, x2, t2c, t2b, t2a, t1, ot2, ot1, ly1, lcy⟩
⟨all, x2, t2c, t2b, t2a, t1, lcy⟩\DeclareUnicodeCharacter{201E}{\quotedblbase}
⟨all, ts1, oms, ly1⟩\DeclareUnicodeCharacter{2020}{\textdagger}
⟨all, ts1, oms, ly1⟩\DeclareUnicodeCharacter{2021}{\textdaggerdbl}
⟨all, ts1, oms, ly1⟩\DeclareUnicodeCharacter{2022}{\textbullet}
⟨all, ly1, utf8⟩\DeclareUnicodeCharacter{2026}{\textellipsis}
⟨*all, x2, ts1, t2c, t2b, t2a, t1, ly1⟩
\DeclareUnicodeCharacter{2030}{\textperthousand}
⟨/all, x2, ts1, t2c, t2b, t2a, t1, ly1⟩
⟨*all, x2, ts1, t2c, t2b, t2a, t1⟩
\DeclareUnicodeCharacter{2031}{\textpertenthousand}
⟨/all, x2, ts1, t2c, t2b, t2a, t1⟩
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{2039}{\guilsinglleft}
⟨all, t1, ly1⟩\DeclareUnicodeCharacter{203A}{\guilsinglright}
⟨all, ts1⟩\DeclareUnicodeCharacter{203B}{\textreferencemark}
⟨all, ts1⟩\DeclareUnicodeCharacter{203D}{\textinterrobang}
⟨all, ts1⟩\DeclareUnicodeCharacter{2044}{\textfractionsolidus}
⟨all, ts1⟩\DeclareUnicodeCharacter{204E}{\textasteriskcentered}
⟨all, ts1⟩\DeclareUnicodeCharacter{2052}{\textdiscount}
⟨all, ts1⟩\DeclareUnicodeCharacter{20A1}{\textcolonmonetary}
⟨all, ts1⟩\DeclareUnicodeCharacter{20A4}{\textlira}
⟨all, ts1⟩\DeclareUnicodeCharacter{20A6}{\textnaira}
⟨all, ts1⟩\DeclareUnicodeCharacter{20A9}{\textwon}
⟨all, ts1⟩\DeclareUnicodeCharacter{20AB}{\textdong}
⟨all, ts1⟩\DeclareUnicodeCharacter{20AC}{\texteuro}
⟨all, ts1⟩\DeclareUnicodeCharacter{20B1}{\textpeso}
⟨all, ts1⟩\DeclareUnicodeCharacter{2103}{\textcelsius}
⟨all, x2, ts1, t2c, t2b, t2a, ot2, lcy⟩\DeclareUnicodeCharacter{2116}{\textnumero}
⟨all, ts1⟩\DeclareUnicodeCharacter{2117}{\textcircledP}
⟨all, ts1⟩\DeclareUnicodeCharacter{211E}{\textrecipe}
⟨all, ts1⟩\DeclareUnicodeCharacter{2120}{\textservicemark}
⟨all, ts1, ly1, utf8⟩\DeclareUnicodeCharacter{2122}{\texttrademark}
⟨all, ts1⟩\DeclareUnicodeCharacter{2126}{\textohm}
⟨all, ts1⟩\DeclareUnicodeCharacter{2127}{\textmho}
⟨all, ts1⟩\DeclareUnicodeCharacter{212E}{\textestimated}
⟨all, ts1⟩\DeclareUnicodeCharacter{2190}{\textleftarrow}
⟨all, ts1⟩\DeclareUnicodeCharacter{2191}{\textuparrow}
⟨all, ts1⟩\DeclareUnicodeCharacter{2192}{\textrightarrow}
⟨all, ts1⟩\DeclareUnicodeCharacter{2193}{\textdownarrow}
⟨all, x2, ts1, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{2329}{\textlangle}
⟨all, x2, ts1, t2c, t2b, t2a⟩\DeclareUnicodeCharacter{232A}{\textrangle}
```

```
⟨all, ts1⟩\DeclareUnicodeCharacter{2422}{\textblank}
⟨all, x2, t2c, t2b, t2a, t1, utf8⟩\DeclareUnicodeCharacter{2423}{\textvisiblespace}
⟨all, ts1⟩\DeclareUnicodeCharacter{25E6}{\textopenbullet}
⟨all, ts1⟩\DeclareUnicodeCharacter{25EF}{\textbigcircle}
⟨all, ts1⟩\DeclareUnicodeCharacter{266A}{\textmusicalnote}
```

3.3 Notes

The following inputs are inconsistent with the 8-bit inputenc files since they will always only produce the 'text character'. This is an area where inputenc is notoriously confused.

```
%<all,ts1,t1,ot1,ly1>\DeclareUnicodeCharacter{00A3}{\textsterling}
%<*all,x2,ts1,t2c,t2b,t2a,oms,ly1>
\DeclareUnicodeCharacter{00A7}{\textsection}
%</all,x2,ts1,t2c,t2b,t2a,oms,ly1>
%<all,ts1,utf8>\DeclareUnicodeCharacter{00A9}{\textcopyright}
%<all,ts1>\DeclareUnicodeCharacter{00B1}{\textpm}
%<all,ts1,oms,ly1>\DeclareUnicodeCharacter{00B6}{\textparagraph}
%<all,ts1,oms,ly1>\DeclareUnicodeCharacter{2020}{\textdagger}
%<all,ts1,oms,ly1>\DeclareUnicodeCharacter{2021}{\textdaggerdbl}
%<all,ly1,utf8>\DeclareUnicodeCharacter{2026}{\textellipsis}
```

The following definitions are in an encoding file but have no direct equivalent in Unicode, or they simply do not make sense in that context (or we have not yet found anything or ...:-). For example, the non-combining accent characters are certainly available somewhere but these are not equivalent to a TEX accent command.

```
\DeclareTextSymbol{\j}{OT1}{17}
\DeclareTextSymbol{\SS}{T1}{223}
\DeclareTextSymbol{\textcompwordmark}{T1}{23}

\DeclareTextAccent{\"}{OT1}{127}
\DeclareTextAccent{\'}{OT1}{19}
\DeclareTextAccent{\.}{OT1}{95}
\DeclareTextAccent{\=}{OT1}{22}
\DeclareTextAccent{\H}{OT1}{125}
\DeclareTextAccent{\^}{OT1}{94}
\DeclareTextAccent{\`}{OT1}{18}
\DeclareTextAccent{\r}{OT1}{23}
\DeclareTextAccent{\u}{OT1}{21}
\DeclareTextAccent{\v}{OT1}{20}
\DeclareTextAccent{\~}{OT1}{126}
\DeclareTextCommand{\b}{OT1}[1]
\DeclareTextCommand{\c}{OT1}[1]
\DeclareTextCommand{\d}{OT1}[1]
\DeclareTextCommand{\k}{T1}[1]
```

3.4 Mappings for OT1 glyphs

This is even more incomplete as again it covers only the single glyphs from `OT1` plus some that have been explicitly defined for this encoding. Everything that is

provided in T1, and that could be provided as composite glyphs via OT1, could and probably should be set up as well. Which leaves the many things that are not provided in T1 but can be provided in OT1 (and in T1) by composite glyphs.

Stuff not mapped (note that \j (*j*) is not equivalent to any Unicode character):

```
\DeclareTextSymbol{\j}{OT1}{17}
\DeclareTextAccent{\"}{OT1}{127}
\DeclareTextAccent{\'}{OT1}{19}
\DeclareTextAccent{\.}{OT1}{95}
\DeclareTextAccent{\=}{OT1}{22}
\DeclareTextAccent{\^}{OT1}{94}
\DeclareTextAccent{\`}{OT1}{18}
\DeclareTextAccent{\~}{OT1}{126}
\DeclareTextAccent{\H}{OT1}{125}
\DeclareTextAccent{\u}{OT1}{21}
\DeclareTextAccent{\v}{OT1}{20}
\DeclareTextAccent{\r}{OT1}{23}
\DeclareTextCommand{\b}{OT1}[1]
\DeclareTextCommand{\c}{OT1}[1]
\DeclareTextCommand{\d}{OT1}[1]
```

3.5 Mappings for OMS glyphs

Characters like \textbackslash are not mapped as they are (primarily) only in the lower 127 and the code here only sets up mappings for UTF-8 characters that are at least 2 octets long.

```
\DeclareTextSymbol{\textbackslash}{OMS}{110}        % "6E
\DeclareTextSymbol{\textbar}{OMS}{106}              % "6A
\DeclareTextSymbol{\textbraceleft}{OMS}{102}        % "66
\DeclareTextSymbol{\textbraceright}{OMS}{103}       % "67
```

But the following (and some others) might actually lurk in Unicode somewhere...

```
\DeclareTextSymbol{\textasteriskcentered}{OMS}{3}   % "03
\DeclareTextCommand{\textcircled}{OMS}
```

3.6 Mappings for TS1 glyphs

Exercise for somebody else.

3.7 Mappings for latex.ltx glyphs

There is also a collection of characters already set up in the kernel, one way or the other. Since these do not clearly relate to any particular font encoding they are mapped when the utf8 support is first set up.

Also there are a number of \providecommands in the various input encoding files which may or may not go into this part.

619 ⟨*utf8⟩
620 `% This space is intentionally empty ...`
621 ⟨/utf8⟩

4 A test document

Here is a very small test document which may or may not survive if the current document is transferred from one place to the other.

```
⟨*test⟩
\documentclass{article}

\usepackage[latin1,utf8]{inputenc}
\usepackage[T1]{fontenc}
\usepackage{trace}

\scrollmode  % to run past the error below

\begin{document}

 German umlauts in UTF-8: ^^c3^^a4^^c3^^b6^^c3^^bc  %%% äöü

\inputencoding{latin1}  % switch to latin1

 German umlauts in UTF-8 but read by latin1 (and will produce one
 error since \verb=\textcurrency= is not provided):
 ^^c3^^a4^^c3^^b6^^c3^^bc

\inputencoding{utf8}    % switch back to utf8

 German umlauts in UTF-8: ^^c3^^a4^^c3^^b6^^c3^^bc


Some codes that should produce errors as nothing is set up
for them: ^^c3F ^^e1^^a4^^b6

And some that are not legal utf8 sequences: ^^c3X ^^e1XY

\showoutput
\tracingstats=2
\stop
⟨/test⟩
```