

Font setup for Greek with XeTeX/LuaTeX

Günter Milde

2020/10/30

The file `tuenc-greek.def` provides support for Greek LICR macros and upcasing of text with XeTeX and LuaTeX. It is loaded automatically by `textalpha`, `alphabet`, and `babel-greek` when used with Unicode fonts (LuaTeX or XeTeX with `fontspec`).

Contents

1	Requirements	2
1.1	<code>fontspec</code> and suitable Unicode fonts	2
2	Usage	2
3	LICR input	2
3.1	Greek alphabet	2
3.2	Diacritics	2
3.2.1	perispomeni vs. tilde	3
3.2.2	combined diacritics	3
3.2.3	sub-iota	3
3.3	Additional Greek symbols	4
3.3.1	symbols for Greek numbers	4
3.3.2	symbol variants	4
3.3.3	Ancient Greek Numbers	6
3.3.4	generic text symbols	6
4	Latin transliteration	6
5	UPPERCASE and lowercase	6
5.1	hiatus	7
6	Character Tables	7
6.1	Greek and Coptic Unicode block	7
6.2	Greek Extended Unicode block	8
6.3	Character Tables: accent + literal base character	9
6.3.1	Greek and Coptic Unicode block	9
6.3.2	Greek Extended Unicode block	10

1 Requirements

1.1 *fontspec* and suitable Unicode fonts

LaTeX sets up the TU Unicode text font encoding if it detects the XeTeX or LuaTeX engines. The user must ensure that the selected font contains Greek glyphs (the default Latin Modern fonts have only capital Greek letters). **There are no errors for missing glyphs**, just warnings in the log file (but not in the console output) and empty spaces in the output document.

The *fontspec* package is the standard tool to select fonts in XeTeX/LuaTeX. *babel* provides a front-end to set up language-specific fonts with the `\babelfont` command.

2 Usage

`tuenc-greek.def` is loaded automatically by *textalpha*, *alphabeta*, or *Babel* (with the language option `greek`) if these packages are used with Unicode-aware TeX engines (XeTeX or LuaTeX). This provides support for Greek LICR input and fixes for `MakeUppercase`. The `\MakeUppercase` implementation introduced in the 2022/06 LaTeX release works uses Greek upcasing rules only for text parts with the *babel* language `greek` and ignores the changes to `uccode` values done in this package.

3 LICR input

The LaTeX internal character representation (LICR) is a verbose, fail-safe 7-bit ASCII encoding that can be used unaltered under both, 8-bit TeX and XeTeX/LuaTeX. Use cases are macro definitions and generated text.

See the source of this document, `test-tuenc-greek.tex` for the input used in the examples below.

3.1 Greek alphabet

Greek letters via LICR macros:

```
A B Γ Δ E Z H Θ I K Λ M N Ξ O Π P Σ T Υ Φ X Ψ Ω
α β γ δ ε ζ η θ ι κ λ μ ν ξ ο π ρ σ τ υ φ χ ψ ω
```

The small sigma is set with a different glyph if it ends a word:

```
σ textsigma
ς textfinalsigma or textvarsigma
```

The `\textautosigma` macro, which automatically chooses the glyph according to the position, relies on `\makelowercase` and only works correctly with LaTeX versions newer than 2022/06: `σι` vs. `ισ`.

3.2 Diacritics

Greek diacritics can be input by named macro or symbol macro:

```
άά ξξ àà ùù üü ãã õõ àà ùù áá ùù
```

XeTeX and LuaTeX (since 2022) normalise base letter and combining diacritics to the corresponding pre-composed character (if such character exists in the *Greek and Coptic* or *Greek Extended* Unicode Blocks).

ά ά ά à ì ì ã ã ã à à á á

3.2.1 perispomeni vs. tilde

The Greek *perispomeni* has the look of a tilde but the semantic of a circumflex accent. The Unicode standard provides distinct code points. Only <Greek letter> + COMBINING GREEK PERISPOMENI is considered equivalent to the pre-composed character <Greek letter> WITH PERISPOMENI.

The “named” `\accperispomeni` macro uses COMBINING GREEK PERISPOMENI, while the standard tilde-accent macro `\~` uses the COMBINING TILDE. Composite definitions for `\~` select the pre-composed character:

ã = ã, ñ = ñ, ï = ï, ù = ù, õ = õ

3.2.2 combined diacritics

Combined accents are defined using combining diacritical characters.

í í í x̄x̄x̄ ì ì x̄x̄x̄ ï ï x̄x̄x̄
 ı ı ı x̄x̄x̄ ı ı x̄x̄x̄ ı ı x̄x̄x̄
 ı ı ı x̄x̄x̄ ı ı x̄x̄x̄ ı ı x̄x̄x̄

Composite diacritics overlap when they are not normalised to a pre-composed character. However, this is not a major problem in normal use as pre-composed characters exist in Unicode for all letters that are used with diacritics in (ancient, polytonic or monotonic) Greek.

3.2.3 sub-iota

The sub-iota is input after the base letter.

- `\ypogegrammeni` sets a COMBINING GREEK YPOGEGRAMMENI: α κ.
 A Greek capital letter followed by COMBINING GREEK YPOGEGRAMMENI is normalised to the corresponding Greek capital letter WITH [... AND] PROSGEGRAMMENI, if a mapping exists in the Unicode standard (by XeTeX but not by LuaTeX)
- `\prosgegrammeni` sets a spacing GREEK PROSGEGRAMMENI: Α Κ..
 Spacing is better with the pre-composed characters for Greek capital letters ... WITH PROSGEGRAMMENI.
 Compare Αι (small letter iota) vs. Α. (spacing prosgegrammeni) vs. Α. (pre-composed).

Test letters with `ypogegrammeni` and `prosgegrammeni` (literal/LICR):

unchanged	make lowercase	make uppercase.
αα./αα.	αα./αα	Α.ΑΙ/Α.Α.
Α.Α./Α.Α.	αα./αα	Α.ΑΙ/Α.Α.
ΛΛ./ΛΛ.	λλ./λλ	ΛΛΙ/ΛΛ.

TeX math		Unicode	
symbol	var symbol	“letter”	“symbol”
π	ϖ	π	\Uppsi
ρ	ϱ	ρ	\Uppsi
θ	ϑ	θ	\Uppsi
ϵ	ε	ϵ	ϵ
ϕ	φ	Φ	Φ
β	<i>missing</i>	β	β
κ	<i>missing</i>	κ	κ
Θ	<i>missing</i>	Θ	Θ

Table 1: Greek symbol variants in TeX and Unicode

3.3 Additional Greek symbols

3.3.1 symbols for Greek numbers

\textkappa textkoppa
 \textKoppa textKoppa
 \textqoppa (archaic koppa)
 \textQoppa (archaic Koppa)
 \textsigma textstigma
 \textStigma (Sigma-Tau-Ligature in CB-fonts)¹
 \textsampi textsampi
 \textSampi textSampi
 \textdigamma textdigamma
 \textDigamma textDigamma
 \textdexiakeraia textdexiakeraia
 $\text{\textaristerikeraia}$ textaristerikeraia

3.3.2 symbol variants

Mathematical notation uses variant shapes of some Greek letters as additional symbols. The variations have no syntactic meaning in Greek text and text fonts may use the variant shapes in place of the “regular” ones as a stylistic choice.

Unicode defines separate code points for the symbol variants. TeX supports some of the variant shape symbols in mathematical mode, but its concept of “standard” vs. “variant” symbols differs from the distinction between “GREEK LETTER ...” vs. “GREEK ... SYMBOL” in the Unicode standard (see Table 1).

`tuenc-greek.def` defines three TextCommands for each of these letters:

\text<name> selects the Unicode GREEK LETTER ... variant,
 \text<name>symbol selects the Unicode GREEK ... SYMBOL variant,
 \textvar<name> selects the variant shape according to TeX’ mathematical mode

See Table 2 for the full list. The *alphabet* package defines short macros that work in text and math mode.

¹the name “stigma” originally applied to a medieval sigma-tau ligature, whose shape was confusingly similar to the cursive digamma

text		mathematics	
macro	output	macro	output
<code>\textpi</code>	π	<code>\pi</code>	π
<code>\textvarpi</code>	ϖ	<code>\varpi</code>	ϖ
<code>\textpisymbol</code>	ϖ		
<code>\textrho</code>	ρ	<code>\rho</code>	ρ
<code>\textvarrho</code>	ϱ	<code>\varrho</code>	ϱ
<code>\textrhosymbol</code>	ϱ		
<code>\texttheta</code>	θ	<code>\theta</code>	θ
<code>\textvartheta</code>	ϑ	<code>\vartheta</code>	ϑ
<code>\textthetasymbol</code>	ϑ		
<code>\textepsilon</code>	ε	<code>\epsilon</code>	ϵ
<code>\textvarepsilon</code>	ε	<code>\varepsilon</code>	ε
<code>\textepsilonsymbol</code>	ϵ		
<code>\textphi</code>	φ	<code>\phi</code>	ϕ
<code>\textvarphi</code>	φ	<code>\varphi</code>	φ
<code>\textphisymbol</code>	ϕ		
<code>\textbeta</code>	β	<code>\beta</code>	β
<code>\textvarbeta</code>	$\bar{\beta}$	<i>missing</i>	
<code>\textbetasymbol</code>	$\bar{\beta}$		
<code>\textkappa</code>	κ	<code>\kappa</code>	κ
<code>\textvarkappa</code>	\varkappa	<code>\varkappa</code>	\varkappa
<code>\textkappasymbol</code>	\varkappa		
<code>\textTheta</code>	Θ	<code>\Theta</code>	Θ
<code>\textvarTheta</code>	Θ	<i>missing</i>	
<code>\textThetasymbol</code>	Θ		

Table 2: Macros for Greek symbol variants

3.3.3 Ancient Greek Numbers

Ancient Greek Numbers are missing in most fonts (including Libertine and Deja Vu). The “FreeSerif” font works fine:

⏏⏏⏏

If the LGR font encoding is loaded via `<fontenc>` in the document preamble, Ancient Greek Numbers (as well as any other character) from LGR encoded 8-bit TeX fonts can be used after a font-encoding switch. (This document defines the `\lgrfont` command for this purpose in the preamble.)

⏏⏏⏏

3.3.4 generic text symbols

There are some LICR macros for some symbols from the 8-bit font encoding LGR that are not confined to Greek but not defined in `tuenc.def` [2018/08/11 v2.0j].

```
; textsemicolon
μ textmicro
ə textschwa
```

The SI unit prefix MICRO SIGN is not upcased with `MakeUppercase`:

```
textmu: μ ↦ M but textmicro: μ ↦ μ.
```

4 Latin transliteration

The Latin transliteration known from LGR encoded 8-bit fonts² does not work with Unicode fonts.

For LuaTeX, there is a `transliteration.omega transform` that applies the transliteration system devised by Yannis Haralambous for the Omega system (cf. the Babel documentation for the [Greek locale](#)).

It is possible to set up LGR encoded fonts parallel to Unicode fonts (see the preamble of the source file `test-tuenc-greek.tex` for an example). After switching the font encoding to LGR, Greek letters can be input via a *Latin transliteration*, e.g. `<logos>` becomes `<λογος>` and `<>αυρη\ 'ια>` becomes `<ἀυρνία>`.

Mark that you cannot use Unicode input with LGR encoded fonts except when running in 8-bit compatibility mode. LICR macros work in both, TU and LGR: compare `Ἰανουαρίου` (TU) vs. `Ἰανουαρίου` (LGR).

5 UPPERCASE and lowercase

According to Greek typesetting conventions, Greek diacritics (except the dialytika and sub-iota) are placed to the left (instead of above) capital letters and dropped if text is set in ALL CAPS, e.g. `μαΐστρος` ↦ `ΜΑΪΣΤΡΟΣ`. The macron and breve accents are handled as in Latin texts.

²See the [teubner](#) package or the file `usage.pdf` from [babel-greek](#) for a description.

The new `\MakeUppercase` implementation (rolled out in the [June 2022 LaTeX release](#)) relies on Unicode data. It upcases according to the Greek conventions **if the text language is set to Greek** with `babel` or `polyglossia`.

For the pre-2022/06 implementation, `tuenc-greek.def` contains code to get the same effect. The `uccode/lccode` corrections (taken from Apostolos Syropoulos `xgreek` package) ensure dropping of accents with `\MakeUppercase` for literal Unicode characters.

`@uclclist` additions ensure that upcasing also drops Greek diacritics. However, when the `tonos`, `varia`, and `perispomeni` accents are input using the symbol macros (`\' \` \~`), this does not work, as they cannot be distinguished from Latin acute, grave, and tilde accents. If these accents should be dropped by `MakeUppercase`, they must be input as named macro:

`άά άά àà àà` \mapsto `AA AÁ AÀ AÀ`

5.1 hiatus

`Tonos` and `dasia` mark a *hiatus* (break-up of a diphthong) if placed on the first of two consecutive vowels (`άι`, `άυ`, `έι`, `έυ`, `ǎι`, `ǎυ`, `ěι`). A `dialytika` must be placed on the second vowel if they are dropped.

The «hiatus» feature works with macro input:

`άυλος` \mapsto `AǎΛΟΣ`, `ǎυλος` \mapsto `AǎΛΟΣ`,
`μάινα` \mapsto `MAǎNA`, `κέικ` \mapsto `KEǎK`, `άυπνία` \mapsto `AǎΠΝΙΑ`.

Since the 2022/06 LaTeX release, it works with monotonic Unicode literals (since 2023/02/10 also some polytonic)

`άι`, `άυ`, `έι`, `έυ`, `ǎι`, `ǎυ`, `ěι`, `άυ` \mapsto `AI`, `AY`, `EI`, `AI`, `AY`, `EI`, `AY`

but not with the named accent-macro + Unicode literals

`άι`, `άυ`, `έι`, `έυ`, `ǎι`, `ǎυ`, `ěι`, `άυ` \mapsto `AI`, `AY`, `EI`, `AI`, `AY`, `EI`, `AY`

nor with short accent-macro + Unicode literals (yet?)

`άι`, `άυ`, `έι`, `έυ`, `ǎι`, `ǎυ`, `ěι`, `άυ` \mapsto `'AI`, `'AY`, `'EI`, `AI`, `AY`, `EI`, `AY`

6 Character Tables

The following tables list the Greek Unicode characters. In the input, the `LICR` macro is followed by the corresponding literal Unicode character.

6.1 Greek and Coptic Unicode block

Only characters that are supported by LGR encoded TeX fonts have a `LICR` definition.

`''''''''`
`''''''''` `AA` `·` `'E'E` `'H'H` `TI` `'O'O` `'Y'Y` `'Ω'Ω`
`úú` `AA` `BB` `ΓΓ` `ΔΔ` `EE` `ZZ` `HH` `ΘΘ` `II` `KK` `ΛΛ` `MM` `NN` `ΞΞ` `OO`
`ΠΠ` `PP` `ΣΣ` `ΤΤ` `ΥΥ` `ΦΦ` `ΧΧ` `ΨΨ` `ΩΩ` `Ϊ` `Ϋ` `άά` `έέ` `ήή` `ύύ`

