Greek Unicode with 8-bit TeX and *inputenc*

Günter Milde

March 2, 2023

Abstract

The definitions in `lgrenc.dfu` provide UTF-8 support for the Greek script based on *inputenc* and the *LaTeX internal character representation* macros (LICRs) defined in the *greek-fontenc* package.

1 Requirements

The *inputenc* standard package enables the use of non-ASCII characters with 8-bit TeX. However, it misses definitions for Greek characters. The *greek-inputenc* package extends *inputenc* to allow the use of Greek literals in the document source.

As with all *inputenc* definitions, this only works if the active font encoding supports the characters. For the Greek script, this is usually the non-standard *LGR* font encoding set up by *greek-fontenc*.

2 Usage

There are several alternatives to activate Greek Unicode input for 8-bit TeX[1] (see also the source document `greek-utf8.tex`):

- Define the LGR font encoding and the UTF8 input encoding (the order does not matter), e.g.,

  ```
  \usepackage[T1,LGR]{fontenc}
  \usepackage[utf8]{inputenc}
  ```

  Ensure that LGR is the active font encoding whenever a Greek character is used in the text (see below).

- For text in the Greek language, it is recommended to use the *Babel* package with the Greek language definitions in *babel-greek*. Babel sets the font encoding automatically to LGR and Greek Unicode characters work as expected. Write in the preamble, e.g.,

[1] The XeTeX and LuaTeX engines use utf8 as native input encoding. They do not require (and, except in 8-bit compatibility mode, do not work with) the *inputenc* and *greek-inputenc* packages.

```
\usepackage[utf8]{inputenc}
\usepackage[LGR,T1]{fontenc}
\usepackage[english,greek,german]{babel}
```

and use `\foreignlanguage` or `\selectlanguage` to set the text language to Greek (see the *babel-greek* documentation for detailled examples).

> Τί φήις; Ἰδὼν ἐνθέδε παῖδ' ἐλευθέραν τὰς πλησίον Νύμφας στεφανοῦσαν, Σώστρατε, ἐρῶν ἀπῆλθες εὐθύς;

- Greek literal characters can also be used in PDF-strings (bookmarks and ToC entries with *hyperref*). See *greek-fontenc* for a hyperref test and usage example.
- In combination with the *textalpha* package from *greek-fontenc*, Greek Unicode characters can be used in text with any font encoding – just like the symbols provided by the "textcomp" package (i.e. with some limitations described in textalpha-doc). With the preamble lines

  ```
  \usepackage[utf8]{inputenc}
  \usepackage{textalpha}
  ```

  it is straightforward to write about π-mesons, γ-radiation, or a 50 kΩ resistor.[2]
- In combination with the *alphabeta* package (also from *greek-fontenc*), Greek Unicode literals can also be used in math mode:

  ```
  \usepackage[utf8]{inputenc}
  \usepackage{alphabeta}
  ```

$$\tan\beta = \frac{\sin\beta}{\cos\beta}.$$

3 Warning: unsafe ASCII input

LGR is no "standard font encoding". Latin characters and some other ASCII symbols are mapped to Greek equivalents if LGR is the active font encoding. (See usage.pdf for a description of this Latin-Greek transliteration.)

This means you need an explicit language and/or font-encoding switch for Latin words and abbreviations in Greek text, e.g., not «ηία αντίσταση 750-κΩ» but «ηία αντίσταση 750-kΩ»

Special care is also required with the question mark characters:

- The Unicode standard says character 003B SEMICOLON and not 037E GREEK QUESTION MARK, is the preferred character for a 'Greek question mark' (erotimatiko),

[2]The MIKRO SIGN and OHM SIGN characters are set up by inputenc with *textcomp* characters for any font encoding while GREEK CAPITAL LETTER OMEGA works only with the LGR font encoding.

- The LGR font encoding maps a SEMICOLON to a middle dot (ano teleia), while the Latin question mark "?" is mapped to the erotimatiko.

As a result, only the deprecated character 037E GREEK QUESTION MARK works with both, Xe/LuaTeX and 8-bit TeX. Compare the source `greek-utf8.tex` and the PDF output:

Latin (T1)	Greek (LGR)	question mark character
Τί φήις;	Τί φήις;	037E GREEK QUESTION MARK
Τί φήις;	Τί φήις·	003B SEMICOLON
Τί φήις?	Τί φήις;	003F QUESTION MARK

With the *textalpha* package's "keep-semicolon" option, the SEMICOLON character can be used for the erotimatiko also with LGR encoded fonts.

4 Supported Characters

Unicode definitions exist for all non-ASCII characters that can be rendered with an LGR-encoded font.

4.1 Greek and Coptic

	0	1	2	3	4	5	6	7	8	9	A	B	C	Δ	E	Φ
370	*	*	*	*	ʹ	͵	*	*				*	*	*	;	
380					΄	΅	Ά	·	Έ	Ή	Ί		Ό		Ύ	Ώ
390	ΐ	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο
3A0	Π	Ρ		Σ	Τ	Υ	Φ	Χ	Ψ	Ω	Ϊ	Ϋ	ά	έ	ή	ί
3B0	ΰ	α	β	γ	δ	ε	ζ	η	ϑ	ι	κ	λ	μ	ν	ξ	ο
3C0	π	ρ	ς	σ	τ	υ	φ	χ	ψ	ω	ϊ	ϋ	ό	ύ	ώ	
3D0	*	*	*	*	*	*	*	*	Ϙ	ϙ	Ϛ	ϛ	F	ϝ	*	ϟ
3E0	Ϡ	ϡ	*	*	*	*	*	*	*	*	*	*	*	*	*	*
3F0	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

legend: * glyph missing in LGR, [space] Unicode point not defined

4.2 Greek Extended

	0	1	2	3	4	5	6	7	8	9	A	B	C	Δ	E	Φ
1F00	ἀ	ἁ	ἂ	ἃ	ἄ	ἅ	ἆ	ἇ	Ἀ	Ἁ	Ἂ	Ἃ	Ἄ	Ἅ	Ἆ	Ἇ
1F10	ἐ	ἑ	ἒ	ἓ	ἔ	ἕ			Ἐ	Ἑ	Ἒ	Ἓ	Ἔ	Ἕ		
1F20	ἠ	ἡ	ἢ	ἣ	ἤ	ἥ	ἦ	ἧ	Ἠ	Ἡ	Ἢ	Ἣ	Ἤ	Ἥ	Ἦ	Ἧ
1F30	ἰ	ἱ	ἲ	ἳ	ἴ	ἵ	ἶ	ἷ	Ἰ	Ἱ	Ἲ	Ἳ	Ἴ	Ἵ	Ἶ	Ἷ
1F40	ὀ	ὁ	ὂ	ὃ	ὄ	ὅ			Ὀ	Ὁ	Ὂ	Ὃ	Ὄ	Ὅ		
1F50	ὐ	ὑ	ὒ	ὓ	ὔ	ὕ	ὖ	ὗ		Ὑ		Ὓ		Ὕ		Ὗ
1F60	ὠ	ὡ	ὢ	ὣ	ὤ	ὥ	ὦ	ὧ	Ὠ	Ὡ	Ὢ	Ὣ	Ὤ	Ὥ	Ὦ	Ὧ
1F70	ὰ	ά	ὲ	έ	ὴ	ή	ὶ	ί	ὸ	ό	ὺ	ύ	ὼ	ώ		
1F80	ᾀ	ᾁ	ᾂ	ᾃ	ᾄ	ᾅ	ᾆ	ᾇ	ᾈ	ᾉ	ᾊ	ᾋ	ᾌ	ᾍ	ᾎ	ᾏ
1F90	ᾐ	ᾑ	ᾒ	ᾓ	ᾔ	ᾕ	ᾖ	ᾗ	ᾘ	ᾙ	ᾚ	ᾛ	ᾜ	ᾝ	ᾞ	ᾟ
1FA0	ᾠ	ᾡ	ᾢ	ᾣ	ᾤ	ᾥ	ᾦ	ᾧ	ᾨ	ᾩ	ᾪ	ᾫ	ᾬ	ᾭ	ᾮ	ᾯ
1FB0	ᾰ	ᾱ	ᾲ	ᾳ	ᾴ		ᾶ	ᾷ	Ᾰ	Ᾱ	Ὰ	Ά	ᾼ	᾽	ι	῀
1FC0	῁	ῂ	ῃ	ῄ	῅		ῆ	ῇ	Ὲ	Έ	Ὴ	Ή	ῌ	῍	῎	῏
1FD0	ῐ	ῑ	ῒ	ΐ			ῖ	ῗ	Ῐ	Ῑ	Ὶ	Ί		῝	῞	῟
1FE0	ῠ	ῡ	ῢ	ΰ	ῤ	ῥ	ῦ	ῧ	Ῠ	Ῡ	Ὺ	Ύ	Ῥ	῭	΅	`
1FF0			ῲ	ῳ	ῴ		ῶ	ῷ	Ὸ	Ό	Ὼ	Ώ	ῼ	´	῾	

4.3 Other Unicode Blocks

Latin-1 Supplement : ¨ « ¯ ´ · »

IPA Extensions : ə LATIN SMALL LETTER SCHWA

Spacing Modifier Letters : ˘α (here followed by letter alpha)

General Punctuation : – — ‘ ’ ‰ ZWNJ (zero width no joiner, prevents kerning and ligatures, e.g. Aϒ vs. Aϒ and ´α vs. ά)

Currency Symbols : €

Letter-like Symbols : Ω

Ancient Greek Numbers : 𝈇 𝈇 𝈉 𝈉 𝈊 𝈊 𝈋 𝈋

5 Test up/downcasing

Capital Greek letters have diacritics (except the dialytika, macron, and breve) to the left (instead of above) and drop them in uppercase, e.g. μαΐστρος ↦ ΜΑΪΣΤΡΟΣ.

Tonos and dasia on the first vowel of a diphthong (άι, άυ, έι) imply a *hiatus*. A dialytika must be placed on the second vowel if they are dropped (AΪ, AΫ, EΪ).

The implementation of `\MakeUppercase` changed in the 2022/06 LaTeX release (cf. LaTeX News 35). Since then, Greek uppercase rules are only applied if the text language is set to “greek” with Babel.

With the “old” `\MakeUppercase`, the auto-hiatus feature in LGR works with the Latin transliteration and LICR macros (AΪ, AΫ, EΪ) but not with literal Unicode characters.

It works also if only the first character of the diphthong is wrapped in `\ensuregreek` (as done by the lgrenc.dfu definition for accented characters)) or input as literal Unicode character (AÏ, AΫ, AÏ) but not if the second character of the diphthong is a Unicode literal (AI, AΥ, EI).

As an ugly workaround, use `\textiota` resp. `\textupsilon` for the character that should get the diaresis: ἀυπνία ↦ AΫΠΝΙΑ.

With the "new" (post 2022/06) `\MakeUppercase`, the auto-hiatus feature in LGR works with literals from the "Greek and Coptic" Unicode block and some from "Greek extended". See the following examples: άυλος ↦ AΥΛΟΣ, ἄυλος ↦ AΥΛΟΣ, μάινα ↦ MAINA, κέικ, ↦ KEIK, ἀυπνία ↦ AΥΠΝΙΑ.

The following subsections test MakeUppercase and MakeLowercase with all characters defined in lgrenc.dfu:

5.1 Greek and Coptic

Characters of the Greek and Coptic Unicode Block:

´„; ´ ˝·΄Ά·Έ·Ή·Ί·Ό·Ύ·Ώ·ΐΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧΨΩΪΫϚϜFϞ
άέήίΰαβγδεζηθικλμνξοπρςστυφχψωϊϋόύώϚϝϜϞ

MakeUppercase:

´„; ¨·΅·ΕΗΙΟΥΩΪΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧΨΩΪΫϚϜFϞ
ΑΕΗΙΥΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΣΤΥΦΧΨΩΪΥΟΥΩϚϜFϞ

Letters and ypogegrammeni upcased, tonos dropped, dialytika kept.

There is no capital Koppa in LGR, therefore ϟ is left unchanged with MakeUppercase.

MakeLowercase:

´„; ´ ΅·ά·έήίόύώΐαβγδεζηθικλμνξοπρστυφχψωϊϋϚϝϜ
άέήίΰαβγδεζηθικλμνξοπρςστυφχψωϊϋόύώϚϝϟϞ

The lowercase of Σ is the «auto-sigma» (`\textautosigma`): ΣΣ ↦ σς. Add a ZWNJ or use the `\noboundary` macro to prevent conversion to final sigma: σσ. The lowercase of GREEK LETTER STIGMA Ϝ is ϝ.

5.2 Greek extended

MakeUppercase:

Α Α Α Α Α Α Α Α Α Α Α Α Α Α Α Α
Ε Ε Ε Ε Ε Ε Ε Ε Ε Ε Ε
Η Η Η Η Η Η Η Η Η Η Η Η Η Η Η Η Η
Ι Ι Ι Ι Ι Ι Ι Ι Ι Ι Ι Ι Ι Ι Ι Ι
Ο Ο Ο Ο Ο Ο Ο Ο Ο Ο Ο Ο
Υ Υ Υ Υ Υ Υ Υ Υ Υ Υ Υ Υ
Ω Ω Ω Ω Ω Ω Ω Ω Ω Ω Ω Ω Ω Ω Ω
Α Α Ε Ε Η Η Ι Ι Ο Ο Υ Υ Ω Ω
A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I A_I
H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I H_I
$Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$ $Ω_I$

Ă Ā Aᵢ Aᵢ Aᵢ A Aᵢ Ă Ā A A Aᵢ ᵢ
¨ Hᵢ Hᵢ Hᵢ H Hᵢ E E H H Hᵢ
Ĭ Ī Ï Ĭ I Ï Ĭ Ī I I
Ῠ Ῡ Ῡ Ῡ P P Υ Ῡ Ῠ Ῡ Υ Υ P ¨¨
Ωᵢ Ωᵢ Ωᵢ Ω Ωᵢ O O Ω Ω Ωᵢ

MakeLowercase:

ἀ ἁ ἂ ἃ ᾰ ᾱ ᾰ ᾱ ἀ ἁ ἂ ἃ ᾰ ᾱ ᾰ ᾱ
ἐ ἑ ἒ ἓ ἔ ἕ ἐ ἑ ἒ ἓ ἔ ἕ
ἠ ἡ ἢ ἣ ᾐ ᾑ ᾒ ᾓ ἠ ἡ ἢ ἣ ᾐ ᾑ ᾒ ᾓ
ἰ ἱ ἲ ἳ ἴ ἵ ἰ ἱ ἲ ἳ ἴ ἵ
ὀ ὁ ὂ ὃ ᾰ ᾱ ὀ ὁ ὂ ὃ ᾰ ᾱ
ὐ ὑ ὒ ὓ ᾰ ᾱ ᾰ ᾱ ὐ ὑ ὒ ὓ ᾰ ᾱ
ὠ ὡ ὢ ὣ ᾰ ᾱ ᾰ ᾱ ὠ ὡ ὢ ὣ ᾰ ᾱ ᾰ ᾱ
ὰ ά ὲ έ ὴ ή ὶ ί ὸ ό ὺ ύ ὼ ώ
ᾀ ᾁ ᾂ ᾃ ᾰ ᾱ ᾰ ᾱ ᾀ ᾁ ᾂ ᾃ ᾰ ᾱ ᾰ ᾱ
ᾐ ᾑ ᾒ ᾓ ᾰ ᾱ ᾰ ᾱ ᾐ ᾑ ᾒ ᾓ ᾰ ᾱ ᾰ ᾱ
ᾠ ᾡ ᾢ ᾣ ᾰ ᾱ ᾰ ᾱ ᾠ ᾡ ᾢ ᾣ ᾰ ᾱ ᾰ ᾱ
ᾰ ᾱ ᾲ ᾳ ᾴ ᾰ ᾱ ᾰ ᾱ ᾲ ᾳ ᾴ ᾳ ᾳ
῁ ῂ ῃ ῄ ῅ ῆ ὲ έ ὴ ή ῆ ῁ ῂ ῃ ῄ ῅
ῐ ῑ ῒ ΐ ῖ ῗ ῐ ῑ ῒ ΐ ῖ ῗ
ῠ ῡ ῢ ΰ ῤ ῥ ῦ ῧ ῠ ῡ ῢ ΰ ῤ ῥ ῦ ῧ
ῲ ῳ ῴ ῷ ῲ ὸ ό ὼ ώ ῳ ῲ ῳ ῴ ῷ

5.2.1 Iota subscript vs. iota adscript

Pre-composed capital letters with *mute iota* decompose to the base letter and COMBINING GREEK YPOGEGRAMMENI (U+0345) even if they are named ... WITH [... AND] PROSGEGRAMMENI for "historic reasons".[3]

Accordingly, the "canonical" LICR for capital letters with mute iota is the base character LICR followed by `\ypogegrammeni`.

Compare letters followed by `\prosgegrammeni` and `\ypogegrammeni` with the pre-composed characters and with character + literal GREEK YPOGEGRAMMENI.

prosgegrammeni: ᾳᾳ ᾼᾳ / AᵢAᵢ AᵢAᵢ / ᾺᵢᾺᵢ Ὰᵢᾼᵢ
MakeUppercase AᵢAᵢ AᵢAᵢ / AᵢAᵢ AᵢAᵢ / AᵢAᵢ AᵢAᵢ
MakeLowercase ᾳᾳ ᾼᾳ / ᾳᾳ ᾼᾳ / ᾲᾲ ᾲᾼ
ypogegrammeni: ᾳᾳ ᾼᾳ / AᵢAᵢ AᵢAᵢ / ᾺᵢᾺᵢ Ὰᵢᾼᵢ
MakeUppercase AᵢAᵢ AᵢAᵢ / AᵢAᵢ AᵢAᵢ / AᵢAᵢ AᵢAᵢ
MakeLowercase ᾳᾳ ᾼᾳ / ᾳᾳ ᾼᾳ / ᾲᾲ ᾲᾼ

5.3 Other Unicode Blocks

MakeUppercase does not change non-letter symbols and the letter shwa (there is a capital Cyrillic schwa in T2A encoded fonts):

¨ « ¯ ′ · » ə ˘A – — ‘ ’ ‰ AΥ ∈ 🄰 🄷 🅇 🄼

[3] cf. Nick Nicholas 'Titlecase and Adscripts'

MakeLowercase does not change non-letter symbols either:

¨ « ¯ ´ · » ə ˘α – — ‘ ’ ‰ αυ € 𝔸 ℍ 𝕏 𝕄

6 Test kerning/ligatures

check for kerning and unwanted ligatures:

Aἀα Aἁα Aἂα Aἃα Aἄα Aἅα Aἆα Aἇα A᾽Aα A῍Aα A῎Aα A῏Aα A῾Aα A῾Aα A῾Aα

Aἐα Aἑα Aἒα Aἓα Aἔα Aἕα A᾽Eα A῍Eα A῎Eα A῏Eα A῾Eα A῾Eα

Aἠα Aἡα Aἢα Aἣα Aἤα Aἥα Aἦα Aἧα A᾽Hα A῍Hα A῎Hα A῏Hα A῾Hα A῾Hα A῾Hα

Aἰα Aἱα Aἲα Aἳα Aἴα Aἵα Aἶα Aἷα A᾽Iα A῍Iα A῎Iα A῏Iα A῾Iα A῾Iα A῾Iα

Aὀα Aὁα Aὂα Aὃα Aὄα Aὅα A᾽Oα A῍Oα A῎Oα A῏Oα A῾Oα A῾Oα

Aὐα Aὑα Aὒα Aὓα Aὔα Aὕα Aὖα Aὗα A῍Yα A῎Yα A῾Yα A῾Yα

Aὠα Aὡα Aὢα Aὣα Aὤα Aὥα Aὦα Aὧα A᾽Ωα A῍Ωα A῎Ωα A῏Ωα A῾Ωα A῾Ωα A῾Ωα

Aὰα Aάα Aὲα Aέα Aὴα Aήα Aὶα Aία Aὸα Aόα Aὺα Aύα Aὼα Aώα

Aᾀα Aᾁα Aᾂα Aᾃα Aᾄα Aᾅα Aᾆα Aᾇα A᾽Aια A῍Aια A῎Aια A῏Aια A῾Aια A῾Aια A῾Aια

Aᾐα Aᾑα Aᾒα Aᾓα Aᾔα Aᾕα Aᾖα Aᾗα A᾽Hια A῍Hια A῎Hια A῏Hια A῾Hια A῾Hια A῾Hια

Aᾠα Aᾡα Aᾢα Aᾣα Aᾤα Aᾥα Aᾦα Aᾧα A᾽Ωια A῍Ωια A῎Ωια A῏Ωια A῾Ωια A῾Ωια A῾Ωια

Aᾰα Aᾱα Aᾲα Aᾳα Aᾴα Aᾶα Aᾷα AᾸα AᾹα AᾺα AΆα AAια A᾽α Aια A᾽α

A῀α A῁α Aῂα Aῃα Aῄα Aῆα Aῇα AῈα AΈα AῊα AΉα AHια A῎α A῾α A῀α

Aῐα Aῑα Aῒα Aΐα Aῖα Aῗα AῘα AῙα AῚα AΊα A῏α A῾α A῀α

Aῠα Aῡα Aῢα Aΰα Aῤα Aῥα Aῦα Aῧα AῨα AῩα AῪα AΎα APα A῞α A῟α A῎α

Aῲα Aωα Aώα Aῶα Aῷα AῺα AΌα Aῼα AΩια A´α A῾α