Greek Unicode with 8-bit TeX

Günter Milde

March 21, 2023

Abstract

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

Contents

1 Introduction

The default input encoding for 8-bit LaTeX changed from 7-bit ASCII to UTF-8 in April 2018.[1] However, the standard setup misses definitions for Greek Unicode characters. *Greek-inputenc* adds definitions to allow the use of literal characters for Greek letters and symbols in the document source.

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

2 Usage

Since 2018, it is no longer necessary to load the *inputenc* package for UTF-8 encoded sources.[2] The character definitions in the file `lgrenc.dfu` are automatically loaded, if the LGR font encoding is loaded by one of the following alternatives:

- With *fontenc*, e.g.,

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

  Ensure that LGR is the active font encoding whenever a Greek character is used in the text (see fntguide.pdf for font encoding switching commands).

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

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

[2] The legacy input encodings *iso-8859-7* and *macgreek* are selected by giving them as options to the *inputenc* package.

- 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.,

  ```
  \usepackage[english,greek,german]{babel}
  ```

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

- 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{textalpha}
  ```

  it is straightforward to write about π-mesons, γ-radiation, or a 50 kΩ resistor.[3] Words and phrases should be wrapped in `\ensuregreek` to preserve kerning or the Babel command `\foreignlanguage{greek}` to also ensure correct hyphenation.

- In combination with the *alphabeta* package (also from *greek-fontenc*), Greek Unicode literals can also be used in math mode:

  ```
  \usepackage{alphabeta}
  ```

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

- 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.

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 U+003B SEMICOLON and not U+037E GREEK QUESTION MARK, is the preferred character for a "Greek question mark" (erotimatiko),

[3] The MICRO SIGN and OHM SIGN characters are set up 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.

Only the deprecated character U+037E GREEK QUESTION MARK works with both, Xe/LuaTeX and 8-bit TeX. However, Unicode treats it as equivalent to U+003B SEMICOLON so a quote copy-pasted from a source using U+037E may end up with U+003B and middle dots instead of erotimatiko! Compare the source `greek-utf8.tex` and the PDF output:

Input	T1	LGR
003F QUESTION MARK	?	;
037E GREEK QUESTION MARK	not defined	;
003B SEMICOLON	;	·
00B7 MIDDLE DOT	·	·

With the babel-greek language attribute "keep-semicolon" or 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	D	E	F
␣	␣	␣	␣	ʼ	ˊ	␣	␣			̓	␣	␣	␣	;	
				ʼ	ˊ	Ά	·	Έ	Ή	Ί		Ό		Ύ	Ώ
ΐ	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο
Π	Ρ		Σ	Τ	Υ	Φ	Χ	Ψ	Ω	Ϊ	Ϋ	ά	έ	ή	ί
ΰ	α	β	γ	δ	ε	ζ	η	ϑ	ι	κ	λ	μ	ν	ξ	ο
π	ρ	ς	σ	τ	υ	φ	χ	ψ	ω	ϊ	ϋ	ό	ύ	ώ	
␣	␣	␣	␣	␣	␣	␣	␣	Ϙ	ϙ	Ϛ	ϛ	F	ϝ	␣	ϟ
Ⅎ	ↄ	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣
␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣

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

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

Currency Symbols: €

Letter-like Symbols: Ω

Ancient Greek Numbers: 𝈅 𝈆 𝈇 𝈈 𝈉 𝈊 𝈋 𝈌

5 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.

The implementation of `\MakeUppercase` changed significantly 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. See *babel-greek* for details and a comprehensive test document.

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

Aἐα Aἑα Aἒα Aἓα Aἔα Aἕα A᾽Εα A῎Εα A῾Εα A᾽Εα A῾Εα A῾Εα

Aἠα Aἡα Aἢα Aἣα Aἤα Aἥα Aἦα Aἧα A᾽Ηα A῎Ηα A῾Ηα A᾽Ηα A῾Ηα A῾Ηα A῾Ηα A῾Ηα

Aἰα Aἱα Aἲα Aἳα Aἴα Aἵα Aἶα Aἷα A᾽Ια A῎Ια A῾Ια A᾽Ια A῾Ια A῾Ια A῾Ια A῾Ια

Aὀα Aὁα Aὂα Aὃα Aὄα Aὅα A᾽Οα A῎Οα A῾Οα A᾽Οα A῾Οα A῾Οα

Aὐα Aὑα Aὒα Aὓα Aὔα Aὕα Aὖα Aὗα A῎Υα A῾Υα A῾Υα A῾Υα

Aὠα Aὡα Aὢα Aὣα Aὤα Aὥα Aὦα Aὧα A᾽Ωα 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 A῾Aια

Aᾐα Aᾑα Aᾒα Aᾓα Aᾔα Aᾕα Aᾖα Aᾗα A᾽Ηια A῎Ηια A῾Ηια A᾽Ηια A῾Ηια A῾Ηια A῾Ηια A῾Ηια

Aᾠα Aᾡα Aᾢα Aᾣα Aᾤα Aᾥα Aᾦα Aᾧα A᾽Ωια 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 A῍α 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 AΡα A῭α A΅α A`α

Aὠα Aὡα Aὢα Aὣα Aὤα Aὥα Aὦα Aὧα AῸα AΌα AῺα AΏα AΩια A´α A῍α