Greek Unicode with 8-bit TeX and *inputenc*

Günter Milde

August 7, 2015

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.,

  ```
  \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).

[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.

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

- 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.

- 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),
- 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

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
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	Ϝ	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	D	E	F
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. AϒA vs. AϒA 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) to the left (instead of above) and drop them in uppercase, e.g. μαΐστρος ↦ MAΪΣΤΡΟΣ.

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 auto-hiatus feature in lgrxenc.def works with the Latin transcription and with character-macros (AΪ, AΫ, EΪ) and also if the first character is wrapped in `\ensuregreek` (as done by the lgrenc.dfu definition for accented characters) or a literal Unicode character (AΪ, AΫ, AΪ) but not if the second character of the diphthong is a Unicode literal (AI, AΥ, EI).

Therefore, the diaresis is missing in the following examples: άυλος ↦ AΥΛΟΣ, ἄυλος ↦ AΥΛΟΣ, μάινα ↦ MAINA, κέικ, ↦ KEIK, ἀυπνία ↦ AΥΠΝΙΑ.

Fixing this shortcoming requires knowledge of what `\LGR@ifnextchar` "sees" when the next character is an upcased Unicode literal.

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

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:

ʹ͵; ʹ ʹ͵Α·ΈΉΊΌΎΏΐΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧΨΩΪΫϚϜFϞ
άέήίΰαβγδεζηθικλμνξοπρςστυφχψωϊϋόύώϚϝFϟϡ

MakeUppercase:

ʹ͵; ¨Α·ΕΗΙΟΥΩΪΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧΨΩΪΫϚϜFϞ
ΑΕΗΙΥΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΣΤΥΦΧΨΩΪΥΟΥΩϚϜFϟϡ

Letters and ypogegrammeni upcased, tonos dropped, dialytika kept.

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

MakeLowercase:

ʹ͵; ʹ ʹά·έήίόύώΐαβγδεζηθικλμνξοπρστυφχψωϊϋϚϝϡ
άέήίΰαβγδεζηθικλμνξοπρςστυφχψωϊϋόύώϚϝFϟϡ

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
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 Ω
Α Α Ε Ε Η Η Ι Ι Ο Ο Υ Υ Ω Ω

Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ Aᵢ
Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ Hᵢ
Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ Ωᵢ
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 Ĭ Ī I Ĭ Ī Ī I I
Ῠ Ῡ Ῠ Ῡ P P Υ Ῡ Ῠ Ῡ Υ Υ P ¨ ¨
Ωᵢ Ωᵢ Ωᵢ Ω Ωᵢ O O Ω Ω Ωᵢ

MakeLowercase:

ả ả ả ả ắ ắ ắ ả ả ả ả ắ ắ ắ
ẻ ẻ ẻ ẻ ế ẻ ẻ ẻ ẻ ế ế
ỉ ỉ ỉ ỉ ố ố ố ỉ ỉ ỉ ỉ ố ố ố
ỏ ỏ ỏ ỏ ố ố ỏ ỏ ỏ ỏ ố ố
ủ ủ ủ ủ ố ố ố ủ ủ ủ ố ố
ở ở ở ở ớ ớ ớ ở ở ở ở ớ ớ ớ
à á è é ỉ ỉ ì í ò ó ù ú ở ở
ợ ợ ợ ợ ợ ợ ợ ợ ợ ợ ợ ợ ợ ợ
ỉ ỉ ỉ ỉ ố ố ố ỉ ỉ ỉ ỉ ố ố ố
ợ ợ ợ ợ ớ ớ ớ ợ ợ ợ ợ ớ ớ ớ
ă ā ợ ợ ợ ã ợ ă ā à á ợ ̓ ̓
˜ ˜ ỉ ỉ ỉ ố ỉ è é ỉ ỉ ỉ ̃ ̃ ̃
ĭ ī ĭ ĭ ĭ ĭ ī ì í ̃ ̃ ̃
ŭ ū ŭ ŭ ợ ợ ũ ũ ŭ ū ù ú ợ ̃ ́ ̀
ở ở ở ũ ở ò ó ở ở ở ́ ̓

5.3 Other Unicode Blocks

MakeUppercase does not change non-letter symbols and the letter shwa:

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

MakeLowercase does not change non-letter symbols, too:

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

6 Test kerning/ligatures

check for kerning and unwanted ligatures:

Aả̆α Aả̈α Aã̆α Aã̈α Aă̆α Aă̈α Aã̆α Aã̈α A̓Aα A̋Aα A̋Aα A̋Aα
A̋Aα A̋Aα A̋Aα A̋Aα

Aẻ̆α Aẻ̈α Aẽ̆α Aẽ̈α Aĕ̆α Aĕ̈α A̓Eα A̋Eα A̋Eα A̋Eα A̋Eα A̋Eα

Aỉ̆α Aỉ̈α Aĩ̆α Aĩ̈α Aĭ̆α Aĭ̈α Aĩ̆α Aĩ̈α A̓Hα A̋Hα A̋Hα A̋Hα
A̋Hα A̋Hα A̋Hα A̋Hα

Aỉ̆α Aỉ̈α Aĩ̆α Aĩ̈α Aĭ̆α Aĭ̈α Aĩ̆α Aĩ̈α A̓Iα A̋Iα A̋Iα A̋Iα A̋Iα A̋Iα
A̋Iα A̋Iα

Aỏ̆α Aỏ̈α Aõ̆α Aṏα Aŏ̆α Aŏ̈α A̓Oα A̋Oα A̋Oα A̋Oα A̋Oα A̋Oα

Aủ̆α Aủ̈α Aũ̆α Aũ̈α Aŭ̆α Aŭ̈α Aũ̆α Aũ̈α A̋Υα A̋Υα A̋Υα A̋Υα

Aở̆α Aở̈α Aỡ̆α Aỡ̈α Aở̆α Aở̈α Aỡ̆α Aỡ̈α A̓Ωα A̋Ωα A̋Ωα A̋Ωα
A̋Ωα A̋Ωα A̋Ωα A̋Ωα

Aà̆α Aá̈α Aè̆α Aé̈α Aỉ̆α Aí̈α Aì̆α Aí̈α Aò̆α Aó̈α Aù̆α Aú̈α Aờ̆α
Aớ̈α

Aợ̆α Aợ̈α Aợ̆α Aợ̈α Aợ̆α Aợ̈α Aợ̆α Aợ̈α A̓Aᵢα A̋Aᵢα A̋Aᵢα A̋Aᵢα
A̋Aᵢα A̋Aᵢα A̋Aᵢα A̋Aᵢα

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

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

Aᾰα Aᾱα Aὰα Aάα Aᾶα Aᾷα AᾸα AᾹα A᾽Aα A῾Aα AAια A᾽α Aια A᾽α

A῀α A῁α Aὴα Aηα Aήα Aῆα Aῇα A᾽Eα A῾Eα A᾽Hα A῾Hα AHια A῍α A῝α A῎α

Aῐα Aῑα Aὶα Aία Aῖα Aΐα AῘα AῙα A᾽Iα A῾Iα A῍α A῝α A῎α

Aῠα Aῡα Aὺα Aύα Aῤα Aῥα Aῦα Aῧα AῨα AῩα A᾽Yα A῾Yα A῍Pα A῝α A῎α A᾽α

Aὼα Aωα Aώα Aῶα Aῷα A᾽Oα A῾Oα A᾽Ωα A῾Ωα AΩια A῾α A῎α