Greek Unicode with 8-bit TeX and *inputenc*

Günter Milde

June 4, 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 Motivation

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 (see greek-fontec).

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 detailed 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 $50\,\mathrm{k\Omega}$ 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-χΩ» 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					ʹ	ͺ	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. ΑΥ vs. ΑΥ and ᾿α vs. ἀ)

Currency Symbols : €

Letter-like Symbols : Ω

Ancient Greek Numbers : 𐅄 𐅆 𐅈 𐅉

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 ↦ 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 (ΑΪ, ΑΫ, ΕΪ).

The auto-hiatus feature in lgrxenc.def works with the Latin transcription and with character-macros (ΑΪ, ΑΫ, ΕΪ) 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 (ΑΪ, ΑΫ, ΑΪ) but not if the second character of the diphthong is a Unicode literal (ΑΙ, ΑΥ, ΕΙ).

Therefore, the diaresis is missing in the following examples: άυλος ↦ ΑΥΛΟΣ, ἄυλος ↦ ΑΥΛΟΣ, μάινα ↦ ΜΑΙΝΑ, κέικ, ↦ ΚΕΙΚ, ἀυπνία ↦ ΑΥΠΝΙΑ.

Fixing this shortcoming requires knowledge of what the `\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 ↦ ΑΫΠΝΙΑ.

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·ΈΉΊΟΎΩϊΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧΨΩΪΫϏϘϜϞ
άέήίϋαβγδεζηθικλμνξοπρϛστυφχψωϊϋόύώϙϝϟ϶

MakeUppercase:

΄΅· ¨Α·ΕΗΙΟΥΩΪΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧΨΩΪΫϏϘϜϞ
ΑΕΗΙΫΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΣΤΥΦΧΨΩΪΫΟΥΩϘϜϞ

Letters and ypogegrammeni upcased, tonos dropped, dialytika kept.
There is no capital Koppa in LGR, therefore ϟ is left unchanged with MakeUppercase.
MakeLowercase:

΄΅· ᾽ ᾽ ά·έήίόύώϊαβγδεζηθικλμνξοπρστυφχψωϊϋϙϝϟϏ
άέήίϋαβγδεζηθικλμνξοπρϛστυφχψωϊϋόύώϙϝϟϏ

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 Ϛ is ϛ.

5.2 Greek extended

MakeUppercase:
ΑΑΑΑΑΑΑΑΑΑΑΑΑΑ
ΕΕΕΕΕΕΕΕΕΕΕ
ΗΗΗΗΗΗΗΗΗΗΗΗΗΗ
ΙΙΙΙΙΙΙΙΙΙΙΙΙΙΙΙ
ΟΟΟΟΟΟΟΟΟΟΟΟ
ΥΥΥΥΥΥΥΥΥΥΥ
ΩΩΩΩΩΩΩΩΩΩΩΩΩΩΩ
ΑΑΕΕΗΗΙΙΟΟΥΥΩΩ

Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍ Α͍
Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍ Η͍
Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍ Ω͍
Ᾰ Ᾱ Ᾱͅ Ᾱͅ Ᾱͅ Α ᾼ Ᾰ Ᾱ Α Α ᾼ ͅ
¨ Η͍ Η͍ Η͍ Η Η͍ Ε Ε Η Η Η͍
Ῐ Ῑ Ϊ Ι Ϊ Ῐ Ῑ Ι Ι
Ῠ Ῡ Ϋ Ϋ Ρ Ρ Υ Ϋ Ϋ Ῡ Υ Υ Ρ ¨˙
Ω͍ Ω͍ Ω͍ Ω Ω͍ Ο Ο Ω Ω Ω͍

MakeLowercase:

ἀ ά ἂ ἃ ἄ ἅ ἆ ἇ ὰ ά ᾄ ᾅ ᾆ ᾇ

ἐ έ ἒ ἓ ἔ ἕ ὲ έ ἒ ἓ ἔ

ἠ ή ἢ ἣ ἤ ἥ ἦ ἧ ὴ ή ᾔ ᾕ ᾖ ᾗ

ἰ ί ἲ ἳ ἴ ἵ ἶ ἷ ὶ ί ἲ ἳ ἴ ἵ ἶ ἷ

ὀ ό ὂ ὃ ὄ ὅ ὸ ό ὂ ὃ ὄ ὅ

ὐ ύ ὒ ὓ ὔ ὕ ὖ ὗ ὺ ύ ὒ ὓ

ὠ ώ ὢ ὣ ὤ ὥ ὦ ὧ ὼ ώ ᾤ ᾥ ᾦ ᾧ ὤ ὥ

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

ᾐ ᾑ ᾒ ᾓ ᾔ ᾕ ᾖ ᾗ ᾐ ᾑ ᾒ ᾓ ᾔ ᾕ

ᾠ ᾡ ᾢ ᾣ ᾤ ᾥ ᾦ ᾧ ᾠ ᾡ ᾢ ᾣ ᾤ ᾥ

ᾰ ᾱ ᾲ ᾳ ᾴ ᾶ ᾷ ᾰ ᾱ ὰ ά ᾳ ᾿ ῀ ῁

῍ ῎ ῂ η ῄ ῆ ῇ ὲ έ ῂ ή ῇ ῍ ῎ ῏

ῐ ῑ ῒ ΐ ῔ ῕ ῖ ῗ ὶ ί ῐ ῑ ῒ ῏ ῐ ῑ

ῠ ῡ ὺ ύ ῥ ῤ ῠ ῡ ῦ ῧ ὺ ύ ῥ ῭ ΅ ``

ῲ ῳ ῴ ῶ ῷ ὸ ό ὼ ώ ῳ ´˙

5.3 Other Unicode Blocks

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

¨ « ¯ ´ · » ə ˘ Α – — ' ' ‰ ΑΥ € Ⓐ Ⓗ Ⓧ Ⓜ

MakeLowercase does not change non-letter symbols, too:

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

6 Test kerning/ligatures

check for kerning and unwanted ligatures:

Αἀα Αἁα Αἂα Αἃα Αἄα Αἅα Αἆα Αἇα Ἀ̓Αα Ἁ̔Αα Ἂ̓̀Αα Ἃ̔̀Αα Ἄ̓́Αα Ἅ̔́Αα Ἆ̓͂Αα Ἇ̔͂Αα

Αἐα Αἑα Αἒα Αἓα Αἔα Αἕα Ἐ̓Εα Ἑ̔Εα Ἒ̓̀Εα Ἓ̔̀Εα Ἔ̓́Εα Ἕ̔́Εα

Αἠα Αἡα Αἢα Αἣα Αἤα Αἥα Αἦα Αἧα Ἠ̓Ηα Ἡ̔Ηα Ἢ̓̀Ηα Ἣ̔̀Ηα Ἤ̓́Ηα Ἥ̔́Ηα Ἦ̓͂Ηα Ἧ̔͂Ηα

Αἰα Αἱα Αἲα Αἳα Αἴα Αἵα Αἶα Αἷα Ἰ̓Ια Ἱ̔Ια Ἲ̓̀Ια Ἳ̔̀Ια Ἴ̓́Ια Ἵ̔́Ια Ἶ̓͂Ια Ἷ̔͂Ια

Αὀα Αὁα Αὂα Αὃα Αὄα Αὅα Ὀ̓Οα Ὁ̔Οα Ὂ̓̀Οα Ὃ̔̀Οα Ὄ̓́Οα Ὅ̔́Οα

Αὐα Αὑα Αὒα Αὓα Αὔα Αὕα Αὖα Αὗα Υ̓̓Υα Ὑ̔Υα Υ̓̀̓̀Υα Ὓ̔̀Υα

Αὠα Αὡα Αὢα Αὣα Αὤα Αὥα Αὦα Αὧα Ὠ̓Ωα Ὡ̔Ωα Ὢ̓̀Ωα Ὣ̔̀Ωα Ὤ̓́Ωα Ὥ̔́Ωα Ὦ̓͂Ωα Ὧ̔͂Ωα

Αὰα Αάα Αὲα Αέα Αὴα Αήα Αὶα Αία Αὸα Αόα Αὺα Αύα Αὼα Αώα

Αᾀα Αᾁα Αᾂα Αᾃα Αᾄα Αᾅα Αᾆα Αᾇα Ἀ̓ᾼα Ἁ̔ᾼα Ἂ̓̀ᾼα Ἃ̔̀ᾼα Ἄ̓́ᾼα Ἅ̔́ᾼα Ἆ̓͂ᾼα Ἇ̔͂ᾼα

Αἠα Αἠα Αῇα Αῇα Αῄα Αῄα Αῄα Αῄα Αˈ Ηι α Αˊ Ηι α Αˆ Ηι α Αˇ Ηι α
Αˇ Ηι α Αˇ Ηι α Αˇ Ηι α Αˇ Ηι α

Αῴα Αῴα Αῴα Αῴα Αῴα Αῴα Αῴα Αῴα Αˈ Ωι α Αˊ Ωι α Αˆ Ωι α
Αˇ Ωι α Αˇ Ωι α Αˇ Ωι α Αˇ Ωι α Αˇ Ωι α

Αᾰα Αᾱα Αὰα Ααα Αάα Αᾶα Αᾳα ΑᾸα ΑᾹα ΑὰΑα ΑΆΑα ΑΑι α
Αˈα Αι α Αˈα

Α˜α Αˇα Αἠα Αηα Αήα Αῆα Αῄα ΑὲΕα ΑΈΕα ΑὴΗα ΑΉΗα ΑΗι α
Αˆα Αˇα Αˇα

Αῐα Αῑα Αὶα Αία Αῐα Αῐα ΑῘα ΑῙα Αˈ Ια Αˊ Ια Αˆα Αˇα Αˇα

Αῠα Αῡα Αὺα Αύα Αῤα Αῤα Αῦα Αῦα ΑῨα ΑῩα Αˈ Υα Αˊ Υα
Αˈ Ρα Αˈα Αˊα Αˈα

Αῴα Αῳα Αῴα Αῶα Αῷα Αˈ Οα Αˊ Οα Αˈ Ωα Αˊ Ωα ΑΩι α Αˊα Αˈα