Greek Unicode with 8-bit TeX and *inputenc*

Abstract

Small test of Unicode input for Greek letters.

Without the "textalpha" or "alphabeta" packages, Greek Unicode characters must be wrapped in `\ensuregreek{}`, `\textgreek{}`, or `\foreignlanguage{greek}{}` or follow the declarations `\greekscript`, `\greektext`, or `\setlanguage{greek}` to set the correct font encoding (LGR).

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

The MIKRO SIGN and OHM SIGN characters are set up by textcomp for any font encoding:

> With a a 50 kΩ resistor, the current is 20 µA.

However, if the similar looking Greek Unicode characters are used, LGR is required for GREEK CAPITAL LETTER OMEGA. On the other hand, GREEK SMALL LETTER MU works in any font encoding because it translates to `\textmu`, the LICR also used by textcomp for the MICRO SIGN.

> With a a 50 kΩ resistor, the current is 20 µA.

1 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.

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

2 Supported Characters

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

2.1 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	A	B	Γ	Δ	E	Z	H	Θ	I	K	Λ	M	N	Ξ	O
3A0	Π	P		Σ	T	Υ	Φ	X	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

2.2 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	´	῾	

2.3 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 : 𝈅 𝈇 𝈉 𝈌

3 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 (AÏ, AΫ, EÏ).

The auto-hiatus feature in lgrxenc.def works with the Latin transcription and with character-macros (AÏ, AΫ, EÏ) 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 (AÏ, AΫ, AÏ) 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 ↦ ΑΥΛΟΣ, μάινα ↦ MAINA, κέικ, ↦ KEIK, ἀυπνία ↦ ΑΥΠΝΙΑ.

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:

3.1 Greek and Coptic

Characters of the Greek and Coptic Unicode Block:

΄„; ΄ ΅Α·ΈΉΊΌΎΏΐΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧΨΩΪΫϚϜFϞ
άέήίΰαβγδεζηθικλμνξοπρςστυφχψωϊϋόύώϚϝFϟϡ

MakeUppercase:

΄„; ¨Α·ΕΗΙΟΥΩΪΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧΨΩΪΫϚϜFϞ
ΑΕΗΙΫΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΣΤΥΦΧΨΩΪΫΟΥΩϚϜFϟϡ

Letters and ypogegrammeni upcased, tonos dropped, dialytika kept.

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

MakeLowercase:

΄„; ΄ ΅ά·έήίόύώΐαβγδεζηθικλμνξοπρστυφχψωϊϋϚϝF
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.

3.2 Greek extended

MakeUppercase:
Ἀ Ἁ Ἂ Ἃ Ἄ Ἅ Ἆ Ἇ ᾈ ᾉ ᾊ ᾋ ᾌ ᾍ ᾎ ᾏ
Ἐ Ἑ Ἒ Ἓ Ἔ Ἕ Ὲ Έ
Ἠ Ἡ Ἢ Ἣ Ἤ Ἥ Ἦ Ἧ ᾘ ᾙ ᾚ ᾛ ᾜ ᾝ ᾞ ᾟ
Ἰ Ἱ Ἲ Ἳ Ἴ Ἵ Ἶ Ἷ Ὶ Ί
Ὀ Ὁ Ὂ Ὃ Ὄ Ὅ Ό
Ὑ Ὓ Ὕ Ὗ Ὺ Ύ
Ὠ Ὡ Ὢ Ὣ Ὤ Ὥ Ὦ Ὧ ᾨ ᾩ ᾪ ᾫ ᾬ ᾭ ᾮ ᾯ
Ὰ Ά
Ᾰ Ᾱ ᾺΑ ΆΑ ᾼ
Ὲ Έ
ῈΕ ΉΕ ῌ
Ῐ Ῑ Ὶ Ί
Ῠ Ῡ Ὺ Ύ
ῬΡ
Ὸ Ό
ῸΟ ΎΟ ῼ

O O O O O O O O O O O O
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
A A E E H H I I O O 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
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
Ă Ā A_I A_I A_I A A A_I Ă Ā A A A_I $_I$
¨ H_I H_I H_I H H_I E E H H H_I
Ĭ Ī Ĭ Ĭ I Ĭ Ĭ Ī I I
ϒ̆ ϒ̄ ϔ ϔ P P ϒ ϔ ϒ̆ ϒ̄ ϒ ϒ P ¨ ¨
Ω_I Ω_I Ω_I Ω Ω_I O O Ω Ω Ω_I

MakeLowercase:

ȧ ά ᾶ ᾶ ă ă ă ă ȧ ά ᾶ ᾶ ă ă ă ă
ȅ é ȇ ȇ ĕ ĕ ȅ é ȇ ȇ ĕ ĕ
ȟ ή ȟ ȟ ĭ ĭ ĭ ĭ ȟ ή ȟ ȟ ĭ ĭ ĭ ĭ
ȉ í ȋ ȋ ĭ ĭ ĭ ĭ ȉ í ȋ ȋ ĭ ĭ ĭ ĭ
ȯ ό ȏ ȏ ŏ ŏ ȯ ό ȏ ȏ ŏ ŏ
ȕ ύ ȗ ȗ ŭ ŭ ŭ ŭ ȕ ύ ȗ ȗ ŭ ŭ
ὼ ώ ὦ ὦ ŏ ŏ ŏ ŏ ὼ ώ ὦ ὦ ŏ ŏ ŏ ŏ
ὰ ά è é ὴ ή ì í ò ó ὺ ύ ὼ ώ
ᾰ ᾱ ᾲ ᾳ ᾴ ᾷ ᾰ ᾱ ᾲ ᾳ ᾴ ᾷ
ῂ ή ῆ ῆ ĭ ĭ ĭ ĭ ῂ ή ῆ ῆ ĭ ĭ ĭ ĭ
ῲ ώ ῶ ῶ ŏ ŏ ŏ ŏ ῲ ώ ῶ ῶ ŏ ŏ ŏ ŏ
ᾶ ᾱ ᾲ ᾳ ᾴ ᾷ ᾶ ᾱ ᾲ ᾳ ᾴ ᾷ
῁ ή ῆ ῆ ĭ ĭ ĭ ĭ ὲ έ ὴ ή ῆ ῆ ĭ ĭ ĭ ĭ
ῒ ΐ ῖ ῖ ĭ ĭ ĭ ĭ ὶ ί ῖ ῖ ĭ ĭ ĭ ĭ
ῠ ῡ ῦ ῦ ῥ ῥ ŭ ŭ ŭ ŭ ὺ ύ ῥ ῥ
ῳ ῳ ῳ ῶ ῶ ò ó ὼ ώ ῳ ῳ ῳ ῳ

Test the iota subscript: A_IA_I A_IA_I A_IA_I A_IA_I A_IA_I A_IA_I
A_IA_I A_IA_I A_IA_I A_IA_I A_IA_I A_IA_I A_I

3.3 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 ∈ 𝔸 ℍ 𝕏 𝕄

MakeLowercase does not change non-letter symbols either:

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

4 Test kerning/ligatures

check for kerning and unwanted ligatures:

Aȧα Aάα Aᾶα Aᾶα Aᾰα Aᾰα Aᾰα Aᾰα A᾽Aα A῍Aα A῎Aα A῎Aα A῏Aα
A῏Aα A῏Aα A῏Aα

Aȅα Aέα Aȇα Aȇα Aĕα Aĕα A᾽Eα A῍Eα A῎Eα A῎Eα A῏Eα A῏Eα

Aἠα Aἡα Aἢα Aἣα Aἤα Aἥα Aἦα Aἧα 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

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

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

Aᾐα Aᾑα Aᾒα Aᾓα Aᾔα Aᾕα Aᾖα Aᾗα 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

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 Aὺα Aύα Aῤα Aῥα Aῦα Aῧα AῨα AῩα AῪα AΎα A῝α A῞α A῟α

Aῲα Aῳα Aώα Aῶα Aῷα AῸα AΌα AῺα AΏα AΩια A´α A῍α