Test the Greek support for Babel

Günter Milde

October 13, 2023

Test the support for the Greek language as defined in the file `greek.ldf` (source `greek.dtx`).

This document is compiled with pdfLaTeX, format version 2023-06-01 patch-level 1, and the L3 programming layer from 2023-08-29. The Greek font encoding is LGR and the language variant is monotoniko.

Contents

1 Language Switch

The declaration `\selectlanguage` switches between languages.

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

The command `\foreignlanguage` sets its second argument in the language specified as first argument. This is intended for short text parts or single words like Βιβλιοθήκη.

Input may use literal Greek characters (α . . . Ω), LICR macros (α . . . Ω), or (if the LGR font encoding is used) the Latin transliteration (α . . . Ω).

Warning: With 8-bit TeX, Latin letters and some symbols in the input are mapped to Greek equivalents!. Without precautions, quotes copied from external sources (like this Wikpedia entry about the question mark) may come out simply wrong:

> Το **ερωτηματικό** (ελλ. ; , λατ. ;) είναι το σημείο στίξης το οποίο τοποθετείται στο τέλος κάθε ευθείας ερωτηματικής πρότασης σε πολλές γλώσσες.

See section 5.1 for remedies.

There should be no inserted space before or after the language switch (this may happen if there are unescaped linebreaks in the font or language definitions):

> Change script with `\ensuregreek`: |δοῦλος|.
> Change language with `\foreignlanguage`: |δοῦλος|.
> Change language with `\selectlanguage`: |δοῦλος|.

2 Auto-strings

Babel defines macros for several autogenerated strings so that they may appear in the choosen language. *Babel-greek* uses LICR[1] macros in order to let the string macros work independent of the font encoding.

Περίληψη

Look for the abstract name. Today is 13 Οκτωβρίου 2023.

Show the auto-strings for language variant "monotoniko".

2.1 Captions

Περίληψη, βλέπε επίσης, Παράρτημα, Βιβλιογραφία,
Κοινοποίηση, Κεφάλαιο, Περιεχόμενα, Συνημμένα,
Σχήμα, Γλωσσάρι, Προς, Ευρετήριο,
Κατάλογος Σχημάτων, Κατάλογος Πινάκων,
Σελίδα, Μέρος, Πρόλογος, Απόδειξη,
Αναφορές, βλέπε, Πίνακας

Test correct upcasing (dropping of accents):

ΠΕΡΙΛΗΨΗ, ΒΛΕΠΕ ΕΠΙΣΗΣ, ΠΑΡΑΡΤΗΜΑ, ΒΙΒΛΙΟΓΡΑΦΙΑ,
ΚΟΙΝΟΠΟΙΗΣΗ, ΚΕΦΑΛΑΙΟ, ΠΕΡΙΕΧΟΜΕΝΑ, ΣΥΝΗΜΜΕΝΑ,
ΣΧΗΜΑ, ΓΛΩΣΣΑΡΙ, ΠΡΟΣ, ΕΥΡΕΤΗΡΙΟ,

[1] LaTeX internal character representation

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ, ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ,
ΣΕΛΙΔΑ, ΜΕΡΟΣ, ΠΡΟΛΟΓΟΣ, ΑΠΟΔΕΙΞΗ,
ΑΝΑΦΟΡΕΣ, ΒΛΕΠΕ, ΠΙΝΑΚΑΣ

2.2 Months

13 Ιανουαρίου 2023 13 ΙΑΝΟΥΑΡΙΟΥ 2023
13 Φεβρουαρίου 2023 13 ΦΕΒΡΟΥΑΡΙΟΥ 2023
13 Μαρτίου 2023 13 ΜΑΡΤΙΟΥ 2023
13 Απριλίου 2023 13 ΑΠΡΙΛΙΟΥ 2023
13 Μαΐου 2023 13 ΜΑΪΟΥ 2023
13 Ιουνίου 2023 13 ΙΟΥΝΙΟΥ 2023
13 Ιουλίου 2023 13 ΙΟΥΛΙΟΥ 2023
13 Αυγούστου 2023 13 ΑΥΓΟΥΣΤΟΥ 2023
13 Σεπτεμβρίου 2023 13 ΣΕΠΤΕΜΒΡΙΟΥ 2023
13 Οκτωβρίου 2023 13 ΟΚΤΩΒΡΙΟΥ 2023
13 Νοεμβρίου 2023 13 ΝΟΕΜΒΡΙΟΥ 2023
13 Δεκεμβρίου 2023 13 ΔΕΚΕΜΒΡΙΟΥ 2023

3 Hyphenation

Patterns for the Greek language variants:

monotonic: `\l@monogreek` = 13
polytonic: `\l@polygreek` = 12
ancient: `\l@ancientgreek` = 7

current: `\l@greek` = 13

Greek paragraph:

mo-no-to-nic: Ευ-ρε-τήριο, ε-πίσης, Α-πόδει-ξη, Θράκη, τρα-γω-δία
" " tran-slit: Ευ-ρε-τή-ριο, ε-πί-σης, Α-πό-δει-ξη, Θρά-κη, τρα-γω-δί-α
polytonic: Ε-ὑρε-τήριο, ἐπίσης, Ἀ-πόδει-ξη, Θρᾴκη, τραγῳδία
" tran-slit: Εὑρετήριο, ἐπί-σης, Ἀ-πό-δει-ξη, Θρᾴκη, τραγῳδία
ancient: Ε-ὑρε-τήριον, ὡσα-ύτως, Ἀ-πόδει-ξις, Θρᾴκη, τραγῳδία

English paragraph with Greek text (`\foreignlanguage{greek}`):

mo-no-to-nic: Ευ-ρε-τήριο, ε-πίσης, Α-πόδει-ξη, Θράκη, τρα-γω-δία
" " tran-slit: Ευ-ρετήριο, επίσης, Απόδειξη, Θράκη, τρα-γωδία
polytonic: Ε-ὑρε-τήριο, ἐπίσης, Ἀ-πόδει-ξη, Θρᾴκη, τραγῳδία
" tran-slit: Εὑρετήριο, ἐπίσης, Ἀπόδειξη, Θρᾴκη, τραγῳδία
ancient: Ε-ὑρε-τήριον, ὡσα-ύτως, Ἀ-πόδει-ξις, Θρᾴκη, τραγῳδία

English paragraph with Greek script (`\ensuregreek`):

mono-tonic: Ευ-ρετήριο, επίσης, Απόδειξη, Θράκη, τραγ-ωδία
" " translit: Ευ-ρετήριο, επίσης, Απόδειξη, Θράκη, τραγωδία
polytonic: Εὑρετήριο, ἐπίσης, Ἀπόδειξη, Θρᾴκη, τραγῳδία
" translit: Εὑρετήριο, ἐπίσης, Ἀπόδειξη, Θρᾴκη, τραγῳδία
ancient: Εὑρετήριον, ὡσαύτως, Ἀπόδειξις, Θρᾴκη, τραγῳδία

4 Greek Numerals (α′ . . . ͵ϡ͵ϟϑϡϟϑ′)

Babel-Greek provides the macros `\greeknumeral` and `\Greeknumeral` for the conversion of Arabic numbers from 1 to 999 999 into their Greek counterparts (α′, β′, γ′, . . . , ͵ϡ͵ϟϑϡϟϑ′). See babel-greek-doc for the formation rules and configuration options and `test-greeknum.pdf` for samples.

Examples:

36 = λϛ′ 94 = ϟδ′ 678 = χοη′ 2002 = ͵ββ′ 923090 = ͵ϡ͵χ͵γϟ′

36 = ΛϚ′ 94 = ϟΔ′ 678 = XOH′ 2002 = ͵BB′ 923090 = ͵Ϡ͵Κ͵Γϟ′

Users can redefine the macros `\greeknumeralsix` and `\greeknumeralSix` as well as `\greeknumeralninety` `\greeknumeralNinety` to configure the used symbols.

If a font misses glyphs for the Greek numeral signs, subsitute characters may be defined with the macros `\textdexiakeraia` and `\textaristerikeraia`.

Example (use "archaic kappa", "varstigma" with pdftex and substitute chars for the numeral signs with Xe/LuaTeX):

36 = λϛ′ 94 = ϙδ′ 678 = χοη′ 2002 = ͵ββ′ 923090 = ͵ϡ͵χ͵γϙ′

36 = Λϛ′ 94 = ϘΔ′ 678 = XOH′ 2002 = ͵BB′ 923090 = ͵Ϡ͵Κ͵ΓϘ′

The macro `\Grtoday` produces the current date with the month and the day as greek numerals. Today is ΙΓ′ Ὀκτωβρίου ͵ΒΚΓ′.

4.1 Alphabetical counters

In line with Greek typographical tradition (and to avoid messed up alphabetical counters with LGR fonts), *babel-greek* changes the internal LaTeX commands `\@alph` and `\@Alph` to use Greek numerals inside Greek text parts (see section 5.2 for an example).

5 Font Encoding

TeX's standard 8-bit text fonts don't provide for Greek characters. Every language switch to `greek` calls the `\extrasgreek` language hook which in turn calls `\greekscript` to ensure a Greek-supporting font encoding (LGR or TU). With the current setup, this document uses

- LGR as `\greekfontencoding`,
- T1 as `\latinencoding`, and T1 inside `\ensureascii`.

If `\greekfontencoding` is LGR, *babel-greek* performs additional setup steps to fix issues with the Latin transliteration (see below). If it is TU, *babel-greek* loads Greek LICR definitions from the file `tuenc-greek.def`[2].

Switching to a font encoding supporting the Greek script is possible without switching the Babel language using the declarations `\greekscript` (no switch if

[2]Provided by *greek-fontenc* since version 0.14 (2020-02-28)

the current encoding supports the Greek script (e.g. the Unicode font encodings TU and PU) or `\greektext` (always switch to LGR) and the corresponding functions `\ensuregreek` and `\lgrfont`.[3] These commands also work in the middle of a paragraph or word: Φίλων τοῦ TeX (ΕΦΤ) – Friends (Φίλων) of TeX.

5.1 LGR's *Latin transliteration*

LGR has Greek characters in the slots reserved for Latin characters and other symbols in a TeX *standard text font encoding*. This allows the use of a *Latin transliteration* for the input of Greek characters[4], however, characters that should be printed as Latin characters must be protected from conversion by a font encoding switch, either selecting a different language or wrapping them with `\ensureascii` (provided by the Babel core), that sets its argument using an ASCII-compatible font encoding. The legacy declaration `\latintext` switches the font encoding to `\latinencoding`.

With the Unicode font encoding TU, Latin characters can be used in Greek text parts and the Latin transliteration does not work (but see the last example below).

The following quote (with the Babel language set to Greek) illustrates the problem:

Literal characters, words in the "foreign" script protected:

> Φίλων τοῦ TeX (ΕΦΤ) – Friends (Φίλων) of TeX.

Unprotected ASCII characters come out as Greek characters with LGR:

> Φίλων τοῦ ΤεΞ (ΕΦΤ) – Φριενδς (Φίλων) οφ ΤεΞ.

The Latin transliteration works in LGR but not TU:

> Φίλων τοῦ TeX (ΕΦΤ) – Friends (Φίλων) of TeX.

The Latin transliteration can also be used with Xe/LuaTeX, if the input text is wrapped in `\lgrfont`[5] but may result in non-matching fonts and wrong hyphenation:

> Φίλων τοῦ TeX (ΕΦΤ) – Friends (Φίλων) of TeX.

5.1.1 The `keep-semicolon` attribute

The LGR font encoding uses the Latin question mark as input for the *erotimatiko* and maps the semicolon to a middle dot (*ano teleia*). As a result, Unicode-encoded texts that use the semicolon as *erotimatiko* end up with an *ano teleia* in its place:

The character 037E GREEK QUESTION MARK works with both, Xe/LuaTeX and 8-bit TeX. However it is deprecated and Unicode normalizes it to 003B SEMICOLON. This means that even texts wich use the GREEK QUESTION

[3] Hyphenation patterns are not changed, check for wrong hyphenations.
[4] see `usage.pdf`
[5] available, if the LGR encoding is loaded with the *fontenc* package

MARK may end up with SEMICOLON after drag-and-drop or other processing and with a middle dot in the final output.

With the `keep-semicolon` language attribute, 003B SEMICOLON is made active and inserts an *erotimatiko* also with LGR-encoded fonts, if the text language is set to Greek (in this document, the semicolon is active).

Input	T1	LGR	Greek language
003F QUESTION MARK	?	;	;
037E GREEK QUESTION MARK	n/d	;	;
003B SEMICOLON	;	·	;
00B7 MIDDLE DOT	·	·	·

n/d: character not defined in T1 encoding.

This attribute is ignored with Unicode fonts (where the SEMICOLON literal always prints a semicolon character).

Test in math mode: English: $ab; a\ b, (a;a;2)$, Greek: $ab; a\ b, (a;\alpha;2)$.

5.1.2 LGR-proofed macros

Babel-greek provides LGR-local variants for some *TextCommand*s that rely on a standard text encoding.[6] The fallback definitions for some *textcomp* symbols compose the symbols out of Latin letters. The fixes must not overwrite the selection of pre-composed symbols from *textcomp* or TU (try copy and paste from the PDF output).

LGR fonts have a middle dot glyph at the place of the ampersand. The new *TextCommand* `\textampersand` always prints an ampersand.

English: (T1) ©®™ A&W
English: (ΛΓΡ) ©®™ A&Ω
Greek: (ΛΓΡ) ©®™ A&Ω

5.2 LGR re-definitions

The generic macro `\&` is re-defined inside Greek text parts to use the original definition in math mode and `\textampersand` in text mode.

5.2.1 Roman numerals

Without fixes, Roman numerals are printed according to the Latin transliteration (including the conversion of "v" to a ZERO WIDTH NON-JOINER) if the font encoding is LGR:

T1: i, ii, iii, iv, v, vi
ΛΓΡ: ι, ιι, ιιι, ι, , ι

Roman numerals are used by the default document classes, e.g., in the third level of enumerations or as page numbers in the frontmatter of a book. They

[6] These workarounds cannot be done in `lgrenc.def` from the *greek-fontenc* package because they are not allowed in a "font encoding definition file" [`fntguide.pdf`].

may move to auto-generated document parts like the ToC, (hyper)references, or an index.

As document authors cannot wrap page numbers in a ToC in `\ensureascii`, *Babel-greek* redefines the internal LaTeX commands `\@roman` and `\@Roman` to make Roman numberals LGR-proof. Unfortunately, this breaks Makeindex (cf. `test-lgr-fixes.tex`).

5.2.2 Example

In Greek text parts, enumerated lists use Greek numerals in the second and fourth level and ASCII-proofed Roman numerals in the third level.

1. ιτεμ 1
   (α′) ιτεμ 1.1
      i . ιτεμ 1.1.1
         Α′. ιτεμ 1.1.1.1
         Β′. ιτεμ 1.1.1.2
         Γ′. ιτεμ 1.1.1.3
      ii . ιτεμ 1.1.2

Setting the language back to English should restore the alphabetic numbering:

1. item 1
   (a) item 1.1
      i. item 1.1.1
         A. item 1.1.1.1
         B. item 1.1.1.2
         C. item 1.1.1.3
      ii. item 1.1.2

More tests of the LGR-redefinitions are in `test-lgr-fixes.tex`.

6 Up- and downcasing in Greek

Capital Greek letters have diacritics (except the dialytika and sub-iota) to the left (instead of above) and drop them in uppercase (except the dialytika), e.g., μαΐστρος → ΜΑΪΣΤΡΟΣ.

Tonos and psili mark a *hiatus* (break-up of a diphthong) if placed on the first vowel of a diphthong. A dialytika must be placed on the second vowel if they are dropped, e.g. άι, άυ, έι, ἄι, ἄυ, ὰυ, ἔι → ΑΪ, ΑΫ́, ΕΪ, ΑΪ, ΑΫ́, ΑΫ́, ΕΪ

Some affected words:

ἄυλος → ΑΫ́ΛΟΣ ἄυλος → ΑΫ́ΛΟΣ
μάινη → ΜΑΪΝΗ

κέικ → ΚΕΪΚ
ἀυπνία → ΑΫ̓ΠΝΙΑ
ρωμέικα → ΡΩΜΕΪΚΑ
ἀυτή → ΑΫ̓ΤΗ
νεράιδα → ΝΕΡΑΪΔΑ[7]

The file `char-list.tex` in the *greek-fontenc* package includes a comprehensive test of case changing for all supported Greek characters and their various input methods.

6.1 Problems and fixes

6.1.1 Input variants

Depending on the LaTeX version and input variant, there are several limitations and problems.

With `\greekfontencoding` LGR, LaTeX version 2023-06-01, and language variant "monotoniko", we get

pre-composed: ῶ, ή, ΐ, ᾷ, άι → Ω, H, Ϊ, $A_{ι}$, AΪ

transliteration: ω, ή, ΐ, ᾷ, άι → Ω, H, Ϊ, ᾽$A_{ι}$, AI

accent macro + LICR: ῶ, ή, ΐ, ᾷ, άι → Ω, H, Ϊ, $A_{ι}$, AΪ

accent macro + transliteration: ῶ, ή, ΐ, ᾷ, άι → Ω, H, Ϊ, $A_{ι}$, AΪ

accent macro + literal: → inputenc Error: Invalid UTF-8 byte sequence

- The implementation of `\MakeUppercase` introduced in the 2022/06 LaTeX release[8] works (almost) fine with pre-composed literal characters but there are Unicode errors (unknown characters) under 8-bit TeX. Fixed with LaTeX 2023, babel-greek 1.13 and greek-fontenc 2.3.
- The new `\MakeUppercase` did not drop accents input as short accent macros or with the Latin transliteration. Fixed with the 2023 LaTeX release, babel-greek 1.14, and greek-fontenc 2.4.
- The *hiatus* feature fails with the Latin transliteration. Use accent macros, e.g., replace `k'eik` with `k\'eik`.

  It also failed with pre-composed characters and LaTeX versions older than 2022/06.[9]

6.1.2 Particuliarities of the Latin Transliteration

To enable correct upcasing of the "Latin transliteration", *babel-greek* changes the uppercase equivalent of some characters with special meaning in LGR. To minimise side-effects, this is done:

[7] This example uses literal input. It fails with pre-2022 `\MakeUppercase`.
[8] cf. LaTeX News 35
[9] Some "pro" Unicode fonts provide this feature on their own, cf. Greek typesetting without the tears

- only if `\greekfontencoding` is a "short macro" expanding to LGR, i.e. not in documents using Unicode fonts (unless `\greekfontencoding` is explicitly set to `LGR` before loading *babel-greek*),
- only for diacritics that are actually required in the selected language variant (i.e. only for the *tonos* `'`, if the language variant is the default `monotoniko`),
- not for the characters "v" (zero-width space) and "c" (final sigma). Use `\textcompwordmark` instead of `v` and autosigma (`s`) instead of `c` in text parts that could/should become upcased, e.g., αυς ↦ ΑΥΣ not αυς ↦ Α˝Υ˜.

Since version 1.13.2, *babel-greek* utilises `\DeclareUppercaseMapping` (a LaTeX kernel command, new in 2023) for the required change. For LaTeX version older than 2022/06, the `\uccode` is set to the "empty" character 0x9f = 159. *Composite command* definitions ensure that combined accents also work for accent characters "upcased" to the charcter No 159 = 0x9f: ΰ ϋ ὐ ύ ᾶ ᾰ ᾱ ᾳ → Ϋ Ϋ Ϋ Ϋ A A AI AI.

With font encoding LGR, LaTeX version 2023-06-01, and language variant "monotoniko", we get for `" | ' ‘ > <` (dialytika[10], sub-iota, tonos/oxia, varia, psili, and dasia):

᾽ ῾ ῀ → ᾽ ῾ ῀

The changed uc/lccodes have strange effects on Latin text parts in Greek paragraphs if only the encoding is switched:

English: Let's see: " | ' ‘ > < → LET'S SEE: " | ' ‘ > <
`\ensureascii`: Let's see: " | ' ‘ > < → LETS SEE: " | ‘ > <

6.1.3 Iota subscript vs. iota adscript

Unicode decomposes letters with *mute iota* (GREEK CAPITAL LETTER ... WITH [... AND] PROSGEGRAMMENI) to the base letter and a COMBINING GREEK YPOGEGRAMMENI (U+0345).[11] Accordingly, the "canonical" LICR for all pre-composed letters with mute iota is the base character LICR followed by `\ypogegrammeni`.

The appearance in the output and upcasing results depend on the chosen font and LaTeX version. Compare letters followed by `\ypogegrammeni` with pre-composed characters

ᾳᾼ ᾳ / AIAI AI / ᾽AI᾽AI ᾽AI
MakeUppercase AIAI AI / AIAI AI / AIAI AI
MakeLowercase ᾳᾳ ᾳ / ᾳᾳ ᾳ / ᾲᾲ ᾲ

and letters followed by `\prosgegrammeni` with literal character + GREEK YPOGEGRAMMENI:

[10] Unless followed by a to-be accented vowel, the quotation mark " is converted to an upper right apostrophe by LGR.
[11] They are named ... WITH PROSGEGRAMMENI for "historic reasons" (cf. Nick Nicholas *Titlecase and Adscripts*).

$\alpha_{\text{ɪ}}\alpha_{\text{ɪ}}$ $\alpha_{\text{ɪ}}$ / $\text{A}_{\text{ɪ}}\text{A}_{\text{ɪ}}$ $\text{A}_{\text{ɪ}}$ / $\acute{\text{A}}_{\text{ɪ}}\acute{\text{A}}_{\text{ɪ}}$ $\acute{\text{A}}_{\text{ɪ}}$
MakeUppercase $\text{A}_{\text{ɪ}}\text{A}_{\text{ɪ}}$ AI / $\text{A}_{\text{ɪ}}\text{A}_{\text{ɪ}}$ AI / $\text{A}_{\text{ɪ}}\text{A}_{\text{ɪ}}$ AI
MakeLowercase $\alpha\!\!\alpha_{\text{ɪ}}$ / $\alpha\!\!\alpha_{\text{ɪ}}$ / $\acute{\alpha}\!\!\acute{\alpha}_{\text{ɪ}}$

See also the Unicode FAQ.