Test the Greek support for Babel

Günter Milde

June 1, 2023

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

Contents

1 Language Switch

The declaration `\selectlanguage` switches between languages.

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

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.

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 the Latin transliteration (α . . . Ω) defined by the LGR font encoding (if LGR is used).

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 1 Ἰουνίου 2023.

Show the auto-strings for language variant “polutoniko”.

2.1 Captions

Περίληψη, βλέπε ἐπίσης, Παράρτημα, Βιβλιογραφία,
Κοινοποίηση, Κεφάλαιο, Περιεχόμενα, Συνημμένα,
Σχῆμα, Γλωσσάρι, Πρὸς, Εὑρετήριο,
Κατάλογος Σχημάτων, Κατάλογος Πινάκων,
Σελίδα, Μέρος, Πρόλογος, Ἀπόδειξη,
Ἀναφορὲς, βλέπε, Πίνακας

Test correct upcasing (dropping of accents):

ΠΕΡΙΛΗΨΗ, ΒΛΕΠΕ ΕΠΙΣΗΣ, ΠΑΡΑΡΤΗΜΑ, ΒΙΒΛΙΟΓΡΑΦΙΑ,
ΚΟΙΝΟΠΟΙΗΣΗ, ΚΕΦΑΛΑΙΟ, ΠΕΡΙΕΧΟΜΕΝΑ, ΣΥΝΗΜΜΕΝΑ,
ΣΧΗΜΑ, ΓΛΩΣΣΑΡΙ, ΠΡΟΣ, ΕΥΡΕΤΗΡΙΟ,
ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ, ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ,
ΣΕΛΙΔΑ, ΜΕΡΟΣ, ΠΡΟΛΟΓΟΣ, ΑΠΟΔΕΙΞΗ,
ΑΝΑΦΟΡΕΣ, ΒΛΕΠΕ, ΠΙΝΑΚΑΣ

[1] LaTeX internal character representation

2.2 Months

1 Ἰανουαρίου 2023 1 ΙΑΝΟΥΑΡΙΟΥ 2023
1 Φεβρουαρίου 2023 1 ΦΕΒΡΟΥΑΡΙΟΥ 2023
1 Μαρτίου 2023 1 ΜΑΡΤΙΟΥ 2023
1 Ἀπριλίου 2023 1 ΑΠΡΙΛΙΟΥ 2023
1 Μαΐου 2023 1 ΜΑΪΟΥ 2023
1 Ἰουνίου 2023 1 ΙΟΥΝΙΟΥ 2023
1 Ἰουλίου 2023 1 ΙΟΥΛΙΟΥ 2023
1 Αὐγούστου 2023 1 ΑΥΓΟΥΣΤΟΥ 2023
1 Σεπτεμβρίου 2023 1 ΣΕΠΤΕΜΒΡΙΟΥ 2023
1 Ὀκτωβρίου 2023 1 ΟΚΤΩΒΡΙΟΥ 2023
1 Νοεμβρίου 2023 1 ΝΟΕΜΒΡΙΟΥ 2023
1 Δεκεμβρίου 2023 1 ΔΕΚΕΜΒΡΙΟΥ 2023

3 Hyphenation

Patterns for the Greek language variants:

monotonic: `\l@monogreek` = 45
polytonic: `\l@polygreek` = 11
ancient: `\l@ancientgreek` = 39

current: `\l@greek` = 11

Greek paragraph:

mo-no-to-nic: Ευ-ρε-τήριο, ε-πίσης, Α-πόδει-ξη
polytonic: Ε-ὑρε-τήριο, ἐπίσης, Ἀ-πόδει-ξη
ancient: Ε-ὑρε-τήριον, ὡσα-ύτως, Ἀ-πόδει-ξις

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

mono-tonic: Ευ-ρε-τήριο, ε-πίσης, Α-πόδει-ξη
polytonic: Ε-ὑρε-τήριο, ἐπίσης, Ἀ-πόδει-ξη
ancient: Ε-ὑρε-τήριον, ὡσα-ύτως, Ἀ-πόδει-ξις

English paragraph with Greek script (`\ensuregreek`): no hypenation

mono-tonic: Ευ-ρετήριο, επίσης, Απόδειξη
polytonic: Εὑρετήριο, ἐπίσης, Ἀπόδειξη
ancient: Εὑρετήριον, ὡσαύτως, Ἀπόδειξις

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′

$36 = \Lambda\text{Ͳ}'$ $94 = \text{ϛ}\Delta'$ $678 = \text{XOH}'$ $2002 = \text{͵BB}'$ $923090 = \text{͵Ϡ͵K͵Γϛ}'$

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):

$36 = \lambda\text{ϛ}'$ $94 = \text{ϙ}\delta'$ $678 = \chi o\eta'$ $2002 = \text{͵}\beta\beta'$ $923090 = \text{͵ϡ͵}\chi\text{͵}\gamma\text{ϙ}'$

$36 = \Lambda\text{ϛ}'$ $94 = \text{Ϙ}\Delta'$ $678 = \text{XOH}'$ $2002 = \text{͵BB}'$ $923090 = \text{͵Ϡ͵K͵Γ}\text{Ϙ}'$

The macro `\Grtoday` produces the current date with the month and the day as greek numerals. Today is Α΄ Ἰουνίου ͵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 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.

[2]Provided by *greek-fontenc* since version 0.14 (2020-02-28)
[3]Hyphenation patterns are not changed, check for wrong hyphenations.

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 be used with also 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 MARK may and 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:

[4] see `usage.pdf`

[5] available, if the LGR encoding is loaded with the *fontenc* package

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

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

Test in math mode: English: ab; ab; $a\,b$, (a;a;2), Greek: αβ; ab; $a\,b$, (a;α;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, . . . , mcmlxxv
ΛΓΡ: ι, ιι, ιιι, ιv, . . . , μcμλξξ

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 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`).

[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`].

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 test 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. For literal Greek characters and accent macros, these rules are implemented and tested in the required package *greek-fontenc*.

Tonos and dasia 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, ἀυ, ἔι ↦ ΑΪ, ΑΫ́, ΕΪ, ΑΪ, ΑΫ́, ΑΫ́, ΕΪ.

Some affected words:

ἄυλος ↦ ΑΫ́ΛΟΣ, ἄυλος ↦ ΑΫ́ΛΟΣ,
μάινα ↦ ΜΑΪΝΑ,
κέικ ↦ ΚΕΪΚ,
ἀυπνία ↦ ΑΫ́ΠΝΙΑ,
ρωμέικα ↦ ΡΩΜΕΪΚΑ.

With the pre-2022/06 `\MakeUppercase`, automatic upcasing of words with *hiatus* works correctly only if the accents are input as macro and the letters as macro or via the Latin transliteration.

With the current (2023/02/10) `\MakeUppercase`, it works for literal Unicode input and named accents. The `\uccode` changes are ignored.

6.1 Iota subscript vs. iota adscript

Pre-composed capital letters with *mute iota* decompose to the base letter and COMBINING GREEK YPOGEGRAMMENI (U+0345) even if they are named ... WITH [... AND] PROSGEGRAMMENI for "historic reasons".[7] Accordingly, the "canonical" LICR for capital letters with mute iota is the base character LICR followed by `\ypogegrammeni`.

Compare letters followed by `\prosgegrammeni` and `\ypogegrammeni` with the pre-composed characters and with character + literal GREEK YPOGEGRAMMENI.

prosgegrammeni:

	ᾳᾼ ᾳ / AIAI AIAI / ᾼ ᾼ ᾼ
MakeUppercase	AIAI AI / AIAI AIAI / AIAI AIAI
MakeLowercase	ᾳᾳ ᾳ / ᾳᾳ ᾳᾳ / ᾳᾳ ᾳᾳ

ypogegrammeni:

	ᾳᾳ ᾳ / AIAI AI / ᾼ ᾼ
MakeUppercase	AIAI AI / AIAI AI / AIAI AI
MakeLowercase	ᾳᾳ ᾳ / ᾳᾳ ᾳ / ᾳᾳ ᾳ

6.2 Changed uccode/lccode values for LGR

LGR encoded fonts define ligatures for characters with diacritics. To enable correct upcasing, *babel-greek* changes the uc/lccodes of some characters. Characters used to input diacritics are mapped to the "empty" character 0x9f = 159.

To minimise side-effects (see below), uc/lccodes are only changed:

- if `\greekfontencoding` is a "short macro" expanding to LGR,[8]
- if the LaTeX version is older than 2022/06, as `\uccode`s are ignored by the `\MakeUppercase` implementation introduced in this version,
- for accents that are actually required in the selcected language variant (i.e. only for the *tonos*, if the language variant is the default "monotonic".

With `\greekfontencoding` LGR, LaTeX version 2020-10-01, and language variant "polutoniko" we get for dialytika[9], sub-iota, tonos/oxia, varia, psili, and dasia:

᾽ ᾽ ῾ ῾ ῾ ῾ ↦ ᾽ ᾽

Composite command definitions ensure that combined accents also work for accent characters "upcased" to the charcter No 159 = 0x9f):

[7] cf. Nick Nicholas 'Titlecase and Adscripts'

[8] i.e. not in documents using Unicode fonts unless `\greekfontencoding` is explicitly set to `LGR` before loading *babel-greek*

[9] Unless followed by a to-be accented vowel, the quotation mark `"` is converted to an upper right apostrophe by LGR.

ύ ὐ ὺ ύ ᾶ ᾰ ᾱ ᾱ ↦ ΰ ΰ ΰ ΰ A A AÏ AÏ

With babel-greek versions up to 1.11, using the tilde for the *perispomeni* accent inside `\MakeUppercase` led to a

```
Package inputenc Error:  Invalid UTF-8 byte "9F
```

(even without loading *inputenc* and with `\UseRawInputEncoding`).[10] Version 1.12 fixes the issue so that all three input variants work again (with LaTeX versions up to 2022/06)[11]

ῦ ῦ ῦ ↦ ϒ ϒ ϒ

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: " | ' ‘ > < ↦ LET§S SEE: " | § § § §

To limit the effect on, e.g., section headings (which are printed capitalised with the "headings" style), uccodes for "v" (zero-width space) and "c" (final sigma) are not changed. Use `\textcompwordmark` instead of `v` and autosigma (`s`) instead of `c` in text parts that could/should become upcased, e.g., αυς ↦ AϒΣ not αυς ↦ A˝ϒ˜.

[10] cf. LaTeX News 28

[11] The implementation of `\MakeUppercase` introduced in the 2022/06 LaTeX release, cf. LaTeX News 35, works (almost) fine with literal input but currently fails with the Latin transliteration and standard accent macros.