TeXaccents
version 1.0.1

Guido Milanese[*]

17th September 2022

Abstract

TeXaccents is a standalone utility designed to convert legacy (La)TeX ligatures and codes for “accented” characters to Unicode equivalents (text mode, no math) . For example, `\={a}` (‘a’ with macron) will be converted to ā.

General information

Even if modern compilers handle Unicode encoding, (La) and files featuring “legacy” encoding for non-Ascii characters are still very common, and users may need to incorporate old code into new texts that make use of modern text encoding.

Several utilities are available online that claim to be able to convert legacy (La) encoding to standard Unicode. See:

- *Simple LaTeX to Text Converter.* A complex programme, able to deal with maths. Insofar as non-Ascii chars are concerned, it fails sometimes, at least according to my tests. See https://pylatexenc.readthedocs.io/en/latest/latexwalker/. Written in Python.
- *LaTeX handler.* Converts non-Ascii (La) encoding to Unicode. However, it does not seem to be able to deal with the legacy encoding, e.g. `{\a}`

---

[*] Università Cattolica d.S.C., Dipartimento di scienze storiche e filologiche, via Trieste 17, I-25121 Brescia

instead of `\{a}` or `\a`. It does not convert simple ligatures as `\ae{}` `\oe{}`. I used the tables provided by this programme as a starting point. Written in Python. See https://github.com/hayk314/LaTex-handler.

- *Pandoc* is the standard programme for any text format conversion (https://pandoc.org/). It converts almost all the accents (thorn and eth missing?), but (if I have checked this correctly) normalises files stripping non-standard fields. This can be a problem for scholars who frequently use non-standard fields, such as e.g. "shorttitle", required by not a few bibliographic styles.

TeXaccents should be able to transform (La) normal text or "accents" (not "math" accents) to their Unicode equivalent. The programme deals with the following codes (*not all the fonts are able to output all the required Unicode glyphs of this table!*):

```
| NAME              | \tex        | Unicode |
|---------------    |-------     |---------|
| Umlaut            | \"{a}      | ä        |
| acute             | \'{a}      | á        |
| double acute      | \H{a}      | a̋        |
| grave             | \`{a}      | à        |
| circumflex        | \^{a}      | â        |
| caron hraceck     | \v{a}      | ⊠        |
| breve             | \u{a}      | ă        |
| cedilla           | \c{c}      | ç        |
| dot               | \.{a}      | ⊠        |
| dot under         | \d{a}      | a        |
| ogonek            | \k{a}      | ą        |
| tilde             | \~{a}      | ã        |
| macron            | \={a}      | ā        |
| bar under         | \b{a}      | a̱        |
| ring over         | \r{a}      | å        |
```

The programme should recognize the following varieties:

`\'a` – `\'{a}` – `{\'a}` – `{{\'a}}`

It transforms also the encoding for : æ œ Æ Œ ð Đ þ Þ ø Ø ł Ł. Checking the page https://www.utf8-chartable.de/unicode-utf8-table.pl?number=512 I could not find a legacy text mode encoding for: **ƀ Ƀ đ Đ ǥ Ǥ ħ Ħ ɨ Ɨ ŧ Ŧ ƶ Ƶ** (some chars are accessible in math mode).

Setup

From source

The programme is written in Snobol (https://en.wikipedia.org/wiki/SNOBOL or https://it.wikipedia.org/wiki/SNOBOL) and should run on any platform. Steps:

1. Install Snobol4 (version 2.3, March 2022) from http://www.regressive.org/snobol4/csnobol4/curr/. Make sure to install the compiler in a folder listed in your PATH or add the folder to your path. On Linux the folder `snobol4` is installed under `/usr/local/bin/`, which is normally listed in the PATH of a standard Linux system.
2. Test the compiler running `snobol4` from the command line. Leave the compiler with `Ctr-C` or writing `end`.
3. Copy `texaccents.sno` and all the provided `*.inc` files

   `compiler.inc` `delete.inc` `grepl.inc` `newline.inc` `systype.inc`

to a folder of your choice (e.g. `/home/<user>/bin`).

4. In this folder, run `snobol4 texaccents.sno testaccents-in testaccents-out` to test the programme. The test file contains all the accents listed above. See the result typing `cat testaccents-out` (Unixes / Powershell) or `type testaccents-out` (Windows/Dos prompt), or open the file with your text editor. The output file name is just a suggestion, of course.

Windows standalone version

If preferred, a Windows EXE standalone file is provided. It was compiled using Spitbol (see https://github.com/spitbol/windows-nt); the source code has been slightly adapted to Spitbol (basically only input/output syntax). From any directory, run `texaccents.exe INPUT OUTPUT`. To test the programme, run `texaccents.exe testaccents-in testaccents-out`. As above, the output file name is just a suggestion.

History

- 25th July 2022. First version (after trying unsuccesfully to convert an old file with existing utilities)
- 17th August 2022. First complete version (0.9).
- 27rd August 2022. This version (1.0) with documentation and comments.
- 17th September 2022. Windows standalone executable. Manual page written. Version message added; help message improved. In the source, a regular shebang according to the recommendation of CTAN (https://tug.org/texlive/pkgcontrib.html) was added. Documentation updated accordingly.

Contacts / todo

Bugs / suggestions / improvements: please write to guido.milanese@unicatt.it using *TeXaccents* as subject of the mail.

Genoa, Italy, 17th September 2022