NAME

tlgu – convert beta code TLG and PHI CD-ROM txt files to Unicode

SYNOPSIS

tlgu [*options*] [*input_file*] [*output_file*]

DESCRIPTION

tlgu will convert an *input_file* from Thesaurus Linguae Graeca (TLG) and Packard Humanities Institute (PHI) representation to a Unicode (UTF-8) *output_file*. If *input_file* is not specified, standard input will be read. If *output_file* is not specified, the Unicode text is directed to standard output. The TLG/PHI representation consists of **beta-code** text and **citation** information.

OPTIONS

−b inserts a form feed and citation information (levels a, b, c, d) on every "book" citation change. By default the program will output line feeds only (see also **−p**).

−p observes paging instructions. By default the program will output line feeds only.

−r primarily Roman text (PHI). Some TLG texts, notably doccan1.txt and doccan2.txt are mostly roman texts lacking explicit language change codes. Setting this option will force a change to roman text after each citation block is encountered.

−v highest-level reference citation is included before each text line (v-level)

−w reference citation is included before each text line (w-level)

−x reference citation is included before each text line (x-level)

−y reference citation is included before each text line (y-level)

−z lowest-level reference citation is included before each text line (z-level).

−Z <custom_citation_format_string>
an arbitrary combination of citation information is included before each text line; see also -e option e.g. "%A/%B/%x/%y/%z\t" will output the contents of the A, B **citation description** levels, followed by x, y, z **citation reference** levels, followed by a TAB character.

−e <custom_blank_citation_string>
if there is no citation information for a citation level defined with the -Z option above, a single right-hand slash is substituted by default; you may define any string with this option e.g. "-" or "[NONE]" are valid inputs

−B inserts blank space (a tab) before each and every line.

−X compact format; v, w, x citations are inserted as they change at the beginning of each section.

−Y compact format; w, x, y citations are inserted as they change at the beginning of each section.

−N no spaces; line ends and hyphens before an ID code are removed while hyphens and spaces before page and column ends are (still) retained.

−C citation debug information is output.

−S special code debug information is output.

−V block processing information is output (verbose).

−U vowels with acute accent are output using the Unicode 0x0370 codes rather than the 0x1F00 ones for compatibility with most current (as of 2020) keyboard encoders.

−W each work (book) is output as a separate file in the form output_file-xxx.txt; if an output file is not specified, this option has no effect.

HISTORY AND INTENDED USE

The purpose of **tlgu** is to translate binary TLG/PHI-format files into readable and editable text. It is based on an earlier program written in 80x86 assembly language (1996) outputting codes for a home-made font which used the prevalent hellenic font encodings of that time complemented by dead accent characters - not very attractive, but readable.

Then came Unicode and a plethora of accented character glyphs; Polytonic fonts are already available (Cardo, Gentium, Athena, Athenian, Porson); new fonts are being created and older fonts are being expanded as special-use code points are included in the Unicode definition (musical symbols, other special symbols). A notable effort since this note was originally drafted is that of the Greek Font Society, now featuring a great, and expanding, selection of open polytonic fonts.

So, at this point in time, **tlgu** will crunch a file which has been formatted according to the published TLG/PHI format and produce codes for most glyphs generally available. No attempt has been made to introduce multi-character sequences or formatting codes (font changes). If a code has not been defined, the program will output the respective "code family" glyph. You may use the **–S** option to check such codes against the published beta code definition.

July 2005 - Troy A. Griffitts (scribe, crosswire org) contributed the arbitrary citation output code and added per-line processing of the input file.

April 2006 - Final sigma will now be output at end-of-line (!) from free-form input text (thank you Jan).

October 2011 - stdout is used if output_file is not specified.

November 2011 - citations (v, w, x) at the start of section changes (e-book option)

May 2012 - Nick White (nick white, durham ac uk) revised the input arguments to use tlgu as a filter; stdin is used if input_file is not specified

May 2020 - Alternate output codes for vowels with acute accent (-U option)

July 2021 - Corrections to citation code

EXAMPLES

./tlgu -r DOCCAN2.TXT doccanu.txt Translate the TLG canon to a unicode text file. Note the use of the **-r** option (this file expects Roman as the default font).

./tlgu -x -y -z TLG1799.TXT tlg1799u.txt
Generate a continuous file with the texts of granpa Euclides. Available citations (-x -y -z) are Book//demonstratio/line as shown in the respective "cit" field of doccan2.txt.

./tlgu -b -B TLG1799.TXT tlg1799u.txt
Generate the same texts, this time with a page feed and book citation information on the first page of each book and a tab before each line (use with OOo versions earlier than 1.1.4).

./tlgu -C TLG1799.TXT tlg1799u.txt
See how the citation information changes within each TLG block.

./tlgu -S TLG1799.TXT tlg1799u.txt | sort > symbols1799.txt
Check out the symbols used in a work. Book and x, y, z references are printed on a separate line for each symbol. Sort / grep the output to locate specific symbols of interest; save in a file for later use.

./tlgu -W TLG0006.TXT tlg0006u
Will produce separate files for each work, named tlg006u-001.txt etc.

./tlgu -Z "%A/%B/%D/%c/%d/%Z/%x/%y/%z\t" -e "-" chr0010.txt chr0010u.txt
Will generate a file with citation description (A, B, D, Z) and citation reference (c, d, x, y, z) levels, separated by "/" followed by a TAB character and the respective text. Blank citation elements will be filled with a single "-" e.g. Asia/Smyrna/1222-1223 ac/IGChAs/Asia Min [Chr]/88/-/2A/7p1 [TAB] inscription text etc.

./tlgu -r -N -X LAT0448.TXT LAT0448.xx.TXT
will produce a compact version of the Gaius Iulius Caesar texts with v and x citations printed as they change; similarly, **./tlgu -r -N -Y LAT2150.TXT LAT2150.yy.TXT** will produce a compact version of Zeno's texts.

POST-PROCESSING EXAMPLES

I use the OpenOffice/LibreOffice suite for most of my work. This example shows one of many possible ways of using the search and replace facility to create a readable version of the Suda lexicon.

./tlgu -B TLG4085.TXT tlg4085u.txt
A Unicode file with the text is created

Open the generated file with Openoffice/LibreOffice:
File | Open | Filename: tlg4085u.txt, File Type: Text Encoded −− Press Open

The ASCII Filter Options window appears. Select the Unicode (UTF-8) character set and a proper Unicode font installed in your machine (e.g. Cardo). Press OK.

Replace angle brackets with expanded text
Lexicon terms are enclosed in <angle brackets>. The actual beta codes indicate the use of expanded text for emphasis. Select Edit | Find & Replace. The **Find & Replace** window appears.

In the **Search For** field, type the following expression: **<[^<>]*>** This means "find any characters between angle brackets, not including angle brackets".

In the **Replace With** window insert a single ampersand: **&** This means that we need to **add** formatting information (this case) or additional text to the text found. Press **More Options**, **Format...** and select the **Position** tab; select Spacing Expanded by 2.0 points. Press OK.

Check the **Regular Expressions** box and press **Replace All**.

You may now replace the angle brackets with nothings.

Repeat the above procedure for titles enclosed in {braces}. Write a macro...

Other useful information
If you are using your wordprocessor with a locale setting other than Hellenic (el_GR), the following invocation with the desired character classification may prove useful for the occasional polytonic editing:

LC_CTYPE=el_GR.UTF-8 /usr/bin/soffice (or **/opt/libreoffice3.4/program/soffice**).

I put my default locale and keyboard definitions in my **.bashrc** or **.profile**:

export LC_CTYPE=el_GR.UTF-8
setxkbmap us,el ,polytonic -option grp:ctrl_shift_toggle -option grp_led:scroll

This way multi-lingual text can be entered; keyboard layout switching is done by pressing Ctrl/Shift; alternate keyboard layout is indicated by the Scroll Lock light on the keyboard.

FURTHER DEVELOPMENT

You may not like the character output for a specific code. Check out the **tlgcodes.h** file containing the special symbol and punctuation codes and select one to suit you better. It will probably be a while before the beta to Unicode correspondence settles down.

Drop me a line, if you need a new feature; let me know if you do find an interesting applications that others can profit from.

REFERENCES

There are several texts describing the internal representation of **PHI** and **TLG** text, ID data, citation data and index files. The originator of this format is the Packard Humanities Institute. The TLG is maintained by UCI – see **www.tlg.uci.edu** – where you may find the latest versions of the **TLG Beta Code Manual** and the **TLG Beta Code Quick Reference Guide**.

Unicode consortium (**www.unicode.org**) publications pertaining to the codification of characters used in Hellenic literature, scientific and musical texts.

The OpenOffice/Libreoffice suite in its various editions (**www.openoffice.org** - apache.org, **www.libreoffice.org**, **www.neooffice.org**) includes a word processor that you can use to load, process and create new polytonic texts.

Greek Font Society: **www.greekfontsociety.gr**

COPYRIGHT

Copyright (C) 2004, 2005, 2011, 2013, 2020, 2021 Dimitri Marinakis (dm, ssa gr).

This file is part of tlgu which is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License (version 2) as published by the Free Software Foundation.

tlgu is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA