NAME

unibetaprep − Pre-process Beta Code files for **beta2uni**(1)

SYNOPSIS

unibetaprep [−i *input_file.pre*] [−o *output_file.beta*]

DESCRIPTION

unibetaprep(1) reads a document encoded using Beta Code that may contain special character codes from the full Beta Code of the Thesaurus Linguae Graecae (TLG) specification, and converts it to a Beta Code file that has those special characters converted to Unicode escape sequences. This departs from the traditional encoding of those special characters in favor of Unicode code point assignments.

Beta Code is an ASCII-only encoding scheme most commonly used for digital representation of polytonic Greek.

Beta Code has become a widely-adopted standard for encoding classical Greek. It was developed by David Packard in the 1970s and adopted by the Thesaurus Linguae Graecae (TLG) Project at the University of California, Irvine shortly thereafter. This encoding was later adopted by the Perseus Project in the 1980s (originally at Harvard University, now at Tufts University) and by many other collections of classical and Koine Greek. Today, the TLG corpus alone contains over 100 million words from classical to Byzantine Greek.

The TLG uses uppercase Latin letters; the Perseus Project uses lowercase. **unibetaprep**(1) will accept either.

Many classicists who use Beta Code have been actively involved in The Unicode Standard, with evolving recommendations for mapping between Beta Code and Unicode. **unibetaprep**(1) provides a capability for GNU/Linux users who wish to convert Beta Code texts to Unicode.

The most notable range of special characters in the TLG specification is the complete range of Byzantine Musical Symbols, in the Unicode range U+1D000 through U+1D0FF, inclusive. This range corresponds to the TLG special character encodings "#2000" through "#2245", respectively. If a character sequence in the TLG Beta Code specification corresponds to a Unicode glyph or glyph combination, **unibetaprep** should handle the translation correctly.

Most of these Beta Code sequences consist of a "#", "%", "<", ">", "[", or "]" character followed by one or more decimal digits. Sequences corresponding to idiosyncratic Beta Code glyphs are not translated to Unicode. The Beta Code quotation mark sequences "1, "2, "4, and "5 are converted to represent Unicode code points U+201E, U+201C, U+201A, and U+201B, respectively. For other special code sequences, consult the *TLG Beta Code Quick Reference Guide,* or examine the flex program source in file unibetaprep.l.

The output of **unibetaprep** is designed to provide the input to **beta2uni**(1), which then produces UTF-8 Unicode output.

Note: Thesaurus Linguae Graecae and TLG are registered trademarks of the University of California.

OPTIONS

−i Specify the input file. The default is STDIN.

−o Specify the output file. The default is STDOUT.

Sample usage:

unibetaprep −i *my_input_file.pre* −o *my_output_file.beta*

The output file, *my_output_file.beta,* can then be used as input for **beta2uni**(1) for conversion into a UTF-8 Unicode document.

FILES

ASCII text files using Beta Code to encode polytonic Greek.

SEE ALSO

beta2uni(1), **uni2beta**(1), **libunibetacode**(3), **unibetacode**(5)

AUTHOR

unibetaprep was written by Paul Hardy.

LICENSE

BUGS

No known bugs exist.