

**NAME**

patgen – generate patterns for TeX hyphenation

**SYNOPSIS**

**patgen** *dictionary\_file pattern\_file patout\_file translate\_file*

**DESCRIPTION**

This manual page is not meant to be exhaustive. See also the Info file or manual *Web2C: A TeX implementation* available as part of the TeX Live distribution or at <http://tug.org/web2c>.

The *patgen* program reads the *dictionary\_file* containing a list of hyphenated words and the *pattern\_file* containing previously-generated patterns (if any) for a particular language (not a complete TeX source file; see below), and produces the *patout\_file* with (previously- plus newly-generated) hyphenation patterns for that language. The *translate\_file* defines language specific values for the parameters *left\_hyphen\_min* and *right\_hyphen\_min* used by T<sub>E</sub>X's hyphenation algorithm and the external representation of the lower and upper case version(s) of all `letters' of that language. Further details of the pattern generation process such as hyphenation levels and pattern lengths are requested interactively from the user's terminal. Optionally *patgen* creates a new dictionary file **patmp.n** showing the good and bad hyphens found by the generated patterns, where *n* is the highest hyphenation level.

The patterns generated by *patgen* can be read by **initex** for use in hyphenating words. For a real-life example of *patgen*'s output, see *\$TEXMFMAIN/tex/generic/hyphen/hyphen.tex*, which contains the patterns T<sub>E</sub>X uses for English by default. At some sites, patterns for (many) other languages may be available, and the local **tex** programs may have them preloaded.

All filenames must be complete; no adding of default extensions or path searching is done.

**FILE FORMATS****Letters**

When **initex** digests hyphenation patterns, T<sub>E</sub>X first expands macros and the result must entirely consist of digits (hyphenation levels), dots (`.', edge of a word), and letters. In pattern files for non-English languages letters are often represented by macros or other expandable constructs. For the purpose of *patgen* these are just character sequences, subject to the condition that no such sequence is a prefix of another one.

**Dictionary file**

A dictionary file contains a weighted list of hyphenated words, one word per line starting in column 1. A digit in column 1 indicates a global word weight (initially =1) applicable to all following words up to the next global word weight. A digit at some intercharacter position indicates a weight for that position only.

The hyphens in a word are indicated by `-', `\*', or `.` (or their replacements as defined in the translate file) for hyphens yet to be found, `good' hyphens (correctly found by the patterns), and `bad' hyphens (erroneously found by the patterns) respectively; when reading a dictionary file `\*' is treated like `-' and `.` is ignored.

**Pattern file**

A pattern file contains only patterns in the format above, e.g., from a previous run of *patgen*. It may *not* contain any T<sub>E</sub>X comments or control sequences. For instance, this is not a valid pattern file:

```
% this is a pattern file read by TeX.
```

```
\patterns{%
...
}
```

It can only contain the actual patterns, i.e., the ‘...’.

### Translate file

A translate file starts with a line containing the values of *left\_hyphen\_min* in columns 1-2, *right\_hyphen\_min* in columns 3-4, and either a blank or the replacement for one of the "hyphen" characters ``-'`, ``*'`, and ``.'` in columns 5, 6, and 7. (Input lines are padded with blanks as for many T<sub>E</sub>X related programs.)

Each following line defines one ``letter'`: an arbitrary delimiter character in column 1, followed by one or more external representations of that character (first the ``lower'` case one used for output), each one terminated by the delimiter and the whole sequence terminated by another delimiter.

If the translate file is empty, the values *left\_hyphen\_min*=2, *right\_hyphen\_min*=3, and the 26 lower case letters **a** . . **z** with their upper case representations **A** . . **Z** are assumed.

### Terminal input

After reading the *translate\_file* and any previously-generated patterns from *pattern\_file*, *patgen* requests input from the user's terminal.

First the integer values of *hyph\_start* and *hyph\_finish*, the lowest and highest hyphenation level for which patterns are to be generated. The value of *hyph\_start* should be larger than any hyphenation level already present in *pattern\_file*.

Then, for each hyphenation level, the integer values of *pat\_start* and *pat\_finish*, the smallest and largest pattern length to be analyzed, as well as *good weight*, *bad weight*, and *threshold*, the weights for good and bad hyphens and a weight threshold for useful patterns.

Finally the decision (``y'` or ``Y'` vs. anything else) whether or not to produce a hyphenated word list.

## FILES

*\$TEXMFMAIN/tex/generic/hyphen/hyphen.tex*

The original hyphenation patterns for English, by Donald Knuth and Frank Liang.

<http://www.ctan.org/pkg/ushyph>

Additional hyphenation patterns for English, extended by Gerard Kuiken.

<http://www.ctan.org/pkg/hyph-utf8>

Collected hyphenation patterns for many languages in many formats.

<http://www.ctan.org/tex-archive/language/>

General CTAN directory for patterns and support for many other languages.

## SEE ALSO

Frank Liang and Peter Breitenlohner, *patgen.web*.

Frank Liang, *Word hy-phen-a-tion by com-puter*, STAN-CS-83-977, Stanford University Ph.D. thesis, 1983, <http://tug.org/docs/liang>.

Donald E. Knuth, *The T<sub>E</sub>X for troffbook*, Addison-Wesley, 1986, ISBN 0-201-13447-0, Appendix H.

## **AUTHORS**

Frank Liang wrote the first version of this program. Peter Breitenlohner made a substantial revision in 1991 for T<sub>E</sub>X 3. The first version was published as the appendix to the *T<sub>E</sub>X for troffware* technical report. Howard Trickey originally ported it to Unix.