

LuaT_EX

Reference

beta 0.33.0



LuaT_EX

Reference

Manual

copyright: LuaT_EX development team
more info: www.luatex.org
version: January 22, 2009

Contents

| | | |
|-------|--|----|
| 1 | Introduction | 5 |
| 2 | Basic T _E X enhancements | 7 |
| 2.1 | Version information | 7 |
| 2.2 | UNICODE text support | 7 |
| 2.3 | Wide math characters | 8 |
| 2.4 | Extended tables | 9 |
| 2.5 | Attribute registers | 9 |
| 2.5.1 | Box attributes | 10 |
| 2.6 | LUA related primitives | 10 |
| 2.6.1 | <code>\directlua</code> | 11 |
| 2.6.2 | <code>\lualua</code> | 12 |
| 2.6.3 | <code>\luaescapestring</code> | 12 |
| 2.6.4 | <code>\closeslua</code> | 12 |
| 2.7 | New E-T _E X primitives | 13 |
| 2.7.1 | <code>\clearmarks</code> | 13 |
| 2.7.2 | <code>\noligs</code> and <code>\nokerns</code> | 13 |
| 2.7.3 | <code>\formatname</code> | 13 |
| 2.7.4 | <code>\scantextokens</code> | 13 |
| 2.7.5 | Catcode tables | 13 |
| 2.7.6 | <code>\suppressfontnotfounderror</code> | 14 |
| 2.7.7 | Font syntax | 15 |
| 2.8 | Debugging | 15 |
| 3 | LUA general | 17 |
| 3.1 | Initialization | 17 |
| 3.1.1 | LUAT _E X as a LUA interpreter | 17 |
| 3.1.2 | LUAT _E X as a LUA byte compiler | 17 |
| 3.1.3 | Other commandline processing | 17 |
| 3.2 | LUA changes | 19 |
| 3.3 | LUA modules | 21 |
| 4 | LUAT _E X LUA Libraries | 23 |
| 4.1 | The <code>tex</code> library | 23 |
| 4.1.1 | Integer parameters | 23 |
| 4.1.2 | Dimension parameters | 25 |
| 4.1.3 | Direction parameters | 25 |
| 4.1.4 | Glue parameters | 25 |
| 4.1.5 | Muglue parameters | 26 |
| 4.1.6 | Tokenlist parameters | 26 |
| 4.1.7 | Convert commands | 26 |
| 4.1.8 | Last item commands | 27 |



| | | |
|--------|---|----|
| 4.1.9 | Attribute, count, dimension, skip and token registers | 27 |
| 4.1.10 | Box registers | 28 |
| 4.1.11 | Special list heads | 29 |
| 4.1.12 | Print functions | 29 |
| 4.1.13 | Helper functions | 31 |
| 4.2 | The <code>token</code> library | 31 |
| 4.2.1 | <code>token.get_next</code> | 32 |
| 4.2.2 | <code>token.is_expandable</code> | 32 |
| 4.2.3 | <code>token.expand</code> | 32 |
| 4.2.4 | <code>token.is_activechar</code> | 32 |
| 4.2.5 | <code>token.create</code> | 32 |
| 4.2.6 | <code>token.command_name</code> | 33 |
| 4.2.7 | <code>token.command_id</code> | 33 |
| 4.2.8 | <code>token.csname_name</code> | 33 |
| 4.2.9 | <code>token.csname_id</code> | 33 |
| 4.3 | The <code>node</code> library | 33 |
| 4.3.1 | Node handling functions | 34 |
| 4.3.2 | Attribute handling | 40 |
| 4.4 | The <code>texio</code> library | 40 |
| 4.4.1 | Printing functions | 41 |
| 4.5 | The <code>pdf</code> library | 41 |
| 4.6 | The <code>img</code> library | 42 |
| 4.7 | The <code>mplib</code> library | 46 |
| 4.7.1 | <code>mplib.new</code> | 46 |
| 4.7.2 | <code>mp:statistics</code> | 47 |
| 4.7.3 | <code>mp:execute</code> | 47 |
| 4.7.4 | <code>mp:finish</code> | 48 |
| 4.7.5 | Result table | 48 |
| 4.7.6 | Subsidiary table formats | 50 |
| 4.7.7 | Character size information | 51 |
| 4.8 | The <code>callback</code> library | 52 |
| 4.8.1 | File discovery callbacks | 53 |
| 4.8.2 | File reading callbacks | 55 |
| 4.8.3 | Data processing callbacks | 57 |
| 4.8.4 | Node list processing callbacks | 58 |
| 4.8.5 | Information reporting callbacks | 61 |
| 4.8.6 | Font-related callbacks | 63 |
| 4.9 | The <code>lua</code> library | 63 |
| 4.9.1 | Variables | 63 |
| 4.9.2 | LUA bytecode registers | 63 |
| 4.10 | The <code>kpse</code> library | 64 |
| 4.10.1 | <code>kpse.set_program_name</code> | 64 |
| 4.10.2 | <code>kpse.find_file</code> | 64 |
| 4.10.3 | <code>kpse.init_prog</code> | 66 |



| | | |
|--------|--|-----|
| 4.10.4 | <code>kpse.readable_file</code> | 66 |
| 4.10.5 | <code>kpse.expand_path</code> | 66 |
| 4.10.6 | <code>kpse.expand_var</code> | 66 |
| 4.10.7 | <code>kpse.expand_braces</code> | 66 |
| 4.10.8 | <code>kpse.show_path</code> | 66 |
| 4.10.9 | <code>kpse.var_value</code> | 67 |
| 4.11 | The <code>status</code> library | 67 |
| 4.12 | The <code>texconfig</code> table | 68 |
| 4.13 | The <code>font</code> library | 69 |
| 4.13.1 | Loading a TFM file | 69 |
| 4.13.2 | Loading a VF file | 70 |
| 4.13.3 | The fonts array | 70 |
| 4.13.4 | Checking a font's status | 70 |
| 4.13.5 | Defining a font directly | 70 |
| 4.13.6 | Projected next font id | 71 |
| 4.13.7 | Currently active font | 71 |
| 4.13.8 | Maximum font id | 71 |
| 4.13.9 | Iterating over all fonts | 71 |
| 4.14 | The <code>fontforge</code> library | 71 |
| 4.14.1 | Getting quick information on a font | 71 |
| 4.14.2 | Loading an OPENTYPE or TRUETYPE file | 72 |
| 4.14.3 | Applying a 'feature file' | 73 |
| 4.14.4 | Applying an 'AFM file' | 73 |
| 4.15 | Fontforge font tables | 73 |
| 4.15.1 | Table types | 73 |
| 4.16 | The <code>lang</code> library | 89 |
| 5 | Languages and characters, fonts and glyphs | 93 |
| 5.1 | Characters and glyphs | 93 |
| 5.2 | The main control loop | 94 |
| 5.3 | Loading patterns and exceptions | 95 |
| 5.4 | Applying hyphenation | 96 |
| 5.5 | Applying ligatures and kerning | 97 |
| 5.6 | Breaking paragraphs into lines | 98 |
| 6 | Font structure | 101 |
| 6.1 | Real fonts | 106 |
| 6.2 | Virtual fonts | 107 |
| 6.2.1 | Artificial fonts | 109 |
| 6.2.2 | Example virtual font | 109 |
| 7 | Nodes | 111 |
| 7.1 | LUA node representation | 111 |
| 7.1.1 | Auxiliary items | 111 |
| 7.1.2 | Main text nodes | 112 |



| | | |
|-------|--|-----|
| 7.1.3 | Math nodes | 116 |
| 7.1.4 | whatsit nodes | 119 |
| 8 | Modifications | 127 |
| 8.1 | Changes from T _E X 3.141592 | 127 |
| 8.2 | Changes from E-T _E X 2.2 | 127 |
| 8.3 | Changes from PDFT _E X 1.40 | 127 |
| 8.4 | Changes from ALEPH RC4 | 128 |
| 8.5 | Changes from standard WEB2C | 129 |
| 9 | Implementation notes | 131 |
| 9.1 | Primitives overlap | 131 |
| 9.2 | Memory allocation | 131 |
| 9.3 | Sparse arrays | 131 |
| 9.4 | Simple single-character csnames | 132 |
| 9.5 | Compressed format | 132 |
| 9.6 | Binary file reading | 132 |
| 10 | Known bugs and limitations | 133 |
| 11 | TODO | 135 |



1 Introduction

This book will eventually become the reference manual of LuaT_EX. At the moment, it simply reports the behaviour of the executable matching the snapshot or beta release date in the title page.

Features may come and go. The current version of LuaT_EX is not meant for production and users cannot depend on stability, nor on functionality staying the same.

Nothing is considered stable just yet. This manual therefore simply reflects the current state of the executable. ***Absolutely nothing*** on the following pages is set in stone. When the need arises, anything can (and will) be changed without prior notice.

If you are not willing to deal with this situation, you should wait for the stable version. Currently we expect the first release with (some) fixed interfaces to be available sometime in the autumn of 2008. Full stabilization will not happen soon, the TODO list is still very large.

LuaT_EX consists of a number of interrelated but (still) distinguishable parts:

- pdfT_EX version 1.40.9
- Aleph RC4 (from the T_EXLive repository)
- Lua 5.1.2 (+ coco 1.1.3)
- dedicated Lua libraries
- various T_EX extensions
- parts of FontForge 2008.11.17
- a still experimental version of the MetaPost library
- newly written compiled source code to glue it all together

Neither Aleph's I/O translation processes, nor tcx files, nor encT_EX can be used, these encoding-related functions are superseded by a Lua-based solution (reader callbacks). Also, some experimental pdfT_EX features are removed. These can be implemented in Lua instead.





2 Basic T_EX enhancements

2.1 Version information

There are three new primitives to test the version of LuaT_EX:

| primitive | explanation |
|-------------------------------|---|
| <code>\luatexversion</code> | a combination of major and minor number, as in pdfT _E X; the current current value is 33 |
| <code>\luatexrevision</code> | the revision, as in pdfT _E X; the current value is 0 |
| <code>\luatexdatestamp</code> | a combination of the local date and hour when the current executable was compiled, the syntax is identical to <code>\luatexrevision</code> ; the value for the executable that generated this document is 2009012209. |

Note that the `\luatexdatestamp` depends on both the compilation time and compilation place of the current executable, it is defined in terms of the local time. The purpose of this primitive is solely to be an aid in the development process, do not use it for anything besides debugging.

2.2 UNICODE text support

Text input and output is now considered to be Unicode text, so input characters can use the full range of Unicode ($2^{20} + 2^{16} = 10FFFF = 1114111$).

Later chapters will talk of characters and glyphs. Although these are not the interchangeable, they are closely related. During typesetting, a character is always converted to a suitable graphic representation of that character in a specific font. However, while processing a list of to-be-typeset nodes, its contents may still be seen as a character. Inside LuaT_EX there is not yet a clear separation between the two concepts. Until this is implemented, please do not be too harsh on us if we make errors in the usage of the terms.

Note: for now, it only makes sense to use values above the base plane ("**0xFFFF**") for `\mathcode` and `\catcode` assignments, since the hyphenation patterns are still limited to max. 16-bit values, so the other commands will not know what to do with those high values.

A few primitives are affected by this, all in a similar fashion: each of them has to accommodate for a larger range of acceptable numbers. For instance, `\char` now accepts values between 0 and 1114111. This should not be a problem for well-behaved input files, but it could create incompatibilities for input that would have generated an error when processed by older T_EX-based engines. The maximum number of allocations is "**10FFFF**" or $2^{20} + 2^{16}$ (21 bits). The maximum value that can be assigned are:

| primitive | bits | hex | numeric |
|-----------------------|------|--------|-------------------|
| <code>\char</code> | 21 | 10FFFF | $2^{20} + 2^{16}$ |
| <code>\chardef</code> | 21 | 10FFFF | $2^{20} + 2^{16}$ |
| <code>\lccode</code> | 21 | 10FFFF | $2^{20} + 2^{16}$ |



| | | | |
|-----------------------|----|--------|-------------------|
| <code>\uccode</code> | 21 | 10FFFF | $2^{20} + 2^{16}$ |
| <code>\sfcode</code> | 15 | 7FFF | 2^{15} |
| <code>\catcode</code> | 4 | F | 2^4 |

As far as the core engine is concerned, all input and output to text files is utf-8 encoded. Input files can be pre-processed using the [reader](#) callback. This will be explained in a later chapter.

Output in byte-sized chunks can be achieved by using characters just outside of the valid unicode range, starting at the value 1.114.112 (0x110000). When the times comes to print a character $c \geq 1.114.112$, Lua_T_EX will actually print the single byte corresponding to $c - 1.114.112$.

Output to the terminal uses `^^` notation for the lower control range ($c < 32$), with the exception of `^^I`, `^^J` and `^^M`. These are considered ‘safe’ and therefore printed as-is.

Normalization of the Unicode input can be handled by a macro package during callback processing (this will be explained in [section 4.8.2](#)).

2.3 Wide math characters

Text handling is now extended up to the full Unicode range. The extension from 8-bit to 16-bit was already present in Aleph by means of a set of extra primitives starting with the `\o` prefix, the extension to full Unicode (the `\U` prefix) is compatible with X₃_T_EX.

The math primitives from T_EX and Aleph are kept mostly as they are, except for the ones that convert from input to math commands like `mathcode` and `omathcode` that now allow for a larger argument on the left hand side of the equals sign.

| primitive | max index/bits | value range (in hex) |
|----------------------------|----------------|----------------------------|
| <code>\mathchardef</code> | 15 | 8000 |
| <code>\mathcode</code> | 8=15 | FF = 800 |
| <code>\delcode</code> | 8=24 | FF = FFFFF |
| <code>\mathchar</code> | 15 | 7FFF |
| <code>\mathaccent</code> | 15 | 7FFF |
| <code>\delimiter</code> | 27 | 7FFFFFFF |
| <code>\radical</code> | 27 | 7FFFFFFF |
| <code>\omathchardef</code> | 27 | 8000000 |
| <code>\omathcode</code> | 21=27 | 10FFFF = 8000000 |
| <code>\odelcode</code> | 21=24+24 | 10FFFF = FFFFFFF + FFFFFFF |
| <code>\omathchar</code> | 27 | 7FFFFFFF |
| <code>\omathaccent</code> | 27 | 7FFFFFFF |
| <code>\odelimiter</code> | 27+24 | 7FFFFFFF + FFFFFFF |
| <code>\oradical</code> | 27+24 | 7FFFFFFF + FFFFFFF |
| <code>\Umathchardef</code> | 3+8+21 | 8+FF+10FFFF |
| <code>\Umathcode</code> | 21=3+8+21 | 10FFFF = 8+FF+10FFFF |
| <code>\Udelcode</code> | 21=8+21 | 10FFFF = FF+10FFFF |
| <code>\Umathchar</code> | 3+8+21 | 8+FF+10FFFF |
| <code>\Umathaccent</code> | 3+8+21 | 8+FF+10FFFF |



| | | |
|----------------------------|--------|---------------------|
| <code>\Udelimiter</code> | 3+8+21 | 8+FF+10FFFF |
| <code>\Uradical</code> | 8+21 | 7FFFFFFF + FFFFFFFF |
| <code>\Umathcharnum</code> | 32 | see below |
| <code>\Umathcodenum</code> | 21=32 | 10FFFF = see below |
| <code>\Udelcodenum</code> | 21=29 | 10FFFF = see below |

Note: for `\Umathcharnum` and `\Umathcodenum`, all information is packed into a single signed integer: the lowest 21 bits are the character code, the 3 bit above that represent the math class, and the family data is kept in the topmost bits (This means that the values for math families 128–255 are actually negative). For `\Udelcodenum` there is no math class; the math family information is stored in the bits directly on top of the character code. Using these three commands is not as natural as using the two- and three-value commands, so unless you know exactly what you are doing and absolutely require the speedup resulting from the faster input scanning, it is better to use the verbose commands instead.

2.4 Extended tables

All traditional T_EX and ϵ -T_EX registers can be 16 bit numbers as in Aleph. The affected commands are:

| | | | |
|----------------------|-------------------------|-----------------------|----------------------|
| <code>\count</code> | <code>\countdef</code> | <code>\unhbox</code> | <code>\ht</code> |
| <code>\dimen</code> | <code>\dimendef</code> | <code>\unvbox</code> | <code>\dp</code> |
| <code>\skip</code> | <code>\skipdef</code> | <code>\copy</code> | <code>\setbox</code> |
| <code>\muskip</code> | <code>\muskipdef</code> | <code>\unhcopy</code> | <code>\vsplit</code> |
| <code>\marks</code> | <code>\toksdef</code> | <code>\unvcopy</code> | |
| <code>\toks</code> | <code>\box</code> | <code>\wd</code> | |

The glyph properties (like `\efcode`) introduced in pdfT_EX that deal with font expansion (hz) and character protruding are also 16 bit. Because font memory management has been rewritten, these character properties are no longer shared among fonts instances that originate from the same metric file

2.5 Attribute registers

Attributes are a completely new concept in LuaT_EX. Syntactically, they behave a lot like counters: attributes obey T_EX's nesting stack and can be used after `\the` etc. just like the normal `\count` registers.

```
\attribute <16-bit number> <optional equals> <31-bit number>
\attributedef <csname> <optional equals> <16-bit number>
```

Conceptually, an attribute is either 'set' or 'unset'. Set attributes can only have values of 0 or more, otherwise they are considered unset and automatically remapped to an special negative value meaning 'unset' (currently that value is -1 , but please test on negativity, not on a specific value). All attributes start out in the 'unset' state (in iniT_EX).

Attributes can be used as extra counter values, but their usefulness comes mostly from the fact that the numbers and values of all 'set' attributes are attached to all nodes created in their scope. These can then be queried from any Lua code that deals with node processing. Future versions of LuaT_EX will probably



be using specific negative attribute ids for internal use. Further information about how to use attributes for node list processing from Lua is given in **chapter 7**.

2.5.1 Box attributes

Nodes typically receive the list of attributes that is in effect when they are created. This moment can be quite asynchronous. For example: in paragraph building, the individual line boxes are created after the `\par` command has been processed, so they will receive the list of attributes that is in effect then, not the attributes that were in effect in, say, the first or third line of the paragraph.

Similar situations happen in LuaTeX regularly. A few of the more obvious problematic cases are dealt with: the attributes for nodes that are created during hyphenation and ligaturing borrow their attributes from their surrounding glyphs, and it is possible to influence box attributes directly.

When you assemble a box in a register, the attributes of the nodes contained in the box are unchanged when such a box is placed, unboxed, or copied. In this respect attributes act the same as characters that have been converted to references to glyphs in fonts. For instance, when you use attributes to implement color support, each node carries information about its color. In that case, unless you implement mechanisms that deal with it, applying a color to already boxed material will have no effect. Keep in mind that this incompatibility is mostly due to the fact that specials and literals are a more unnatural approach to colors than attributes.

Many other inserted nodes, like the nodes resulting from math mode and alignments, are processed ‘out of order’, and will have the attributes that are in effect at the precise moment of creation (which is often later than expected). This area needs studying, and is in fact one of the reasons for a beta at this moment.

It is possible to fine-tune the list of attributes that are applied to a `hbox`, `vbox` or `vtop` by the use of the keyword `attr`. An example:

```
\attribute2=5
\setbox0=\hbox {Hello}
\setbox2=\hbox attr1=12 attr2=-1{Hello}
```

This will set the attribute list of box 2 to $1 = 12$, and the attributes of box 0 will be $2 = 5$. As you can see, assigning a negative value causes an attribute to be ignored.

The `attr` keyword(s) should come before a `to` or `spread`, if that is also specified.

2.6 LUA related primitives

In order to merge Lua code with TeX input, a few new primitives are needed. LuaTeX has support for 65536 separate Lua interpreter states. States are automatically created based on the integer argument to the primitives `\directlua` and `\latelua`.



2.6.1 `\directlua`

The primitive `\directlua` is used to execute Lua code immediately. The syntax is

```
\directlua <16-bit number> <general text>
\directlua name <general text> <16-bit number> <general text>
```

The second `<general text>` is expanded fully, and then fed into the Lua interpreter state indicated by the `<16-bit number>`. If the state does not exist yet, it will be initialized automatically. After reading and expansion has been applied to the `<general text>`, the resulting token list is converted to a string as if it was displayed using `\the\toks`. On the Lua side, each `\directlua` block is treated as a separate chunk. In such a chunk you can use the `local` directive to keep your variables from interfering with those used by the macro package.

The conversion from and to a token list means that you normally can not use Lua line comments (starting with `--`) within the argument, as there typically will be only one ‘line’, so that comment will then run on until the end of the input. You will either need to use T_EX-style line comments (starting with `%`), or change the T_EX category codes locally. Another possibility is to say:

```
\begingroup
\endlinechar=10
\directlua ...
\endgroup
```

Then Lua line comments can be used, since T_EX does not replace line endings with spaces.

`name <general text>` specifies the name of the Lua chunk, mainly shown in the stack backtrace of error messages created by Lua code. The `<general text>` is expanded fully, thus macros can be used to generate the chunk name, i.e.

```
\directlua name{\jobname:\the\inputlineno} ...
```

to include the name of the input file as well as the input line into the chunk name.

The chunk name should not start with a `@`, or it will be displayed as a file name (this is a quirk in the current Lua implementation).

The `\directlua` command is expandable: the results of the Lua code become effective immediately. As an example, the following input:

```
$\pi = \directlua0{tex.print(math.pi)}$
```

will result in

```
 $\pi = 3.1415926535898$ 
```

Because the `<general text>` is a chunk, the normal Lua error handling is triggered if there is a problem in the included code. The Lua error messages should be clear enough, but the contextual information is still pretty bad. Typically, you will only see the line number of the right brace at the end of the code.



While on the subject of errors: some of the things you can do inside Lua code can break up LuaTeX pretty bad. If you are not careful while working with the node list interface, you may even end up with assertion errors from within the T_EX portion of the executable.

2.6.2 `\latelua`

`\latelua` stores Lua code in a whatsit that will be processed inside the output routine. Its intended use is a cross between `\pdfliteral` and `\write`. Within the Lua code, you can print pdf statements directly to the pdf file via `tex.print`, or you can write to other output streams via `texio.write` or simply using lua's I/O routines.

`\latelua` <16-bit number> <general text>

Expansion of macros etcetera in the <general text> is delayed until just before the whatsit is executed (like in `\write`). With regard to PDF output stream `\latelua` behaves as `\pdfliteral` page.

2.6.3 `\luaescapestring`

This primitive converts a T_EX token sequence so that it can be safely used as the contents of a Lua string: embedded backslashes, double and single quotes, and newlines and carriage returns are escaped. This is done by prepending an extra token consisting of a backslash with category code 12, and for the line endings, converting them to `n` and `r` respectively. The token sequence is fully expanded.

`\luaescapestring` <general text>

Most often, this command is not actually the best way to deal with the differences between the T_EX and Lua. In very short bits of Lua code it is often not needed, and for longer stretches of Lua code it is easier to keep the code in a separate file and load it using Lua's `dofile`:

```
\directlua0 { dofile('mysetups.lua')}
```

2.6.4 `\closeslua`

This primitive allows you to close a Lua state, freeing all of its used memory.

`\closeslua` <16-bit number>

You cannot close the initial Lua state 0, attempts to do so will be silently ignored. States are never closed automatically except when a fatal out of memory error occurs, at which point LuaTeX will exit anyway. Also be aware that Lua states are not closed immediately, but only when the `\output` routine comes into play next (because there may be pending `\latelua` calls).



2.7 New E-TeX primitives

2.7.1 `\clearmarks`

This primitive clears a marks class completely, resetting all three connected mark texts to empty.

```
\clearmarks <16-bit number>
```

2.7.2 `\noligs` and `\nokerns`

These primitives prohibit ligature and kerning insertion at the time when the initial node list is built by LuaTeX's main control loop. They are part of a temporary trick and will be removed in the near future. For now, you need to enable these primitives when you want to do node list processing of 'characters', where TeX's normal processing would get in the way.

```
\noligs <integer>
\nokerns <integer>
```

These primitives can now be implemented by overloading the ligature building and kerning functions, i.e. by assigning dummy functions to their associated callbacks.

2.7.3 `\formatname`

`\formatname`'s syntax is identical to `\jobname`.

In `iniTeX`, the expansion is empty. Otherwise, the expansion is the value that `\jobname` had during the `iniTeX` run that dumped the currently loaded format.

2.7.4 `\scantextokens`

The syntax of `\scantextokens` is identical to `\scantokens`. This primitive is a slightly adapted version of ϵ -TeX's `\scantokens`. The differences are:

- The last (and usually only) line does not have a `\endlinechar` appended
- `\scantextokens` never raises an EOF error, and it does not execute `\everyeof` tokens.
- The '... while end of file ...' error tests are not executed, allowing the expansion to end on a different grouping level or while a conditional is still incomplete.

2.7.5 Catcode tables

Catcode tables are a new feature that allows you to switch to a predefined catcode regime in a single statement. You can have a practically unlimited number of different tables.



The subsystem is backward compatible: if you never use the following commands, your document will not notice any difference in behavior compared to traditional T_EX.

The contents of each catcode table is independent from any other catcode tables, and their contents is stored and retrieved from the format file.

2.7.5.1 `\catcodetable`

`\catcodetable` <16-bit number>

The `\catcodetable` switches to a different catcode table. Such a table has to be previously created using one of the two primitives below, or it has to be zero. Table zero is initialized by `iniTEX`.

2.7.5.2 `\initcatcodetable`

`\initcatcodetable` <16-bit number>

The `\initcatcodetable` creates a new table with catcodes identical to those defined by `iniTEX`:

| | | | | |
|----|-----------------|--------|--------------|------------------------|
| 0 | - | | escape | |
| 5 | ^^M | return | car_ret | (this name may change) |
| 9 | ^^@ | null | ignore | |
| 10 | <space> | space | spacer | |
| 11 | a – z | | letter | |
| 11 | A – Z | | letter | |
| 12 | everything else | | other | |
| 14 | % | | comment | |
| 15 | ^^? | delete | invalid_char | |

The new catcode table is allocated globally: it will not go away after the current group has ended. If the supplied number is identical to the currently active table, an error is raised.

2.7.5.3 `\savecatcodetable`

`\savecatcodetable` <16-bit number>

`\savecatcodetable` copies the current set of catcodes to a new table with the requested number. The definitions in this new table are all treated as if they were made in the outermost level.

The new table is allocated globally: it will not go away after the current group has ended. If the supplied number is the currently active table, an error is raised.

2.7.6 `\suppressfontnotfounderror`

`\suppressfontnotfounderror = 1`



If this new integer parameter is non-zero, then LuaT_EX will not complain about font metrics that are not found. Instead it will silently skip the font assignment, making the requested csname for the font `\ifx` equal to `\nullfont`, so that it can be tested against that without bothering the user.

2.7.7 Font syntax

LuaT_EX will accept a braced argument as a font name:

```
\font\myfont = {cmr10}
```

This allows for embedded spaces, without the need for double quotes. Macro expansion takes place inside the argument.

2.8 Debugging

If `\tracingonline` is larger than 2, the node list display will also print the node number of the nodes.





3 LUA general

3.1 Initialization

3.1.1 L^AT_EX as a LUA interpreter

There are some situations that make L^AT_EX behave like a standalone Lua interpreter:

- if a `--luaonly` option is given on the commandline, or
- if the executable is named `texlua` (or `luatexlua`), or
- if the only non-option argument (file) on the commandline has the extension `lua` or `luc`.

In this mode, it will set Lua's `arg[0]` to the found script name, pushing preceding options in negative values and the rest of the commandline in the positive values, just like the Lua interpreter.

L^AT_EX will exit immediately after executing the specified Lua script and is, in effect, a somewhat bulky standalone Lua interpreter with a bunch of extra preloaded libraries.

3.1.2 L^AT_EX as a LUA byte compiler

There are two situations that make L^AT_EX behave like the Lua byte compiler:

- if a `--luaonly` option is given on the commandline, or
- if the executable is named `texluac`

In this mode, L^AT_EX is exactly like `luac` from the standalone Lua distribution, except that it does not have the `-l` switch, and that it accepts (but ignores) the `--luaonly` switch.

3.1.3 Other commandline processing

When the L^AT_EX executable starts, it looks for the `--lua` commandline option. If there is no `--lua` option, the commandline is interpreted in a similar fashion as in traditional pdf_T_EX and Aleph. But if the option is present, L^AT_EX will enter an alternative mode of commandline parsing in comparison to the standard web2c programs.

In this mode, a small series of actions is taken in order. At first, it will only interpret a small subset of the commandline directly:

| | |
|-------------------------|--|
| <code>--lua=s</code> | load and execute a Lua initialization script |
| <code>--safer</code> | disable easily exploitable Lua commands |
| <code>--nosocket</code> | disable the Lua socket library |
| <code>--help</code> | display help and exit |
| <code>--version</code> | display version and exit |



Now it searches for the requested Lua initialization script. If it can not be found using the actual name given on the commandline, a second attempt is made by prepending the value of the environment variable `LUATEXDIR`, if that variable is defined.

Then it checks the `--safer` switch. You can use that to disable some Lua commands that can easily be abused by a malicious document. At the moment, this switch `nils` the following functions:

library functions

```
os      execute exec setenv rename remove tmpdir
io      popen output tmpfile
lfs      rmdir mkdir chdir lock touch
```

And it makes `io.open()` fail on files that are opened for anything besides reading.

Next the initialization script is loaded and executed. From within the script, the entire commandline is available in the Lua table `arg`, beginning with `arg[0]`, containing the name of the executable.

Commandline processing happens very early on. So early, in fact, that none of T_EX's initializations have taken place yet. For that reason, the tables that deal with typesetting, like `tex`, `token`, `node` and `pdf`, are off-limits during the execution of the startup file (they are nilled). Special care is taken that `texio.write` and `texio.write_nl` function properly, so that you can at least report your actions to the log file when (and if) it eventually becomes opened (note that T_EX does not even know its `\jobname` yet at this point). See **chapter 4** for more information about the LuaT_EX-specific Lua extension tables.

The Lua initialization script is loaded into Lua state 0, and everything you do will remain visible during the rest of the run, with the exception of the aforementioned `tex`, `token`, `node` and `pdf` tables: those will be initialized to their documented state after the execution of the script. You should not store anything in variables or within tables with these four global names, as they will be overwritten completely.

We recommend you use the startup file only for your own T_EX-independent initializations (if you need any), to parse the commandline, set values in the `texconfig` table, and register the callbacks you need. LuaT_EX will fetch some of the other commandline options from the `texconfig` table at the end of script execution (see the description of the `texconfig` table later on in this document for more details on which ones exactly).

Unless the `texconfig` table tells LuaT_EX not to initialize kpathsea at all (set `texconfig.kpse_init` to `false` for that), LuaT_EX acts on three more commandline options after the initialization script is finished:

| flag | meaning |
|---------------------------|--------------------------------------|
| <code>--fmt=s</code> | set the format name |
| <code>--progname=s</code> | set the progname (only for kpathsea) |
| <code>--ini</code> | enable iniT _E X mode |

In order to initialize the built-in kpathsea library properly, LuaT_EX needs to know the correct `progname` to use, and for that it needs to check `--progname` (and `--ini` and `--fmt`, if `--progname` is missing).



3.2 LUA changes

The C coroutine (coco) patches from luajit are applied to the Lua core, the used version is 1.1.3. See <http://luajit.org/coco.html> for details.

The `read("*line")` function from the io library has been adjusted so that it is line-ending neutral: any of LF, CR or CR+LF are acceptable line endings.

The `tostring()` printer for numbers has been changed so that it returns 0 instead of something like `2e-5` (which confused T_EX enormously) when the value is so small that T_EX cannot distinguish it from zero.

Dynamic loading of `.so` and `.dll` files is disabled on all platforms.

`luafilesystem` has been extended with two extra boolean functions (`isdir(filename)` and `isfile(filename)`) and one extra string field in its attributes table (`permissions`).

The `string` library has an extra function: `string.explode(s[,m])`. This function returns an array containing the string argument `s` split into substrings based on the value of the string argument `m`. The second argument is a string that is either empty (this splits the string into characters), a single character (this splits on each occurrence of that character, possibly introducing empty strings), or a single character followed by the plus sign `+` (this special version does not create empty substrings). The default value for `m` is `+` (multiple spaces).

Note: `m` is not hidden by surrounding braces (as it would be if this function was written in T_EX macros).

The `string` library also has six extra iterators that return strings piecemeal:

- `string.utfvalues(s)` (returns an integer value in the Unicode range)
- `string.utfcharacters(s)` (returns a string with a single utf-8 token in it)
- `string.characters(s)` (a string containing one byte)
- `string.characterpairs(s)` (two strings each containing one byte) will produce an empty second string in the string length was odd.
- `string.bytes(s)` (a single byte value)
- `string.bytepairs(s)` (two byte values) Will produce nil instead of a number as its second return value if the string length was odd.

The `string.characterpairs()` and `string.bytepairs()` are useful especially in the conversion of UTF-16 encoded data into UTF-8.

Note: The `string` library functions `find` etc. are not Unicode-aware. In cases where this is required (i.e. because the pattern used for searching contains characters above code point 127), the corresponding functions from `unicode.utf8` should be used.

The `os` library has a few extra functions and variables:

- `os.exec(commandline)` is a variation on `os.execute`.
The `commandline` can be either a single string or a single table.



If the argument is a table: LuaTeX first checks if there is a value at integer index zero. If there is, this is the command to be executed. Otherwise, it will use the value at integer index one. (if neither are present, nothing at all happens).

The set of consecutive values starting at integer 1 in the table are the arguments that are passed on to the command (the value at index 1 becomes `argv[0]`). The command is searched for in the execution path, so there is normally no need to pass on a fully qualified pathname.

If the argument is a string, then it is automatically converted into a table by splitting on whitespace. In this case, it is impossible for the command and first argument to differ from each other.

In the string argument format, whitespace can be protected by putting (part of) an argument inside single or double quotes. One layer of quotes is interpreted by LuaTeX, and all occurrences of `\`, `'` or `\\` within the quoted text are un-escaped. In the table format, there is no string handling taking place.

This function normally does not return control back to the Lua script: the command will replace the current process. However, it will return the two values `nil` and `'error'` if there was a problem while attempting to execute the command.

On windows, the current process is actually kept in memory until after the execution of the command has finished. This prevents crashes in situations where TeX Lua scripts are run inside integrated TeX environments.

The original reason for this command is that it cleans out the current process before starting the new one, making it especially useful for use in TeX Lua.

- `os.spawn(commandline)` is a returning version of `os.exec`, with otherwise identical calling conventions.

If the command ran ok, then the return value is the exit status of the command. Otherwise, it will return the two values `nil` and `'error'`.

- `os.setenv('key', 'value')` This sets a variable in the environment. Passing `nil` instead of a value string will remove the variable.
- `os.env` This is a hash table containing a dump of the variables and values in the process environment at the start of the run. It is writeable, but the actual environment is *not* updated automatically.
- `os.gettimeofday()` Returns the current 'Unix time', but as a float. This function is not available on the SunOS platforms, so do not use this function for portable documents.
- `os.times()` Returns the current process times cf. the Unix C library 'times' call in seconds. This function is not available on the MS Windows and SunOS platforms, so do not use this function for portable documents.
- `os.tmpdir()` This will create a directory in the 'current directory' with the name `luatex.XXXXXX` where the X-es are replaced by a unique string. The function also returns this string, so you can `lfs.chdir()` into it, or `nil` if it failed to create the directory. The user is responsible for cleaning up at the end of the run, it does not happen automatically.
- `os.type` This is a string that gives a global indication of the class of operating system. The possible values are currently `windows`, `unix`, and `msdos` (you are unlikely to find this value 'in the wild').
- `os.name` This is a string that gives a more precise indication of the operating system. These possible values are not yet fixed, and for `os.type` values `windows` and `msdos`, the `os.name` values are simply `windows` and `msdos`



The list for the type `unix` is more precise: `linux`, `freebsd`, `openbsd`, `solaris`, `sunos` (pre-solaris), `hpux`, `irix`, `macosx`, `bsd` (unknown, but bsd-like), `sysv` (unknown, but sysv-like), `generic` (unknown).

(`os.version` is planned as a future extension)

In stock Lua, many things depend on the current locale. In LuaT_EX, we can't do that, because it makes documents unportable. While LuaT_EX is running it forces the following locale settings:

```
LC_CTYPE=C
LC_COLLATE=C
LC_NUMERIC=C
```

3.3 LUA modules

Some modules that are normally external to Lua are statically linked in with LuaT_EX, because they offer useful functionality:

- `slnunicode`, from the `Selene` libraries, <http://luaforge.net/projects/sln>. (version 1.1)
This library has been slightly extended so that the `unicode.utf8.*` functions also accept the first 256 values of plane 18. This is the range LuaT_EX uses for raw binary output, as explained above,
- `luazip`, from the kepler project, <http://www.keplerproject.org/luazip/>. (version 1.2.1, but patched for compilation with Lua 5.2)
- `luafilesystem`, also from the kepler project, <http://www.keplerproject.org/luafilesystem/>. (version 1.4.1)
- `lpeg`, by Roberto Ierusalimsky, <http://www.inf.puc-rio.br/~roberto/lpeg.html>. (version 0.8.1)
Note: `lpeg` is not Unicode-aware, but interprets strings on a byte-per-byte basis. This mainly means that `lpeg.S` cannot be used with characters above code point 127, since those characters are encoded using two bytes, and thus `lpeg.S` will look for one of those two bytes when matching, not the combination of the two.
The same is true for `lpeg.R`, although the latter will display an error message if used with characters above code point 127: i.e. `lpeg.R('ä')` results in the message `bad argument #1 to 'R' (range must have two characters)`, since to `lpeg`, `ä` is two 'characters' (bytes), so `ä` totals three.
- `lzlib`, by Tiago Dionizio, <http://mega.ist.utl.pt/~tngd/lua/>. (version 0.2)
- `md5`, by Roberto Ierusalimsky <http://www.inf.puc-rio.br/~roberto/md5/md5-5/md5.html>.
- `luasocket`, by Diego Nehab <http://www.tecgraf.puc-rio.br/~diego/professional/luasocket/> (version 2.0.2).

Note: the `.lua` support modules from `luasocket` are also preloaded inside the executable, there are no external file dependencies.





4 L^AT_EX LUA Libraries

The interfacing between T_EX and Lua is facilitated by a set of library modules. The Lua libraries in this chapter are all defined and initialized by the LuaT_EX executable. Together, they allow Lua scripts to query and change a number of T_EX's internal variables, run various internal functions T_EX, and set up LuaT_EX's hooks to execute Lua code.

4.1 The tex library

The `tex` table contains a large list of virtual internal T_EX parameters that are partially writable.

The designation 'virtual' means that these items are not properly defined in Lua, but are only frontends that are handled by a metatable that operates on the actual T_EX values. As a result, most of the Lua table operators (like `pairs` and `#`) do not work on such items.

At the moment, it is possible to access almost every parameter that has these characteristics:

- You can use it after `\the`
- It is a single token.
- Some special others, see the list below

This excludes parameters that need extra arguments, like `\the\scriptfont`.

The subset comprising simple integer and dimension registers are writable as well as readable (stuff like `\tracingcommands` and `\parindent`).

4.1.1 Integer parameters

The integer parameters accept and return Lua numbers.

Read-write:

| | |
|---------------------------------------|--------------------------------------|
| <code>tex.adjdemerits</code> | <code>tex.errorcontextlines</code> |
| <code>tex.binoppenalty</code> | <code>tex.escapechar</code> |
| <code>tex.brokenpenalty</code> | <code>tex.exhyphenpenalty</code> |
| <code>tex.catcodetable</code> | <code>tex.fam</code> |
| <code>tex.clubpenalty</code> | <code>tex.finalhyphendemerits</code> |
| <code>tex.day</code> | <code>tex.floatingpenalty</code> |
| <code>tex.defaultthyphenchar</code> | <code>tex.globaldefs</code> |
| <code>tex.defaultskewchar</code> | <code>tex.hangafter</code> |
| <code>tex.delimiterfactor</code> | <code>tex.hbadness</code> |
| <code>tex.displaywidowpenalty</code> | <code>tex.holdinginserts</code> |
| <code>tex.doublehyphendemerits</code> | <code>tex.hyphenpenalty</code> |
| <code>tex.endlinechar</code> | <code>tex.interlinepenalty</code> |



| | |
|---|--------------------------------------|
| <code>tex.language</code> | <code>tex.pdfuniqueresname</code> |
| <code>tex.lastlinefit</code> | <code>tex.postdisplaypenalty</code> |
| <code>tex.lefthyphenmin</code> | <code>tex.predisplaydirection</code> |
| <code>tex.linepenalty</code> | <code>tex.predisplaypenalty</code> |
| <code>tex.localbrokenpenalty</code> | <code>tex.pretolerance</code> |
| <code>tex.localinterlinepenalty</code> | <code>tex.relpenalty</code> |
| <code>tex.looseness</code> | <code>tex.righthyphenmin</code> |
| <code>tex.mag</code> | <code>tex.savinghyphcodes</code> |
| <code>tex.maxdeadcycles</code> | <code>tex.savingvdiscards</code> |
| <code>tex.month</code> | <code>tex.showboxbreadth</code> |
| <code>tex.newlinechar</code> | <code>tex.showboxdepth</code> |
| <code>tex.outputpenalty</code> | <code>tex.time</code> |
| <code>tex.pausing</code> | <code>tex.tolerance</code> |
| <code>tex.pdfadjustinterwordglue</code> | <code>tex.tracingassigns</code> |
| <code>tex.pdfadjustspacing</code> | <code>tex.tracingcommands</code> |
| <code>tex.pdfappendkern</code> | <code>tex.tracinggroups</code> |
| <code>tex.pdfcompresslevel</code> | <code>tex.tracingifs</code> |
| <code>tex.pdfdecimaldigits</code> | <code>tex.tracinglostchars</code> |
| <code>tex.pdfgamma</code> | <code>tex.tracingmacros</code> |
| <code>tex.pdfgentounicode</code> | <code>tex.tracingnesting</code> |
| <code>tex.pdfimageapplygamma</code> | <code>tex.tracingonline</code> |
| <code>tex.pdfimagegamma</code> | <code>tex.tracingoutput</code> |
| <code>tex.pdfimagehicolor</code> | <code>tex.tracingpages</code> |
| <code>tex.pdfimageresolution</code> | <code>tex.tracingparagraphs</code> |
| <code>tex.pdfinclusionerrorlevel</code> | <code>tex.tracingrestores</code> |
| <code>tex.pdfminorversion</code> | <code>tex.tracingscantokens</code> |
| <code>tex.pdfobjcompresslevel</code> | <code>tex.tracingstats</code> |
| <code>tex.pdfoutput</code> | <code>tex.uchyph</code> |
| <code>tex.pdfpagebox</code> | <code>tex.vbadness</code> |
| <code>tex.pdfpkresolution</code> | <code>tex.widowpenalty</code> |
| <code>tex.pdfprependkern</code> | <code>tex.year</code> |
| <code>tex.pdfprotrudechars</code> | |
| <code>tex.pdftracingfonts</code> | |



Read-only:

| | | |
|----------------------------------|---------------------------|------------------------------|
| <code>tex.deadcycles</code> | <code>tex.parshape</code> | <code>tex.spacefactor</code> |
| <code>tex.insertpenalties</code> | <code>tex.prevgraf</code> | |

4.1.2 Dimension parameters

The dimension parameters accept Lua numbers (signifying scaled points) or strings (with included dimension). The result is always a number in scaled points.

Read-write:

| | | |
|-------------------------------------|-------------------------------------|----------------------------------|
| <code>tex.boxmaxdepth</code> | <code>tex.pagebottomoffset</code> | <code>tex.pdfpageheight</code> |
| <code>tex.delimitershortfall</code> | <code>tex.pageheight</code> | <code>tex.pdfpagewidth</code> |
| <code>tex.displayindent</code> | <code>tex.pageleftoffset</code> | <code>tex.pdfpxdimen</code> |
| <code>tex.displaywidth</code> | <code>tex.pagerightoffset</code> | <code>tex.pdfthreadmargin</code> |
| <code>tex.emergencystretch</code> | <code>tex.pagetopoffset</code> | <code>tex.pdfvorigin</code> |
| <code>tex.hangindent</code> | <code>tex.pagewidth</code> | <code>tex.predisplaysize</code> |
| <code>tex.hfuzz</code> | <code>tex.parindent</code> | <code>tex.scriptspace</code> |
| <code>tex.hoffset</code> | <code>tex.pdfdestmargin</code> | <code>tex.splitmaxdepth</code> |
| <code>tex.hsize</code> | <code>tex.pdfeachlinedepth</code> | <code>tex.vfuzz</code> |
| <code>tex.lineskiplimit</code> | <code>tex.pdfeachlineheight</code> | <code>tex.voffset</code> |
| <code>tex.mathsurround</code> | <code>tex.pdffirstlineheight</code> | <code>tex.vsize</code> |
| <code>tex.maxdepth</code> | <code>tex.pdfhorigin</code> | |
| <code>tex.nulldelimiterspace</code> | <code>tex.pdflastlinedepth</code> | |
| <code>tex.overfullrule</code> | <code>tex.pdflinkmargin</code> | |

Read-only:

| | | |
|------------------------------------|------------------------------|----------------------------|
| <code>tex.pagedepth</code> | <code>tex.pagegoal</code> | <code>tex.prevdepth</code> |
| <code>tex.pagefillllstretch</code> | <code>tex.pageshrink</code> | |
| <code>tex.pagefillstretch</code> | <code>tex.pagestretch</code> | |
| <code>tex.pagefilstretch</code> | <code>tex.pagetotal</code> | |

4.1.3 Direction parameters

The direction parameters are read-only and return a Lua string.

| | | |
|--------------------------|--------------------------|--------------------------|
| <code>tex.bodydir</code> | <code>tex.pagedir</code> | <code>tex.textdir</code> |
| <code>tex.mathdir</code> | <code>tex.pardir</code> | |

4.1.4 Glue parameters

All glue parameters are read-only and return a userdata object that represents a `glue_spec` node.

| | | |
|---|-----------------------------------|---|
| <code>tex.abovedisplayshort-skip</code> | <code>tex.abovedisplayskip</code> | <code>tex.belowdisplayshort-skip</code> |
| | <code>tex.baselineskip</code> | |



| | | |
|-----------------------------------|-------------------------------|-----------------------------|
| <code>tex.belowdisplayskip</code> | <code>tex.rightskip</code> | <code>tex.xspaceskip</code> |
| <code>tex.leftskip</code> | <code>tex.spaceskip</code> | |
| <code>tex.lineskip</code> | <code>tex.splittopskip</code> | |
| <code>tex.parfillskip</code> | <code>tex.tabskip</code> | |
| <code>tex.parskip</code> | <code>tex.topskip</code> | |

4.1.5 Muglue parameters

All muglue parameters are read-only and return a Lua string.

| | |
|------------------------------|-----------------------------|
| <code>tex.medmuskip</code> | <code>tex.thinmuskip</code> |
| <code>tex.thickmuskip</code> | |

4.1.6 Tokenlist parameters

All tokenlist parameters are read-only and return a Lua string.

| | | |
|-------------------------------|----------------------------|-----------------------------------|
| <code>tex.errhelp</code> | <code>tex.everyjob</code> | <code>tex.pdfpageattr</code> |
| <code>tex.everycr</code> | <code>tex.everymath</code> | <code>tex.pdfpageresources</code> |
| <code>tex.everydisplay</code> | <code>tex.everypar</code> | <code>tex.pdfpagesattr</code> |
| <code>tex.everyeof</code> | <code>tex.everyvbox</code> | <code>tex.pdfpkmode</code> |
| <code>tex.everyhbox</code> | <code>tex.output</code> | |

4.1.7 Convert commands

All ‘convert’ commands are read-only and return a Lua string. The supported commands at this moment are:

| | |
|-----------------------------------|---|
| <code>tex.AlephVersion</code> | <code>tex.pdfTeXrevision</code> |
| <code>tex.Alephrevision</code> | <code>tex.fontname(number)</code> |
| <code>tex.OmegaVersion</code> | <code>tex.pdfFontname(number)</code> |
| <code>tex.Omegarevision</code> | <code>tex.pdfFontobjnum(number)</code> |
| <code>tex.eTeXVersion</code> | <code>tex.pdfFontSize(number)</code> |
| <code>tex.eTeXrevision</code> | <code>tex.uniformdeviate(number)</code> |
| <code>tex.formatname</code> | <code>tex.number(number)</code> |
| <code>tex.jobname</code> | <code>tex.roman numeral(number)</code> |
| <code>tex.luatexrevision</code> | <code>tex.pdfPageRef(number)</code> |
| <code>tex.luatexdatestamp</code> | <code>tex.pdfXformname(number)</code> |
| <code>tex.pdfNormaldeviate</code> | <code>tex.fontIdentifier(number)</code> |
| <code>tex.pdfTeXBanner</code> | |

If you are wondering why this list looks haphazard; these are all the cases of the ‘convert’ internal command that do not require an argument, as well as the ones that require only a simple numeric value.

The special (lua-only) case of `tex.fontIdentifier` returns the `csname` string that matches a font id number (if there is one).



4.1.8 Last item commands

All ‘last item’ commands are read-only and return a number.

The supported commands at this moment are:

| | | |
|--------------------------------------|------------------------------------|-------------------------------------|
| <code>tex.lastpenalty</code> | <code>tex.pdfplastannot</code> | <code>tex.eTeXversion</code> |
| <code>tex.lastkern</code> | <code>tex.pdfplastxpos</code> | <code>tex.currentgrouplevel</code> |
| <code>tex.lastskip</code> | <code>tex.pdflastypos</code> | <code>tex.currentgrouptype</code> |
| <code>tex.lastnodetype</code> | <code>tex.pdfrandomseed,</code> | <code>tex.currentiflevel</code> |
| <code>tex.inputlineno</code> | <code>tex.pdfplastlink</code> | <code>tex.currentiftype</code> |
| <code>tex.badness</code> | <code>tex.luatexversion</code> | <code>tex.currentifbranch</code> |
| <code>tex.pdfptextversion</code> | <code>tex.Alephversion</code> | <code>tex.pdfplastximagecol-</code> |
| <code>tex.pdfplastobj</code> | <code>tex.Omegaversion</code> | <code>ordepth</code> |
| <code>tex.pdfplastxform</code> | <code>tex.Alephminorversion</code> | |
| <code>tex.pdfplastximage</code> | <code>tex.Omegaminorversion</code> | |
| <code>tex.pdfplastximagepages</code> | <code>tex.eTeXminorversion</code> | |

4.1.9 Attribute, count, dimension, skip and token registers

TeX’s attributes (`\attribute`), counters (`\count`), dimensions (`\dimen`), skips (`\skip`) and token (`\toks`) registers can be accessed and written to using five virtual sub-tables of the `tex` table:

| | | |
|----------------------------|------------------------|-----------------------|
| <code>tex.attribute</code> | <code>tex.dimen</code> | <code>tex.toks</code> |
| <code>tex.count</code> | <code>tex.skip</code> | |

It is possible to use the names of relevant `\attributedef`, `\countdef`, `\dimendef`, `\skipdef`, or `\toksdef` control sequences as indices to these tables:

```
tex.count.scratchcounter = 0
enormous = tex.dimen['maxdimen']
```

In this case, LuaTeX looks up the value for you on the fly. You have to use a valid `\countdef` (or `\attributedef`, or `\dimendef`, or `\skipdef`, or `\toksdef`), anything else will generate an error (the intent is to eventually also allow `<chardef tokens>` and even macros that expand into a number).

The attribute and count registers accept and return Lua numbers.

The dimension registers accept Lua numbers (in scaled points) or strings (with an included absolute dimension; `em` and `ex` and `px` are forbidden). The result is always a number in scaled points.

The token registers accept and return Lua strings. Lua strings are converted to and from token lists using `\the\toks` style expansion: all category codes are either space (10) or other (12).

The skip registers accept and return `glue_spec` userdata node objects (see the description of the node interface elsewhere in this manual).

As an alternative to array addressing, there are also accessor functions defined:



```

tex.setdimen(<number> n, <string> s)
tex.setdimen(<string> s, <string> s)
tex.setdimen(<number> n, <number> n)
tex.setdimen(<string> s, <number> n)
<number> n = tex.getdimen(<number> n)
<number> n = tex.getdimen(<string> s)

tex.setcount(<number> n, <number> n)
tex.setcount(<string> s, <number> n)
<number> n = tex.getcount(<number> n)
<number> n = tex.getcount(<string> s)

tex.settoks (<number> n, <string> s)
tex.settoks (<string> s, <string> s)
<string> s = tex.gettoks (<number> n)
<string> s = tex.gettoks (<string> s)

tex.setskip (<number> n, <node> s)
tex.setskip (<string> s, <node> s)
<node> s = tex.getskip (<number> n)
<node> s = tex.getskip (<string> s)

```

4.1.10 Box registers

The current dimensions of \box registers can be read and altered using three other virtual sub-tables :

```

tex.wd
tex.ht
tex.dp

```

Boxes are indexed by number or by name. In macro packages [chardef](#) is normally used to refer to allocated box registers and LuaTeX is able to deal with these symbolic names.

The box size registers accept Lua numbers (in scaled points) or strings (with included dimension). The result is always a number in scaled points.

As an alternative to array addressing, there are also accessor functions defined:

```

tex.setboxwd(<number> n, <number> n)
<number> n = tex.getboxwd(<number> n)

tex.setboxht(<number> n, <number> n)
<number> n = tex.getboxht(<number> n)

tex.setboxdp(<number> n, <number> n)
<number> n = tex.getboxdp(<number> n)

```



It is also possible to set and query actual boxes, using the node interface as defined in the [node](#) library:

```
tex.box
```

for array access, or

```
tex.setbox(<number> n, <node> s)
<node> n = tex.getbox(<number> n)
```

for function-based access.

Be warned that an assignment like

```
tex.box[0] = tex.box[2]
```

does not copy the node list, it just duplicates a node pointer. If `\box2` will be cleared by T_EX commands later on, the contents of `\box0` becomes invalid as well. To prevent this from happening, always use `node.copy_list()` unless you are assigning to a temporary variable:

```
tex.box[0] = node.copy_list(tex.box[2])
```

4.1.11 Special list heads

The virtual table `tex.lists` contains the set of internal registers that keep track of building page lists.

| field | description |
|-----------------|---|
| page_ins_head | circular list of pending insertions |
| contrib_head | the recent contributions |
| page_head | the page-so-far |
| hold_head | used for heldover items for next page |
| adjust_head | head of the current <code>\adjust</code> list |
| pre_adjust_head | head of the current <code>\adjust pre</code> list |

4.1.12 Print functions

The `tex` table also contains the three print functions that are the major interface from Lua scripting to T_EX.

The arguments to these three functions are all stored in an in-memory virtual file that is fed to the T_EX scanner as the result of the expansion of `\directlua`.

The total amount of returnable text from a `\directlua` command is only limited by available system ram. However, each separate printed string has to fit completely in T_EX's input buffer.

The result of using these functions from inside callbacks is undefined at the moment.



4.1.12.1 `tex.print`

```
tex.print(<string> s, ...)  
tex.print(<number> n, <string> s, ...)
```

Each string argument is treated by T_EX as a separate input line.

The optional parameter can be used to print the strings using the catcode regime defined by `\catcodetable n`. If `n` is not a valid catcode table, then it is ignored, and the currently active catcode regime is used instead.

The very last string of the very last `tex.print()` command in a `\directlua` will not have the `\endlinechar` appended, all others do.

4.1.12.2 `tex.sprint`

```
tex.sprint(<string> s, ...)  
tex.sprint(<number> n, <string> s, ...)
```

Each string argument is treated by T_EX as a special kind of input line that makes it suitable for use as a partial line input mechanism:

- T_EX does not switch to the ‘new line’ state, so that leading spaces are not ignored.
- No `\endlinechar` is inserted.
- Trailing spaces are not removed.
(Note that this does not prevent T_EX itself from eating spaces as result of interpreting the line. For example, in

```
before\directlua0{tex.sprint("\relax")tex.sprint(" inbetween")}after
```

the space before `inbetween` will be gobbled as a result of the ‘normal’ scanning of `\relax`).

4.1.12.3 `tex.write`

```
tex.write(<string> s, ...)
```

Each string argument is treated by T_EX as a special kind of input line that makes is suitable for use as a quick way to dump information:

- All catcodes on that line are either ‘space’ (for ‘ ’) or ‘character’ (for all others).
- There is no `\endlinechar` appended.



4.1.13 Helper functions

4.1.13.1 `tex.round`

```
<number> n = tex.round(<number> o)
```

Rounds Lua number `o`, and returns a number that is in the range of a valid T_EX register value. If the number starts out of range, it generates a ‘number to big’ error as well.

4.1.13.2 `tex.scale`

```
<number> n = tex.scale(<number> o, <number> delta)
table n = tex.scale(table o, <number> delta)
```

Multiplies the Lua numbers `o` and `delta`, and returns a rounded number that is in the range of a valid T_EX register value. In the table version, it creates a copy of the table with all numeric top-level values scaled in that manner. If the multiplied number(s) are of range, it generates ‘number to big’ error(s) as well.

4.1.13.3 `tex.definefont`

```
tex.definefont(<string> csname, <number> fontid)
tex.definefont(<boolean> global, <string> csname, <number> fontid)
```

Associates `csname` with the internal font number `fontid`. The definition is global if (and only if) `global` is specified and true (the setting of `globaldefs` is not taken into account).

4.2 The token library

The `token` table contains interface functions to T_EX’s handling of tokens. These functions are most useful when combined with the `token_filter` callback, but they could be used standalone as well.

A token is represented in Lua as a small table. For the moment, this table consists of three numeric entries:

| index | meaning | description |
|-------|---------------------|---|
| 1 | command code | this is a value between 0 and 130 (approximately) |
| 2 | command modifier | this is a value between 0 and 2 ²¹ |
| 3 | control sequence id | for commands that are not the result of control sequences, like letters and characters, it is zero, otherwise, it is a number pointing into the ‘equivalence table’ |



4.2.1 `token.get_next`

```
token t = token.get_next()
```

This fetches the next input token from the current input source, without expansion.

4.2.2 `token.is_expandable`

```
<boolean> b = token.is_expandable(token t)
```

This tests if the token `t` could be expanded.

4.2.3 `token.expand`

```
token.expand()
```

If a token is expandable, this will expand one level of it, so that the first token of the expansion will now be the next token to be read by `token.get_next()`.

4.2.4 `token.is_activechar`

```
<boolean> b = token.is_activechar(token t)
```

This is a special test that is sometimes handy. Discovering whether some control sequence is the result of an active character turned out to be very hard otherwise.

4.2.5 `token.create`

```
token t = token.create(<string> csname)
token t = token.create(<number> charcode)
token t = token.create(<number> charcode, <number> catcode)
```

This is the token factory. If you feed it a string, then it is the name of a control sequence (without leading backslash), and it will be looked up in the equivalence table.

If you feed it number, then this is assumed to be an input character, and an optional second number gives its category code. This means it is possible to overrule a character's category code, with a few exceptions: the category codes 0 (escape), 9 (ignored), 13 (active), 14 (comment), and 15 (invalid) cannot occur inside a token. The values 0, 9, 14 and 15 are therefore illegal as input to `token.create()`, and active characters will be resolved immediately.

Note: unknown string sequences and never defined active characters will result in a token representing an 'undefined control sequence' with a near-random name. It is *not* possible to define brand new control sequences using `token.create`!



4.2.6 token.command_name

```
<string> commandname = token.command_name(<token> t)
```

This returns the name associated with the ‘command’ value of the token in LuaTeX. There is not always a direct connection between these names and primitives. For instance, all `\ifxxx` tests are grouped under `if_fest`, and the ‘command modifier’ defines which test is to be run.

4.2.7 token.command_id

```
<number> i = token.command_id(<string> commandname)
```

This returns a number that is the inverse operation of the previous command, to be used as the first item in a token table.

4.2.8 token.csname_name

```
<string> csname = token.csname_name(<token> t)
```

This returns the name associated with the ‘equivalence table’ value of the token in LuaTeX. It returns the string value of the command used to create the current token, or an empty string if there is no associated control sequence.

Keep in mind that there are potentially two control sequences that return the same csname string: single character control sequences and active characters have the same ‘name’.

4.2.9 token.csname_id

```
<number> i = token.csname_id(<string> csname)
```

This returns a number that is the inverse operation of the previous command, to be used as the third item in a token table.

4.3 The node library

The `node` library contains functions that facilitate dealing with (lists of) nodes and their values. They allow you to create, alter, copy, delete, and insert LuaTeX node objects, the core objects within the typesetter.

LuaTeX nodes are represented in Lua as userdata with the metadata type `luatex.node`. The various parts within a node can be accessed using named fields.



Each node has at least the three fields `next`, `id`, and `subtype`:

- The `next` field returns the userdata object for the next node in a linked list of nodes, or nil, if there is no next node.
- The `id` indicates T_EX's 'node type'. The field `id` has a numeric value for efficiency reasons, but some of the library functions also accept a string value instead of `id`.
- The `subtype` is another number. It often gives further information about a node of a particular `id`, but it is most important when dealing with 'whatsits', because they are differentiated solely based on their `subtype`.

The other available fields depend on the `id` (and for 'whatsits', the `subtype`) of the node. Further details on the various fields and their meanings are given in **chapter 7**.

T_EX's math nodes are not yet supported: there is not yet an interface to the internals of the math list and it is not possible to create them from Lua. Support for `unset` (alignment) nodes is partial: they can be queried and modified from Lua code, but not created.

Nodes can be compared to each other, but: you are actually comparing indices into the node memory. This means that equality tests can only be trusted under very limited conditions. It will not work correctly in any situation where one of the two nodes has been freed and/or reallocated: in that case, there will be false positives.

At the moment, memory management of nodes should still be done explicitly by the user. Nodes are not 'seen' by the Lua garbage collector, so you have to call the node freeing functions yourself when you are no longer in need of a node (list). Nodes form linked lists without reference counting, so you have to be careful that when control returns back to LuaT_EX itself, you have not deleted nodes that are still referenced from a `next` pointer elsewhere, and that you did not create nodes that are referenced more than once.

There are statistics available with regards to the allocated node memory, which can be handy for tracing.

4.3.1 Node handling functions

4.3.1.1 `node.types`

```
table t = node.types()
```

This function returns an array that maps node id numbers to node type strings, providing an overview of the possible top-level `id` types.

4.3.1.2 `node.whatsits`

```
table t = node.whatsits()
```

T_EX's 'whatsits' all have the same `id`. The various subtypes are defined by their `subtype`. The function is much like `node.types`, except that it provides an array of `subtype` mappings.



4.3.1.3 node.id

```
<number> id = node.id(<string> type)
```

This converts a single type name to its internal numeric representation.

4.3.1.4 node.subtype

```
<number> subtype = node.subtype(<string> type)
```

This converts a single whatsit name to its internal numeric representation ([subtype](#)).

4.3.1.5 node.type

```
<string> type = node.type(<number> id)
```

This converts a internal numeric representation to an external string representation.

4.3.1.6 node.fields

```
table t = node.fields(<number> id)
table t = node.fields(<number> id, <number> subtype)
```

This function returns an array of valid field names for a particular type of node. If you want to get the valid fields for a ‘whatsit’, you have to supply the second argument also. In other cases, any given second argument will be silently ignored.

This function accepts string [id](#) and [subtype](#) values as well.

4.3.1.7 node.has_field

```
<boolean> t = node.has_field(<node> n, <string> field)
```

This function returns a boolean that is only true if [n](#) is actually a node, and it has the field.

4.3.1.8 node.new

```
<node> n = node.new(<number> id)
<node> n = node.new(<number> id, <number> subtype)
```

Creates a new node. All of the new node’s fields are initialized to either zero or nil except for [id](#) and [subtype](#) (if supplied). If you want to create a new whatsit, then the second argument is required, otherwise it need not be present. As with all node functions, this function creates a node on the T_EX level.



This function accepts string `id` and `subtype` values as well.

4.3.1.9 `node.free`

```
node.free(<node> n)
```

Removes the node `n` from T_EX's memory. Be careful: no checks are done on whether this node is still pointed to from a register or some `next` field: it is up to you to make sure that the internal data structures remain correct.

4.3.1.10 `node.flush_list`

```
node.flush_list(<node> n)
```

Removes the node list `n` and the complete node list following `n` from T_EX's memory. Be careful: no checks are done on whether any of these nodes is still pointed to from a register or some `next` field: it is up to you to make sure that the internal data structures remain correct.

4.3.1.11 `node.copy`

```
<node> m = node.copy(<node> n)
```

Creates a deep copy of node `n`, including all nested lists as in the case of a hlist or vlist node. Only the `next` field is not copied.

4.3.1.12 `node.copy_list`

```
<node> m = node.copy_list(<node> n)
```

Creates a deep copy of the node list that starts at `n`.

4.3.1.13 `node.hpack`

```
<node> h = node.hpack(<node> n)
<node> h = node.hpack(<node> n, <number> w, <string> info)
```

This function creates a new hlist by packaging the list that begins at node `n` into a horizontal box. With only a single argument, this box is created using the natural width of its components. In the three argument form, `info` must be either `additional` or `exactly`, and `w` is the additional (`\hbox spread`) or exact (`\hbox to`) width to be used.

Caveat: at this moment, there can be unexpected side-effects to this function, like updating some of the `\marks` and `\inserts`. Also note that the content of `h` is the original node list `n`: if you call



`node.free(h)` you will also free the node list itself, unless you explicitly set the `list` field to `nil` beforehand. And in a similar way, calling `node.free(n)` will invalidate `h` as well!

4.3.1.14 `node.mlist_to_hlist`

```
<node> h = node.mlist_to_hlist(<node> n,  
                               <string> displaytype, <boolean> penalties)
```

This runs the internal mlist to hlist conversion, converting the math list in `n` into the horizontal list `h`. The interface is exactly the same as for the callback `mlist_to_hlist`.)

4.3.1.15 `node.slide`

```
<node> m = node.slide(<node> n)
```

Returns the last node of the node list that starts at `n`. As a side-effect, it also creates a reverse chain of `prev` pointers between nodes.

4.3.1.16 `node.length`

```
<number> i = node.length(<node> n)  
<number> i = node.length(<node> n, <node> m)
```

Returns the number of nodes contained in the node list that starts at `n`. If `m` is also supplied it stops at `m` instead of at the end of the list. The node `m` is not counted.

4.3.1.17 `node.count`

```
<number> i = node.count(<number> id, <node> n)  
<number> i = node.count(<number> id, <node> n, <node> m)
```

Returns the number of nodes contained in the node list that starts at `n` that have an matching `id` field. If `m` is also supplied, counting stops at `m` instead of at the end of the list. The node `m` is not counted.

This function also accept string `id`'s.

4.3.1.18 `node.traverse`

```
<node> t = node.traverse(<node> n)
```

This is an iterator that loops over the node list that starts at `n`.



4.3.1.19 node.traverse_id

```
<node> t = node.traverse_id(<number> id, <node> n)
```

This is an iterator that loops over all the nodes in the list that starts at `n` that have a matching `id` field.

4.3.1.20 node.remove

```
<node> head, current = node.remove(<node> head, <node> current)
```

This function removes the node `current` from the list following `head`. It is your responsibility to make sure it is really part of that list. The return values are the new `head` and `current` nodes. The returned `current` is the node in the calling argument, and is only passed back as a convenience (its `next` field will be cleared). The returned `head` is more important, because if the function is called with `current` equal to `head`, it will be changed.

4.3.1.21 node.insert_before

```
<node> head, new = node.insert_before(<node> head, <node> current, <node> new)
```

This function inserts the node `new` before `current` into the list following `head`. It is your responsibility to make sure that `current` is really part of that list. The return values are the (potentially mutated) `head` and the `new`, set up to be part of the list (with correct `next` field). If `head` is initially `nil`, it will become `new`.

4.3.1.22 node.insert_after

```
<node> head, new = node.insert_after(<node> head, <node> current, <node> new)
```

This function inserts the node `new` after `current` into the list following `head`. It is your responsibility to make sure that `current` is really part of that list. The return values are the `head` and the `new`, set up to be part of the list (with correct `next` field). If `head` is initially `nil`, it will become `new`.

4.3.1.23 node.first_character

```
<node> n = node.first_character(<node> n)
<node> n = node.first_character(<node> n, <node> m)
```

Returns the first node that is a glyph node with a subtype indicating it is a character, or `nil`.



4.3.1.24 node.ligaturing

```
<node> h, <node> t, <boolean> success = node.ligaturing(<node> n)
<node> h, <node> t, <boolean> success = node.ligaturing(<node> n, <node> m)
```

Apply T_EX-style ligaturing to the specified nodelist. The tail node `m` is optional. The two returned nodes `h` and `t` are the new head and tail (both `n` and `m` can change into a new ligature).

4.3.1.25 node.kerning

```
<node> h, <node> t, <boolean> success = node.kerning(<node> n)
<node> h, <node> t, <boolean> success = node.kerning(<node> n, <node> m)
```

Apply T_EX-style kerning to the specified nodelist. The tail node `m` is optional. The two returned nodes `h` and `t` are the head and tail (either one of these can be an inserted kern node, because special kernings with word boundaries are possible).

4.3.1.26 node.unprotect_glyphs

```
node.unprotect_glyphs(<node> n)
```

Subtracts 256 from all glyph node subtypes. This and the next function are helpers to convert from `characters` to `glyphs` during node processing.

4.3.1.27 node.protect_glyphs

```
node.protect_glyphs(<node> n)
```

Adds 256 to all glyph node subtypes in the node list starting at `n`, except that if the value is 1, it adds only 255. The special handling of 1 means that `characters` will become `glyphs` after subtraction of 256.

4.3.1.28 node.last_node

```
<node> n = node.last_node()
```

This function pops the last node from T_EX's 'current list'. It returns that node, or `nil` if the current list is empty.

4.3.1.29 node.write

```
node.write(<node> n)
```



This is an experimental function that will append a node list to T_EX's ‘current list’. There is no error checking yet!

4.3.2 Attribute handling

Attributes appear as linked list of userdata objects in the `attr` field of individual nodes. They can be handled individually, but it is much safer and more efficient to use the dedicated functions associated with them.

4.3.2.1 `node.has_attribute`

```
<number> v = node.has_attribute(<node> n, <number> id)
<number> v = node.has_attribute(<node> n, <number> id, <number> val)
```

Tests if a node has the attribute with number `id` set. If `val` is also supplied, also tests if the value matches `val`. It returns the value, or, if no match is found, `nil`.

4.3.2.2 `node.set_attribute`

```
node.set_attribute(<node> n, <number> id, <number> val)
```

Sets the attribute with number `id` to the value `val`. Duplicate assignments are ignored. *[needs explanation]*

4.3.2.3 `node.unset_attribute`

```
<number> v = node.unset_attribute(<node> n, <number> id, <number> val)
<number> v = node.unset_attribute(<node> n, <number> id)
```

Unsets the attribute with number `id`. If `val` is also supplied, it will only perform this operation if the value matches `val`. Missing attributes or attribute-value pairs are ignored.

If the attribute was actually deleted, returns its old value. Otherwise, returns `nil`.

4.4 The `texio` library

This library takes care of the low-level I/O interface.



4.4.1 Printing functions

4.4.1.1 `texio.write`

```
texio.write(<string> target, <string> s, ...)  
texio.write(<string> s, ...)
```

Without the `target` argument, writes all given strings to the same location(s) \TeX writes messages to at this moment. If `\batchmode` is in effect, it writes only to the log, otherwise it writes to the log and the terminal.

The optional `target` can be one of three possibilities: `term`, `log` or `term and log`.

Note: If several strings are given, and if the first of these strings is or might be one of the targets above, the `target` must be specified explicitly to prevent Lua from interpreting the first string as the target.

4.4.1.2 `texio.write_nl`

```
texio.write_nl(<string> target, <string> s, ...)  
texio.write_nl(<string> s, ...)
```

This function behaves like `texio.write`, but make sure that the given strings will appear at the beginning of a new line. You can pass a single empty string if you only want to move to the next line.

4.5 The pdf library

This table contains the current `h` and `v` values that define the location on the output page. The values can be queried and set using scaled points as units.

```
pdf.v  
pdf.h
```

The associated function calls are

```
pdf.setv(<number> n)  
<number> n = pdf.getv()  
pdf.seth(<number> n)  
<number> n = pdf.geth()
```

It also holds a print function to write stuff to the pdf document that can be used from within a `\lualatex` argument. This function is not to be used inside `\directlua` unless you know *exactly* what you are doing.



`pdf.print`

```
pdf.print(<string> s)
pdf.print(<string> type, <string> s)
```

The optional parameter can be used to mimic the behavior of `\pdfliteral`: the `type` is `direct` or `page`.

4.6 The `img` library

The `img` library can be used as an alternative to `\pdfximage` and `\pdfrefximage`, and the associated ‘satellite’ commands like `\pdfximagebbox`. Image objects can also be used within virtual fonts via the `image` command listed in [section 6.2](#).

`img.new`

```
<image> var = img.new()
<image> var = img.new(image_spec)
```

This function creates a userdata object of type ‘image’. The `image_spec` argument is optional. If it is given, it must be a table, and that table must contain a `filename` key. A number of other keys can also be useful, these are explained below.

You can either say

```
a=img.new()
```

followed by

```
a.filename="foo.png"
```

or you can put the file name (and some or all of the other keys) into a table directly, like so:

```
a=img.new{filename='foo.pdf',page=1}
```

The generated `<image>` userdata object allows access to a set of user-specified values as well as a set of values that are normally filled in and updated automatically by LuaT_EX itself. Some of those are derived from the actual image file, others are updated to reflect the pdf output status of the object.

There is one required user-specified field: the file name (`filename`). It can optionally be augmented by the requested image dimensions (`width`, `depth`, `height`), user-specified image attributes (`attr`), the requested pdf page identifier (`page`), the requested boundingbox (`pagebox`) for pdf inclusion, the requested color space object (`colorspace`).

The function `img.new` does not access the actual image file, it just creates the `<image>` userdata object and initializes some memory structures. The `<image>` object and its internal structures are automatically garbage collected.



Once the image is scanned, all the values in the `<image>` become frozen, and you cannot change them any more.

`img.keys`

```
<table> keys = img.keys()
```

This function returns a list of all the possible `image_spec` keys, both user-supplied and automatic ones.

| field name | type | description |
|------------|--------|---|
| depth | number | the image depth for LuaTeX (in scaled points) |
| height | number | the image height for LuaTeX (in scaled points) |
| width | number | the image width for LuaTeX (in scaled points) |
| transform | number | the image transform, integer number 0..7 |
| attr | string | the image attributes for LuaTeX |
| filename | string | the image file name |
| stream | string | the raw stream data for an <code>/XObject /Form</code> object |
| page | ?? | the identifier for the requested image page (type is number or string, default is the number 1) |
| pagebox | string | the requested bounding box, one of <code>none</code> , <code>media</code> , <code>crop</code> , <code>bleed</code> , <code>trim</code> , <code>art</code> |
| bbox | table | table with 4 boundingbox dimensions <code>llx</code> , <code>lly</code> , <code>urx</code> , and <code>ury</code> overruling the <code>pagebox</code> entry |
| filepath | string | the full (expanded) file name of the image |
| colordepth | number | the number of bits used by the color space |
| colorspace | number | the color space object number |
| imagetype | string | one of <code>pdf</code> , <code>png</code> , <code>jpg</code> , <code>jbig2</code> , or <code>nil</code> |
| objnum | number | the pdf image object number |
| index | number | the pdf image name suffix |
| pages | number | the total number of available pages |
| xsize | number | the natural image width |
| ysize | number | the natural image height |
| xres | number | the horizontal natural image resolution (in dpi) |
| gres | number | the vertical natural image resolution (in dpi) |

A running (undefined) dimension in `width`, `height`, or `depth` is represented as `nil` in Lua, so if you want to load an image at its ‘natural’ size, you do not have to specify any of those three fields.

The `stream` parameter allows to fabricate an `/XObject /Form` object from a string giving the stream contents, e.g., for a filled rectangle:

```
a.stream = "0 0 20 10 re f"
```

When writing the image, an `/XObject /Form` object is created, like with embedded pdf file writing. The object is written out only once. The `stream` key requires that also the `bbox` table is given. The `stream` key conflicts with the `filename` key. The `transform` key works as usual also with `stream`.



The `bbox` key needs a table with four boundingbox values, e.g.:

```
a.bbox = {"30bp", 0, "225bp", "200bp"}
```

This replaces and overrules any given `pagebox` value; with given `bbox` the box dimensions coming with an embedded pdf file are ignored. The `xsize` and `ysize` dimensions are set accordingly, when the image is scaled. The `bbox` parameter is ignored for non-pdf images.

The `transform` allows to mirror and rotate the image in steps of 90 deg. The default value 0 gives an unmirrored, unrotated image. Values 1–3 give counterclockwise rotation by 90, 180, or 270 degrees, whereas with values 4–7 the image is first mirrored and then rotated counterclockwise by 90, 180, or 270 degrees. The `transform` operation gives the same visual result as if you would externally preprocess the image by a graphics tool and then use it by LuaTeX. If a pdf file to be embedded already contains a `/Rotate` specification, the rotation result is the combination of the `/Rotate` rotation followed by the `transform` operation.

`img.scan`

```
<image> var = img.scan(<image> var)
<image> var = img.scan(image_spec)
```

When you say `img.scan(a)` for a new image, the file is scanned, and variables such as `xsize`, `ysize`, image `type`, number of `pages`, and the resolution are extracted. Each of the `width`, `height`, `depth` fields are set up according to the image dimensions, if they were not given an explicit value already. An image file will never be scanned more than once for a given image variable. With all subsequent `img.scan(a)` calls only the dimensions are again set up (if they have been changed by the user in the meantime).

For ease of use, you can do right-away a

```
<image> a = img.scan { filename = "foo.png" }
```

without a prior `img.new`.

Nothing is written yet at this point, so you can do `a=img.scan`, retrieve the available info like image width and height, and then throw away `a` again by saying `a=nil`. In that case no image object will be reserved in the PDF, and the used memory will be cleaned up automatically.

`img.copy`

```
<image> var = img.copy(<image> var)
<image> var = img.copy(image_spec)
```

If you say `a = b`, then both variables point to the same `<image>` object. if you want to write out an image with different sizes, you can do a `b=img.copy(a)`.



Afterwards, `a` and `b` still reference the same actual image dictionary, but the dimensions for `b` can now be changed from their initial values that were just copies from `a`.

`img.write`

```
<image> var = img.write(<image> var)
<image> var = img.write(image_spec)
```

By `img.write(a)` a pdf object number is allocated, and a whatsit node of subtype `pdf_refximage` is generated and put into the output list. By this the image `a` is placed into the page stream, and the image file is written out into an image stream object after the shipping of the current page is finished.

Again you can do a terse call like

```
img.write { filename = "foo.png" }
```

The `<image>` variable is returned in case you want it for later processing.

`img.immediatewrite`

```
<image> var = img.immediatewrite(<image> var)
<image> var = img.immediatewrite(image_spec)
```

By `img.immediatewrite(a)` a pdf object number is allocated, and the image file for image `a` is written out immediately into the pdf file as an image stream object (like with `\immediate\pdfximage`). The object number of the image stream dictionary is then available by the `objnum` key. No `pdf_refximage` whatsit node is generated. You will need a `img.write(a)` or `img.node(a)` call to let the image appear on the page, or reference it by another trick; else you will have a dangling image object in the pdf file.

Also here you can do a terse call like

```
a = img.immediatewrite { filename = "foo.png" }
```

The `<image>` variable is returned and you will most likely need it.

`img.node`

```
<node> n = img.node(<image> var)
<node> n = img.node(image_spec)
```

This function allocates a pdf object number and returns a whatsit node of subtype `pdf_refximage`, filled with the image parameters `width`, `height`, `depth`, and `objnum`. Also here you can do a terse call like:

```
n = img.node { filename = "foo.png" }
```



This example outputs an image:

```
node.write(img.node{filename="foo.png"})
```

`img.types`

```
<table> types = img.types()
```

This function returns a list with the supported image file type names, currently these are `pdf`, `png`, `jpg`, and `jbig2`.

`img.bboxes`

```
<table> bboxes = img.bboxes()
```

This function returns a list with the supported pdf page box names, currently these are `media`, `crop`, `bleed`, `trim`, and `art` (all in lowercase letters).

4.7 The `mplib` library

The MetaPost library interface registers itself in the table `mplib`. It is based on MPlib version 1.102.

4.7.1 `mplib.new`

To create a new metapost instance, call

```
<mpinstance> mp = mplib.new({...})
```

This creates the `mp` instance object. The argument hash can have a number of different fields, as follows:

| name | type | description | default |
|--------------------------|---------|---|------------------------|
| <code>error_line</code> | number | error line width | 79 |
| <code>print_line</code> | number | line length in ps output | 100 |
| <code>main_memory</code> | number | total memory size | 5000 |
| <code>hash_size</code> | number | hash size | 16384 |
| <code>param_size</code> | number | max. active macro parameters | 150 |
| <code>max_in_open</code> | number | max. input file nestings | 10 |
| <code>random_seed</code> | number | the initial random seed | variable |
| <code>interaction</code> | string | the interaction mode, one of <code>batch</code> , <code>nonstop</code> , <code>scroll</code> , <code>errorstop</code> | <code>errorstop</code> |
| <code>ini_version</code> | boolean | the <code>-ini</code> switch | true |
| <code>mem_name</code> | string | <code>--mem</code> | <code>plain</code> |



| | | | |
|------------------------|----------|--------------------------|--------------------|
| <code>job_name</code> | string | <code>--jobname</code> | <code>mpout</code> |
| <code>find_file</code> | function | a function to find files | only local files |

The `find_file` function should be of this form:

```
<string> found = finder (<string> name, <string> mode, <string> type)
```

with:

`name` the requested file

`mode` the file mode: `r` or `w`

`type` the kind of file, one of: `mp`, `mem`, `tfm`, `map`, `pfb`, `enc`

Return either the full pathname of the found file, or `nil` if the file cannot be found.

4.7.2 `mp:statistics`

You can request statistics with:

```
<table> stats = mp:statistics()
```

This function returns the vital statistics for an MPlib instance. There are four fields, giving the maximum number of used items in each of the four statically allocated object classes:

| | | |
|--------------------------|--------|-------------------------------|
| <code>main_memory</code> | number | memory size |
| <code>hash_size</code> | number | hash size |
| <code>param_size</code> | number | simultaneous macro parameters |
| <code>max_in_open</code> | number | input file nesting levels |

4.7.3 `mp:execute`

You can ask the MetaPost interpreter to run a chunk of code by calling

```
local rettable = mp:execute('metapost language chunk')
```

for various bits of Metapost language input. Be sure to check the `rettable.status` (see below) because when a fatal MetaPost error occurs the MPlib instance will become unusable thereafter.

Generally speaking, it is best to keep your chunks small, but beware that all chunks have to obey proper syntax, like each of them is a small file. For instance, you cannot split a single statement over multiple chunks.

In contrast with the normal standalone `mpost` command, there is *no* implied ‘input’ at the start of the first chunk.



4.7.4 mp:finish

```
local rettable = mp:finish()
```

If for some reason you want to stop using an MPLib instance while processing is not yet actually done, you can call `mp:finish`. Eventually, used memory will be freed and open files will be closed by the Lua garbage collector, but an explicit `mp:finish` is the only way to capture the final part of the output streams.

4.7.5 Result table

The return value of `mp:execute` and `mp:finish` is a table with a few possible keys (only `status` is always guaranteed to be present).

| | | |
|--------|--------|--|
| log | string | output to the 'log' stream |
| term | string | output to the 'term' stream |
| error | string | output to the 'error' stream (only used for 'out of memory') |
| status | number | the return value: 0=good, 1=warning, 2=errors, 3=fatal error |
| fig | table | an array of generated figures (if any) |

When `status` equals 3, you should stop using this MPLib instance immediately, it is no longer capable of processing input.

If it is present, each of the entries in the `fig` array is a userdata representing a figure object, and each of those has a number of object methods you can call:

| | | |
|--------------|----------|--|
| boundingbox | function | returns the bounding box, as an array of 4 values |
| postscript | function | return a string that is the ps output of the <code>fig</code> |
| objects | function | returns the actual array of graphic objects in this <code>fig</code> |
| copy_objects | function | returns a deep copy of the array of graphic objects in this <code>fig</code> |
| filename | function | the filename this <code>fig</code> 's PostScript output would have written to in standalone mode |
| width | function | the <code>charwd</code> value |
| height | function | the <code>charht</code> value |
| depth | function | the <code>chardp</code> value |
| italcorr | function | the <code>charic</code> value |
| charcode | function | the (rounded) <code>charcode</code> value |

NOTE: you can call `fig:objects()` only once for any one `fig` object!

When the boundingbox represents a 'negated rectangle', i.e. when the first set of coordinates is larger than the second set, the picture is empty.

Graphical objects come in various types that each have a different list of accessible values. The types are: `fill`, `outline`, `text`, `start_clip`, `stop_clip`, `start_bounds`, `stop_bounds`, `special`.

There is helper function (`mplib.fields(obj)`) to get the list of accessible values for a particular object, but you can just as easily use the tables given below).



All graphical objects have a field `type` that gives the object type as a string value, that not explicit mentioned in the tables. In the following, `numbers` are PostScript points represented as a floating point number, unless stated otherwise. Field values that are of `table` are explained in the next section.

4.7.5.1 fill

| | | |
|-------------------------|---------------------|---|
| <code>path</code> | <code>table</code> | the list of knots |
| <code>htap</code> | <code>table</code> | the list of knots for the reversed trajectory |
| <code>pen</code> | <code>table</code> | knots of the pen |
| <code>color</code> | <code>table</code> | the object's color |
| <code>linejoin</code> | <code>number</code> | line join style (bare number) |
| <code>miterlimit</code> | <code>number</code> | miterlimit |
| <code>prescript</code> | <code>string</code> | the prescript text |
| <code>postscript</code> | <code>string</code> | the postscript text |

The entries `htap` and `pen` are optional.

There is helper function (`mplib.pen_info(obj)`) that returns a table containing a bunch of vital characteristics of the used pen (all values are floats):

| | | |
|--------------------|---------------------|------------------|
| <code>width</code> | <code>number</code> | width of the pen |
| <code>rx</code> | <code>number</code> | x scale |
| <code>sx</code> | <code>number</code> | xy multiplier |
| <code>sy</code> | <code>number</code> | yx multiplier |
| <code>ry</code> | <code>number</code> | y scale |
| <code>tx</code> | <code>number</code> | x offset |
| <code>ty</code> | <code>number</code> | y offset |

4.7.5.2 outline

| | | |
|-------------------------|---------------------|-------------------------------|
| <code>path</code> | <code>table</code> | the list of knots |
| <code>pen</code> | <code>table</code> | knots of the pen |
| <code>color</code> | <code>table</code> | the object's color |
| <code>linejoin</code> | <code>number</code> | line join style (bare number) |
| <code>miterlimit</code> | <code>number</code> | miterlimit |
| <code>linecap</code> | <code>number</code> | line cap style (bare number) |
| <code>dash</code> | <code>table</code> | representation of a dash list |
| <code>prescript</code> | <code>string</code> | the prescript text |
| <code>postscript</code> | <code>string</code> | the postscript text |

The entry `dash` is optional.



4.7.5.3 text

| | | |
|------------|--------|-----------------------|
| text | string | the text |
| font | string | font tfm name |
| dsize | number | font size |
| color | table | the object's color |
| width | number | |
| height | number | |
| depth | number | |
| transform | table | a text transformation |
| prescript | string | the prescript text |
| postscript | string | the postscript text |

4.7.5.4 special

prescript string special text

4.7.5.5 start_bounds, start_clip

path table the list of knots

4.7.5.6 stop_bounds, stop_clip

Here are no fields available.

4.7.6 Subsidiary table formats

4.7.6.1 Paths and pens

Paths and pens (that are really just a special type of paths as far as MPLib is concerned) are represented by an array where each entry is a table that represents a knot.

| | | |
|------------|--------|--|
| left_type | string | when present: 'endpoint', but ususally absent |
| right_type | string | like left_type |
| x_coord | number | X coordinate of this knot |
| y_coord | number | Y coordinate of this knot |
| left_x | number | X coordinate of the precontrol point of this knot |
| left_y | number | Y coordinate of the precontrol point of this knot |
| right_x | number | X coordinate of the postcontrol point of this knot |
| right_y | number | Y coordinate of the postcontrol point of this knot |



There is one special case: pens that are (possibly transformed) ellipses have an extra string-valued key `type` with value `elliptical` besides the array part containing the knot list.

4.7.6.2 Colors

A color is an integer array with 0, 1, 3 or 4 values:

| | | |
|---|--------------|--|
| 0 | marking only | no values |
| 1 | greyscale | one value in the range (0,1), 'black' is 0 |
| 3 | rgb | three values in the range (0,1), 'black' is 0,0,0 |
| 4 | cmyk | four values in the range (0,1), 'black' is 0,0,0,1 |

If the color model of the internal object was `unitialized`, then it was initialized to the values representing 'black' in the colorspace `defaultcolormodel` that was in effect at the time of the `shipout`.

4.7.6.3 Transforms

Each transform is a six-item array.

| | | |
|---|--------|---------------|
| 1 | number | represents x |
| 2 | number | represents y |
| 3 | number | represents xx |
| 4 | number | represents yx |
| 5 | number | represents xy |
| 6 | number | represents yy |

Note that the translation (index 1 and 2) comes first. This differs from the ordering in PostScript, where the translation comes last.

4.7.6.4 Dashes

Each `dash` is two-item hash, using the same model as PostScript for the representation of the dashlist. `dashes` is an array of 'on' and 'off', values, and `offset` is the phase of the pattern.

| | | |
|--------|--------|----------------------------|
| dashes | hash | an array of on-off numbers |
| offset | number | the starting offset value |

4.7.7 Character size information

These functions find the size of a glyph in a defined font. The `fontname` is the same name as the argument to `infont`; the `char` is a glyph id in the range 0 to 255; the returned `w` is in AFM units.



4.7.7.1 mp.char_width

```
<number> w = mp.char_width(<string> fontname, <number> char)
```

4.7.7.2 mp.char_height

```
<number> w = mp.char_height(<string> fontname, <number> char)
```

4.7.7.3 mp.char_depth

```
<number> w = mp.char_depth(<string> fontname, <number> char)
```

4.8 The callback library

This library has functions that register, find and list callbacks.

The `callback` library is only available in Lua state zero (0).

```
id, error = callback.register(<string> callback_name,function callback_func)
id, error = callback.register(<string> callback_name,nil)
```

where the `callback_name` is a predefined callback name, see below. The function returns the internal `id` of the callback or `nil`, if the callback could not be registered. In the latter case, `error` contains an error message, otherwise it is `nil`.

LuaTeX internalizes the callback function in such a way that it does not matter if you redefine a function accidentally.

Callback assignments are always global. You can use the special value `nil` instead of a function for clearing the callback.

Currently, callbacks are not dumped into the format file.

```
table info = callback.list()
```

The keys in the table are the known callback names, the value is a boolean where `true` means that the callback is currently set (active).

```
function f = callback.find(callback_name)
```

If the callback is not set, `callback.find` returns `nil`.



4.8.1 File discovery callbacks

4.8.1.1 find_read_file and find_write_file

Your callback function should have the following conventions:

```
<string> actual_name = function (number id_number, <string> asked_name)
```

Arguments:

id_number

This number is zero for the log or `\input` files. For \TeX 's `\read` or `\write` the number is incremented by one, so `\read0` becomes 1.

asked_name

This is the user-supplied filename, as found by `\input`, `\openin` or `\openout`.

Return value:

actual_name

This is the filename used. For the very first file that is read in by \TeX , you have to make sure you return an `actual_name` that has an extension and that is suitable for use as `jobname`. If you don't, you will have to manually fix the name of the log file and output file after \LaTeX is finished, and an eventual format filename will become mangled. That is because these file names depend on the `jobname`.

You have to return `nil` if the file cannot be found.

4.8.1.2 find_font_file

Your callback function should have the following conventions:

```
<string> actual_name = function (<string> asked_name)
```

The `asked_name` is an `otf` or `tfm` font metrics file.

Return `nil` if the file cannot be found.

4.8.1.3 find_output_file

Your callback function should have the following conventions:

```
<string> actual_name = function (<string> asked_name)
```

The `asked_name` is the `pdf` or `dvi` file for writing.



4.8.1.4 find_format_file

Your callback function should have the following conventions:

```
<string> actual_name = function (<string> asked_name)
```

The `asked_name` is a format file for reading (the format file for writing is always opened in the current directory).

4.8.1.5 find_vf_file

Like `find_font_file`, but for virtual fonts. This applies to both Aleph's ovf files and traditional Knuthian vf files.

4.8.1.6 find_ocp_file

Like `find_font_file`, but for ocp files.

4.8.1.7 find_map_file

Like `find_font_file`, but for map files.

4.8.1.8 find_enc_file

Like `find_font_file`, but for enc files.

4.8.1.9 find_sfd_file

Like `find_font_file`, but for subfont definition files.

4.8.1.10 find_pk_file

Like `find_font_file`, but for pk bitmap files. The argument `name` is a bit special in this case. Its form is

```
<base res>dpi/<fontname>.<actual res>pk
```

So you may be asked for `600dpi/manfnt.720pk`. It is up to you to find a 'reasonable' bitmap file to go with that specification.



4.8.1.11 find_data_file

Like `find_font_file`, but for embedded files (`\pdfobj file '...`).

4.8.1.12 find_opentype_file

Like `find_font_file`, but for OpenType font files.

4.8.1.13 find_truetype_file and find_type1_file

Your callback function should have the following conventions:

```
<string> actual_name = function (<string> asked_name)
```

The `asked_name` is a font file. This callback is called while LuaTeX is building its internal list of needed font files, so the actual timing may surprise you. Your return value is later fed back into the matching `read_file` callback.

Strangely enough, `find_type1_file` is also used for OpenType (otf) fonts.

4.8.1.14 find_image_file

Your callback function should have the following conventions:

```
<string> actual_name = function (<string> asked_name)
```

The `asked_name` is an image file. Your return value is used to open a file from the harddisk, so make sure you return something that is considered the name of a valid file by your operating system.

4.8.2 File reading callbacks

4.8.2.1 open_read_file

Your callback function should have the following conventions:

```
<table> env = function (<string> file_name)
```

Argument:

`file_name`

The filename returned by a previous `find_read_file` or the return value of `kpse.find_file()` if there was no such callback defined.

Return value:



env

This is a table containing at least one required and one optional callback function for this file. The required field is `reader` and the associated function will be called once for each new line to be read, the optional one is `close` that will be called once when LuaTeX is done with the file.

LuaTeX never looks at the rest of the table, so you can use it to store your private per-file data. Both the callback functions will receive the table as their only argument.

4.8.2.1.1 reader

LuaTeX will run this function whenever it needs a new input line from the file.

```
function(<table> env)
    return <string> line
end
```

Your function should return either a string or `nil`. The value `nil` signals that the end of file has occurred, and will make TeX call the optional `close` function next.

4.8.2.1.2 close

LuaTeX will run this optional function when it decides to close the file.

```
function(<table> env)
    return
end
```

Your function should not return any value.

4.8.2.2 General file readers

There is a set of callbacks for the loading of binary data files. These all use the same interface:

```
function(<string> name)
    return <boolean> success, <string> data, <number> data_size
end
```

The `name` will normally be a full path name as it is returned by either one of the file discovery callbacks or the internal version of `kpse.find_file()`.

success

Return false when a fatal error occurred (e.g. when the file cannot be found, after all).

data

The bytes comprising the file.

data_size

The length of the `data`, in bytes.



Return an empty string and zero if the file was found but there was a reading problem.

The list of functions is as follows:

| | |
|---------------------------------|--|
| <code>read_font_file</code> | ofm or tfm files |
| <code>read_vf_file</code> | virtual fonts |
| <code>read_ocp_file</code> | ocp files |
| <code>read_map_file</code> | map files |
| <code>read_enc_file</code> | encoding files |
| <code>read_sfd_file</code> | subfont definition files |
| <code>read_pk_file</code> | pk bitmap files |
| <code>read_data_file</code> | embedded files (<code>\pdfobj file ...</code>) |
| <code>read_truetype_file</code> | TrueType font files |
| <code>read_type1_file</code> | Type1 font files |
| <code>read_opentype_file</code> | OpenType font files |

4.8.3 Data processing callbacks

4.8.3.1 `process_input_buffer`

This callback allows you to change the contents of the line input buffer just before LuaTeX actually starts looking at it.

```
function(<string> buffer)
    return <string> adjusted_buffer
end
```

If you return `nil`, LuaTeX will pretend like your callback never happened. You can gain a small amount of processing time from that.

4.8.3.2 `token_filter`

This callback allows you to replace the way LuaTeX fetches lexical tokens.

```
function()
    return <table> token
end
```

The calling convention for this callback is a bit more complicated than for most other callbacks. The function should either return a Lua table representing a valid to-be-processed token or tokenlist, or something else like `nil` or an empty table.

If your Lua function does not return a table representing a valid token, it will be immediately called again, until it eventually does return a useful token or tokenlist (or until you reset the callback value to `nil`). See the description of `token` for some handy functions to be used in conjunction with this callback.



If your function returns a single usable token, then that token will be processed by LuaT_EX immediately. If the function returns a token list (a table consisting of a list of consecutive token tables), then that list will be pushed to the input stack at a completely new token list level, with its token type set to ‘inserted’. In either case, the returned token(s) will not be fed back into the callback function.

4.8.4 Node list processing callbacks

The description of nodes and node lists is in [chapter 7](#).

4.8.4.1 `buildpage_filter`

This callback is called whenever LuaT_EX is ready to move stuff to the main vertical list. You can use this callback to do specialized manipulation of the page building stage like imposition or column balancing.

```
function(<string> extrainfo)
end
```

The string `extrainfo` gives some additional information about what T_EX’s state is with respect to the ‘current page’. The possible values are:

| value | explanation |
|-----------------------------|--|
| <code>alignment</code> | a (partial) alignment is being added |
| <code>after_output</code> | an output routine has just finished |
| <code>box</code> | a typeset box is being added |
| <code>new_graf</code> | the beginning of a new paragraph |
| <code>vmode_par</code> | <code>\par</code> was found in vertical mode |
| <code>hmode_par</code> | <code>\par</code> was found in horizontal mode |
| <code>insert</code> | an insert is added |
| <code>penalty</code> | a penalty (in vertical mode) |
| <code>before_display</code> | immediately before a display starts |
| <code>after_display</code> | a display is finished |
| <code>end</code> | LuaT _E X is terminating (it’s all over) |

4.8.4.2 `pre_linebreak_filter`

This callback is called just before LuaT_EX starts converting a list of nodes into a stack of `\hboxes`. The removal of a possible final skip and the subsequent insertion of `\parfillskip` has not happened yet at that moment.

```
function(<node> head, <string> groupcode)
    return true | false | <node> newhead
end
```



The string called `groupcode` identifies the nodelist's context within T_EX's processing. The range of possibilities is given in the table below, but not all of those can actually appear in `pre_linebreak_filter`, some are for the `hpack_filter` and `vpack_filter` callbacks that will be explained in the next two paragraphs.

| value | explanation |
|---------------|---------------------------------|
| <empty> | main vertical list |
| hbox | \hbox in horizontal mode |
| adjusted_hbox | \hbox in vertical mode |
| vbox | \vbox |
| vtop | \vtop |
| align | \halign or \valign |
| disc | discretionaries |
| insert | packaging an insert |
| vcenter | \vcenter |
| local_box | \localleftbox or \localrightbox |
| split_off | top of a \vsplit |
| split_keep | remainder of a \vsplit |
| align_set | alignment cell |
| fin_row | alignment row |

4.8.4.3 linebreak_filter

This callback replaces LuaT_EX's line breaking algorithm.

```
function(<node> head, <boolean> is_display)
    return <node> newhead
end
```

The returned node is the head of the list that will be added to the main vertical list, the boolean argument is true if this paragraph is interrupted by a following math display.

If you return something that is not a `<node>`, LuaT_EX will apply the internal linebreak algorithm on the list that starts at `<head>`.

4.8.4.4 post_linebreak_filter

This callback is called just after LuaT_EX has converted a list of nodes into a stack of \hboxes.

```
function(<node> head, <string> groupcode)
    return true | false | <node> newhead
end
```



4.8.4.5 hpack_filter

This callback is called when T_EX is ready to start boxing some horizontal mode material. Math items and line boxes are ignored at the moment.

```
function(<node> head, <string> groupcode, <number> size, <string> packtype)
    return true | false | <node> newhead
end
```

The `packtype` is either `additional` or `exactly`. If `additional`, then the `size` is a `\hbox spread ...` argument. If `exactly`, then the `size` is a `\hbox to` In both cases, the number is in scaled points.

4.8.4.6 vpack_filter

This callback is called when T_EX is ready to start boxing some vertical mode material. Math displays are ignored at the moment.

This function is very similar to the `hpack_filter`. Besides the fact that it is called at different moments, there is an extra variable that matches T_EX's `\maxdepth` setting.

```
function(<node> head, <string> groupcode, <number> size, <string> packtype,
<number> maxdepth)
    return true | false | <node> newhead
end
```

4.8.4.7 pre_output_filter

This callback is called when T_EX is ready to start boxing the box 255 for `\output`.

```
function(<node> head, <string> groupcode, <number> size, <string> packtype,
<number> maxdepth)
    return true | false | <node> newhead
end
```

4.8.4.8 hyphenate

```
function(<node> head, <node> tail)
end
```

No return values. This callback has to insert discretionary nodes in the node list it receives.

4.8.4.9 ligaturing




```
function(<node> head, <node> tail)
end
```

No return values. This callback has to apply ligaturing to the node list it receives.

You don't have to worry about return values because the `head` node that is passed on to the callback is guaranteed not to be a `glyph_node` (if need be, a temporary node will be prepended), and therefore it cannot be affected by the mutations that take place. After the callback, the internal value of the 'tail of the list' will be recalculated.

The `next` of `head` is guaranteed to be non-nil.

The `next` of `tail` is guaranteed be nil, and therefore the second callback argument can often be ignored. It is provided for orthogonality, and because it can sometimes be handy when special processing has to take place.

4.8.4.10 kerning

```
function(<node> head, <node> tail) end
```

No return values. This callback has to apply kerning between the nodes in the node list it receives. See [ligaturing](#) for calling conventions.

4.8.4.11 mlist_to_hlist

This callback replaces LuaT_EX's math list to node list conversion algorithm.

```
function(<node> head, <string> displaytype, <boolean> need_penalties)
    return <node> newhead
end
```

The returned node is the head of the list that will be added to the vertical or horizontal list, the string argument is either 'text' or 'display' depending on the current math mode, the boolean argument is `true` if penalties have to be inserted in this list, `false` otherwise.

4.8.5 Information reporting callbacks

4.8.5.1 start_run

```
function()
```

Replaces the code that prints LuaT_EX's banner.



4.8.5.2 stop_run

```
function()
```

Replaces the code that prints LuaTeX's statistics and 'output written to' messages.

4.8.5.3 start_page_number

```
function()
```

Replaces the code that prints the [and the page number at the begin of \shipout. This callback will also override the printing of box information that normally takes place when \tracingoutput is positive.

4.8.5.4 stop_page_number

```
function()
```

Replaces the code that prints the] at the end of \shipout.

4.8.5.5 show_error_hook

```
function()  
    return  
end
```

This callback is run from inside the T_EX error function, and the idea is to allow you to do some extra reporting on top of what T_EX already does (none of the normal actions are removed). You may find some of the values in the [status](#) table useful.

message

is the formal error message T_EX has given to the user. (the line after the '!').

indicator

is either a filename (when it is a string) or a location indicator (a number) that can mean lots of different things like a token list id or a \read number.

lineno

is the current line number.

This is an investigative item for 'testing the water' only. The final goal is the total replacement of T_EX's error handling routines, but that needs lots of adjustments in the web source because T_EX deals with errors in a somewhat haphazard fashion. This is why the exact definition of [indicator](#) is not given here.



4.8.6 Font-related callbacks

4.8.6.1 define_font

```
function(<string> name, <number> size, <number> id) return <table> font end
```

The string `name` is the filename part of the font specification, as given by the user.

The number `size` is a bit special:

- if it is positive, it specifies an ‘at size’ in scaled points.
- if it is negative, its absolute value represents a ‘scaled’ setting relative to the designsize of the font.

The internal structure of the `font` table that is to be returned is explained in **chapter 6**. That table is saved internally, so you can put extra fields in the table for your later Lua code to use.

4.9 The lua library

This library contains two read-only items:

4.9.1 Variables

```
<number> n = lua.id
```

This returns the id number of the instance.

```
<string> s = lua.version
```

This returns a LuaTeX version identifier string. The value is currently `lua.version`, but it is soon to be replaced by something more elaborate.

4.9.2 LUA bytecode registers

Lua registers can be used to communicate Lua functions across Lua states. The accepted values for assignments are functions and `nil`. Likewise, the retrieved value is either a function or `nil`.

```
lua.bytecode[n] = function () .. end  
lua.bytecode[n]()
```

The contents of the `lua.bytecode` array is stored inside the format file as actual Lua bytecode, so it can also be used to preload Lua code.

Note: The function must not contain any upvalues. Currently, functions containing upvalues can be stored (and their upvalues are set to `nil`), but this is an artefact of the current Lua implementation and thus subject to change.



The associated function calls are

```
function f = lua.getbytecode(<number> n)
lua.setbytecode(<number> n, <function> f)
```

Note: Since a Lua file loaded using `loadfile(filename)` is essentially an anonymous function, a complete file can be stored in a bytecode register like this:

```
lua.bytecode[n] = loadfile(filename)
```

Now all definitions (functions, variables) contained in the file can be created by executing this bytecode register:

```
lua.bytecode[n]()
```

Note that the path of the file is stored in the Lua bytecode to be used in stack backtraces and therefore dumped into the format file if above code is used in `iniTEX`. If it contains private information, i.e. the user name, this information is then contained in the format file as well. This should be kept in mind when preloading files into a bytecode register in `iniTEX`.

4.10 The kpse library

This library provides an interface to the `kpathsea` file search method.

Before the search library can be used at all, its database has to be initialized. When `LuaTEX` is used to typeset documents, this happens automatically (that is, unless explicitly prohibited by the user's startup script. See [section 3.1](#) for more details). In `TEX Lua` mode, the initialization has to be done explicitly via the `kpse.set_program_name` function.

4.10.1 `kpse.set_program_name`

Sets the `kpathsea` executable (and optionally program) name.

```
kpse.set_program_name(<string> name)
kpse.set_program_name(<string> name, <string> proname)
```

The second argument controls the use of the 'dotted' values in the `texmf.cnf` configuration file, and defaults to the first argument.

4.10.2 `kpse.find_file`

The most often used function in the library is `find_file`:

```
<string> f = kpse.find_file(<string> filename)
<string> f = kpse.find_file(<string> filename, <string> ftype)
```



```

<string> f = kpse.find_file(<string> filename, <boolean> mustexist)
<string> f = kpse.find_file(<string> filename, <string> ftype, <boolean>
mustexist)
<string> f = kpse.find_file(<string> filename, <string> ftype, <number> dpi)

```

Arguments:

filename

the name of the file you want to find, with or without extension.

ftype

maps to the `-format` argument of `kpsewhich`. The supported `ftype` values are the same as the ones supported by the standalone `kpsewhich` program:

| | |
|--------------------|----------------------------|
| 'gf' | 'tex' |
| 'pk' | 'TeX system documentation' |
| 'bitmap font' | 'texpool' |
| 'tfm' | 'TeX system sources' |
| 'afm' | 'PostScript header' |
| 'base' | 'Troff fonts' |
| 'bib' | 'type1 fonts' |
| 'bst' | 'vf' |
| 'cnf' | 'dvips config' |
| 'ls-R' | 'ist' |
| 'fmt' | 'truetype fonts' |
| 'map' | 'type42 fonts' |
| 'mem' | 'web2c files' |
| 'mf' | 'other text files' |
| 'mfpool' | 'other binary files' |
| 'mft' | 'misc fonts' |
| 'mp' | 'web' |
| 'mppool' | 'cweb' |
| 'MetaPost support' | 'enc files' |
| 'ocp' | 'cmap files' |
| 'ofm' | 'subfont definition files' |
| 'opl' | 'opentype fonts' |
| 'otp' | 'pdftex config' |
| 'ovf' | 'lig files' |
| 'ovp' | 'texmfscripts' |
| 'graphic/figure' | |

The default type is `tex`.

mustexist

is similar to `kpsewhich`'s `-must-exist`, and the default is `false`. If you specify `true` (or a non-zero integer), then the `kpse` library will search the disk as well as the `ls-R` databases.

dpi

This is used for the size argument of the formats `pk`, `gf`, and `bitmap font`.



4.10.3 `kpse.init_prog`

Extra initialization for programs that need to generate bitmap fonts.

```
kpse.init_prog(<string> prefix, <number> base_dpi, <string> mfmode)
kpse.init_prog(<string> prefix, <number> base_dpi, <string> mfmode, <string>
fallback)
```

4.10.4 `kpse.readable_file`

Test if an (absolute) file name is a readable file

```
<string> f = kpse.readable_file(<string> name)
```

The return value is the actual absolute filename you should use, because the disk name is not always the same as the requested name, due to aliases and system-specific handling under e.g. msdos.

Returns `nil` if the file does not exist or is not readable.

4.10.5 `kpse.expand_path`

Like `kpsewhich`'s `-expand-path`:

```
<string> r = kpse.expand_path(<string> s)
```

4.10.6 `kpse.expand_var`

Like `kpsewhich`'s `-expand-var`:

```
<string> r = kpse.expand_var(<string> s)
```

4.10.7 `kpse.expand_braces`

Like `kpsewhich`'s `-expand-braces`:

```
<string> r = kpse.expand_braces(<string> s)
```

4.10.8 `kpse.show_path`

Like `kpsewhich`'s `-show-path`:

```
<string> r = kpse.show_path(<string> ftype)
```



4.10.9 kpse.var_value

Like kpsewhich's `-var-value`:

```
<string> r = kpse.var_value(<string> s)
```

4.11 The status library

This contains a number of run-time configuration items that you may find useful in message reporting, as well as an iterator function that gets all of the names and values as a table.

```
<table> info = status.list()
```

The keys in the table are the known items, the value is the current value. Almost all of the values in `status` are fetched through a metatable at run-time whenever they are accessed, so you cannot use `pairs` on `status`, but you *can* use `pairs` on `info`, of course. If you do not need the full list, you can also ask for a single item by using its name as an index into `status`.

The current list is:

| key | explanation |
|------------------|---|
| pdf_gone | written pdf bytes |
| pdf_ptr | not yet written pdf bytes |
| dvi_gone | written dvi bytes |
| dvi_ptr | not yet written dvi bytes |
| total_pages | number of written pages |
| output_file_name | name of the pdf or dvi file |
| log_name | name of the log file |
| banner | terminal display banner |
| var_used | variable (one-word) memory in use |
| dyn_used | token (multi-word) memory in use |
| str_ptr | number of strings |
| init_str_ptr | number of init _{TEX} strings |
| max_strings | maximum allowed strings |
| pool_ptr | string pool index |
| init_pool_ptr | init _{TEX} string pool index |
| pool_size | current size allocated for string characters |
| node_mem_usage | a string giving insight into currently used nodes |
| var_mem_max | number of allocated words for nodes |
| fix_mem_max | number of allocated words for tokens |
| fix_mem_end | maximum number of used tokens |
| cs_count | number of control sequences |
| hash_size | size of hash |
| hash_extra | extra allowed hash |
| font_ptr | number of active fonts |



| | |
|---------------------------------|--|
| <code>max_in_stack</code> | max used input stack entries |
| <code>max_nest_stack</code> | max used nesting stack entries |
| <code>max_param_stack</code> | max used parameter stack entries |
| <code>max_buf_stack</code> | max used buffer position |
| <code>max_save_stack</code> | max used save stack entries |
| <code>stack_size</code> | input stack size |
| <code>nest_size</code> | nesting stack size |
| <code>param_size</code> | parameter stack size |
| <code>buf_size</code> | current allocated size of the line buffer |
| <code>save_size</code> | save stack size |
| <code>obj_ptr</code> | max pdf object pointer |
| <code>obj_tab_size</code> | pdf object table size |
| <code>pdf_os_cntr</code> | max pdf object stream pointer |
| <code>pdf_os_objidx</code> | pdf object stream index |
| <code>pdf_dest_names_ptr</code> | max pdf destination pointer |
| <code>dest_names_size</code> | pdf destination table size |
| <code>pdf_mem_ptr</code> | max pdf memory used |
| <code>pdf_mem_size</code> | pdf memory size |
| <code>largest_used_mark</code> | max referenced marks class |
| <code>filename</code> | name of the current input file |
| <code>inputid</code> | numeric id of the current input |
| <code>linenumber</code> | location in the current input file |
| <code>lasterrorstring</code> | last error string |
| <code>luabytecodes</code> | number of active Lua bytecode registers |
| <code>luabytecode_bytes</code> | number of bytes in Lua bytecode registers |
| <code>luastates</code> | number of active Lua interpreters |
| <code>luastate_bytes</code> | number of bytes in use by Lua interpreters |
| <code>output_active</code> | <code>true</code> if the <code>\output</code> routine is active |
| <code>callbacks</code> | total number of executed callbacks so far |
| <code>indirect_callbacks</code> | number of those that were themselves a result of other callbacks (e.g. file readers) |

4.12 The `texconfig` table

This is a table that is created empty. A startup Lua script could fill this table with a number of settings that are read out by the executable after loading and executing the startup file.

| key | type | default | explanation |
|-------------------------------|---------|---------|--|
| <code>kpse_init</code> | boolean | true | <code>false</code> totally disables <code>kpathsea</code> initialisation, and enables interpretation of the following numeric key–value pairs. (only ever unset this if you implement <i>all</i> file find callbacks!) |
| <code>string_vacancies</code> | number | 75000 | cf. <code>web2c</code> docs |
| <code>pool_free</code> | number | 5000 | cf. <code>web2c</code> docs |
| <code>max_strings</code> | number | 15000 | cf. <code>web2c</code> docs |



| | | | |
|-------------------------------|---------|-------|--|
| <code>strings_free</code> | number | 100 | cf. web2c docs |
| <code>nest_size</code> | number | 50 | cf. web2c docs |
| <code>max_in_open</code> | number | 15 | cf. web2c docs |
| <code>param_size</code> | number | 60 | cf. web2c docs |
| <code>save_size</code> | number | 4000 | cf. web2c docs |
| <code>stack_size</code> | number | 300 | cf. web2c docs |
| <code>dvi_buf_size</code> | number | 16384 | cf. web2c docs |
| <code>error_line</code> | number | 79 | cf. web2c docs |
| <code>half_error_line</code> | number | 50 | cf. web2c docs |
| <code>max_print_line</code> | number | 79 | cf. web2c docs |
| <code>ocp_list_size</code> | number | 1000 | cf. web2c docs |
| <code>ocp_buf_size</code> | number | 1000 | cf. web2c docs |
| <code>ocp_stack_size</code> | number | 1000 | cf. web2c docs |
| <code>hash_extra</code> | number | 0 | cf. web2c docs |
| <code>pk_dpi</code> | number | 72 | cf. web2c docs |
| <code>trace_file_names</code> | boolean | true | <code>false</code> disables T _E X's normal file open-close feedback (the assumption is that callbacks will take care of that) |
| <code>file_line_error</code> | boolean | false | do <code>file:line</code> style error messages |
| <code>halt_on_error</code> | boolean | false | abort run on the first encountered error |
| <code>formatname</code> | string | | if no format name was given on the commandline, this key will be tested first instead of simply quitting |
| <code>jobname</code> | string | | if no input file name was given on the commandline, this key will be tested first instead of simply giving up |

Note: the numeric values that match web2c parameters are only used if `kpse_init` is explicitly set to `false`. In all other cases, the normal values from `texmf.cnf` are used.

4.13 The font library

The font library provides the interface into the internals of the font system, and also it contains helper functions to load traditional T_EX font metrics formats. Other font loading functionality is provided by the `fontforge` library that will be discussed in the next section.

4.13.1 Loading a TFM file

```
<table> fnt = font.read_tfm(<string> name, <number> s)
```

The number is a bit special:

- if it is positive, it specifies an ‘at size’ in scaled points.
- if it is negative, its absolute value represents a ‘scaled’ setting relative to the designsize of the font.

The internal structure of the metrics font table that is returned is explained in **chapter 6**.



4.13.2 Loading a VF file

```
<table> vf_fnt = font.read_vf(<string> name, <number> s)
```

The meaning of the number `s`, and the format of the returned table is similar to the one returned by the `read_tfm()` function.

4.13.3 The fonts array

The whole table of T_EX fonts is accessible from Lua using a virtual array.

```
font.fonts[n] = { ... }  
<table> f = font.fonts[n]
```

See [chapter 6](#) for the structure of the tables. Because this is a virtual array, you cannot call `pairs` on it, but see below for the `font.each` iterator.

The two metatable functions implementing the virtual array are:

```
<table> f = font.getfont(<number> n)  
font.setfont(<number> n, <table> f)
```

Also note the following: assignments can only be made to fonts that have already been defined in T_EX, but have not been accessed *at all* since that definition. This limits the usability of the write access to `font.fonts` quite a lot, a less stringent ruleset will likely be implemented later.

4.13.4 Checking a font's status

You can test for the status of a font by calling this function:

```
<boolean> f = font.frozen(<number> n)
```

The return value is one of true (unassignable), false (can be changed) or nil (not a valid font at all).

4.13.5 Defining a font directly

You can define your own font into `font.fonts` by calling this function:

```
<number> i = font.define(<table> f)
```

The return value is the internal id number of the defined font (the index into `font.fonts`). If the font creation fails, an error is raised. The table is a font structure, as explained in [chapter 6](#).



4.13.6 Projected next font id

```
number i = font.nextid();
```

This returns the font id number that would be returned by a `font.define` call if it was executed at this spot in the code flow. This is useful for virtual fonts that need to reference themselves.

4.13.7 Currently active font

```
<number> i = font.current();  
font.current(<number> i);
```

This gets or sets the currently used font number.

4.13.8 Maximum font id

```
<number> i = font.max();
```

This is the largest used index in `font.fonts`.

4.13.9 Iterating over all fonts

```
for i,v in font.each() do  
    ...  
end
```

This is an iterator over each of the defined T_EX fonts. The first returned value is the index in `font.fonts`, the second the font itself, as a Lua table. The indices are listed incrementally, but they do not always form an array of consecutive numbers: in some cases there can be holes in the sequence.

4.14 The fontforge library

4.14.1 Getting quick information on a font

```
local info = fontforge.info('filename')
```

This function returns either `nil`, or a `table`, or an array of small tables (in the case of a TrueType collection). The returned table(s) will contain six fairly interesting information items from the font(s) defined by the file:



| key | type | explanation |
|-------------|--------|---|
| fontname | string | the PostScript name of the font |
| fullname | string | the formal name of the font |
| famillyname | string | the family name this font belongs to |
| weight | string | a string indicating the color value of the font |
| version | string | the internal font version |
| italicangle | float | the slant angle |

Getting information through this function is (sometimes much) more efficient than loading the font properly, and is therefore handy when you want to create a dictionary of available fonts based on a directory contents.

4.14.2 Loading an OPENTYPE or TRUETYPE file

If you want to use an OpenType font, you have to get the metric information from somewhere. Using the `fontforge` library, the basic way to get that information is thus:

```
function load_font (filename)
  local metrics = nil
  local font = fontforge.open(filename)
  if font then
    metrics = fontforge.to_table(font)
    fontforge.close(font)
  end
  return metrics
end

myfont = load_font('/opt/tex/texmf/fonts/data/arial.ttf')
```

The main function call is

```
f, w = fontforge.open('filename')
```

The first return value is a table representation of the font. The second return value is a table containing any warnings and errors reported by fontforge while opening the font. In normal typesetting, you would probably ignore the second argument, but it can be useful for debugging purposes.

For TrueType collections (when filename ends in 'ttc'), you have to use a second string argument to specify which font you want from the collection. Use one of the `fullname` strings that are returned by `fontforge.info` for that.

```
f, w = fontforge.open('filename', 'fullname')
```

The font file is parsed and partially interpreted by the font loading routines from FontForge. The file format can be OpenType, TrueType, TrueType Collection, cff, or Type1.

There are a few advantages to this approach compared to reading the actual font file ourselves:



- The font is automatically re-encoded, so that the `metrics` table for TrueType and OpenType fonts is using Unicode for the character indices.
- Many features are pre-processed into a format that is easier to handle than just the bare tables would be.
- PostScript-based OpenType fonts do not store the character height and depth in the font file, so the character boundingbox has to be calculated in some way.
- In the future, it may be interesting to allow Lua scripts access to the font program itself, perhaps even creating or changing the font.

4.14.3 Applying a ‘feature file’

You can apply a ‘feature file’ to a loaded font:

```
fontforge.apply_featurefile(f, 'filename')
```

A ‘feature file’ is a textual representation of the features in an OpenType font. See http://www.adobe.com/devnet/opentype/afdko/topic_feature_file_syntax.html and <http://fontforge.sourceforge.net/featurefile.html> for a more detailed description of feature files.

4.14.4 Applying an ‘AFM file’

You can apply an ‘afm file’ to a loaded font:

```
fontforge.apply_afmfile(f, 'filename')
```

An afm file is a textual representation of (some of) the metainformation in a Type1 font. See http://www.adobe.com/devnet/font/pdfs/5004.AFM_Spec.pdf for more information about afm files.

Note: If you `fontforge.open()` a Type1 file named `font.pfb`, the library will automatically search for and apply `font.afm` if it exists in the same directory as the file `font.pfb`. In that case, there is no need for an explicit call to `apply_afmfile()`.

4.15 Fontforge font tables

4.15.1 Table types

4.15.1.1 Top-level

The top-level keys in the returned table are (the explanations in this part of the documentation are not yet finished):



| key | type | explanation |
|------------------------------|--------|--|
| table_version | number | indicates the metrics version (currently 0.3) |
| fontname | string | PostScript font name |
| fullname | string | official font name |
| familyname | string | family name |
| weight | string | weight indicator |
| copyright | string | copyright information |
| filename | string | the file name |
| version | string | font version |
| italicangle | float | slant angle |
| units_per_em | number | 1000 for PostScript-based fonts, usually 2048 for TrueType |
| ascent | number | height of ascender in units_per_em |
| descent | number | depth of descender in units_per_em |
| upos | float | |
| uwidth | float | |
| uniqueid | number | |
| glyphcnt | number | number of included glyphs |
| glyphs | array | |
| glyphmax | number | maximum used index the glyphs array |
| hasvmetrics | number | |
| onlybitmaps | number | |
| serifcheck | number | |
| isserif | number | |
| issans | number | |
| encodingchanged | number | |
| strokedfont | number | |
| use_typo_metrics | number | |
| weight_width_slope_only | number | |
| head_optimized_for_cleartype | number | |
| uni_interp | enum | unset , none , adobe , greek , japanese , trad_chinese , simp_chinese , korean , ams |
| origname | string | the file name, as supplied by the user |
| map | table | |
| private | table | |
| xuid | string | |
| pfminfo | table | |
| names | table | |
| cidinfo | table | |
| subfonts | array | |
| commmments | string | |
| fontlog | string | |
| cvt_names | string | |
| anchor_classes | table | |



| | |
|---------------------|--------|
| ttf_tables | table |
| ttf_tab_saved | table |
| kerns | table |
| vkerns | table |
| texdata | table |
| lookups | table |
| gpos | table |
| gsub | table |
| sm | table |
| features | table |
| mm | table |
| chosename | string |
| macstyle | number |
| fondname | string |
| design_size | number |
| fontstyle_id | number |
| fontstyle_name | table |
| design_range_bottom | number |
| design_range_top | number |
| strokewidth | float |
| mark_classes | array |
| mark_class_names | array |
| creationtime | number |
| modificationtime | number |
| os2_version | number |
| sfd_version | number |
| math | table |
| validation_state | table |
| horiz_base | table |
| vert_base | table |
| extrema_bound | number |

4.15.1.2 Glyph items

The `glyphs` is an array containing the per-character information (quite a few of these are only present if nonzero).

| key | type | explanation |
|--------------|--------|---|
| name | string | the glyph name |
| unicode | number | unicode code point, or -1 |
| boundingbox | array | array of four numbers |
| width | number | only for horizontal fonts |
| vwidth | number | only for vertical fonts |
| lsidebearing | number | only if nonzero and not equal to boundingbox[1] |



| | | |
|-------------------|--------|--|
| class | string | one of "automatic", "none" ,"base", "ligature","mark", "component" |
| kerns | array | only for horizontal fonts, if set |
| vkerns | array | only for vertical fonts, if set |
| dependents | array | linear array of glyph name strings, only if nonempty |
| lookups | table | only if nonempty |
| ligatures | table | only if nonempty |
| anchors | table | only if set |
| comment | string | only if set |
| tex_height | number | only if set |
| tex_depth | number | only if set |
| italic_correction | number | only if set |
| top_accent_horiz | number | only if set |
| is_extended_shape | number | only if this character is part of a math extension list |
| altuni | table | alternate unicode items |
| vert_variants | table | |
| horiz_variants | table | |
| mathkern | table | |

The [kerns](#) and [vkerns](#) are linear arrays of small hashes:

| key | type | explanation |
|--------|--------|-------------|
| char | string | |
| off | number | |
| lookup | string | |

The [lookups](#) is a hash, based on lookup subtable names, with the value of each key inside that a linear array of small hashes:

| key | type | explanation |
|---------------|-------|--|
| type | enum | position , pair , substitution , alternate , multiple , ligature , lcaret , kerning , vkerning , anchors , contextpos , contextsub , chainpos , chainsub , reversesub , max , kernback , vkernback |
| specification | table | extra data |

For the first seven values of [type](#), there can be additional sub-information, stored in the sub-table [specification](#):

| value | type | explanation |
|--------------|-------|--|
| position | table | a table of the offset_specs type |
| pair | table | one string: paired , and an array of one or two offset_specs tables: offsets |
| substitution | table | one string: variant |
| alternate | table | one string: components |
| multiple | table | one string: components |
| ligature | table | two strings: components , char |
| lcaret | array | linear array of numbers |



Tables for `offset_specs` contain up to four number-valued fields: `x` (a horizontal offset), `y` (a vertical offset), `h` (an advance width correction) and `v` (an advance height correction).

The `ligatures` is a linear array of small hashes:

| key | type | explanation |
|-------------------------|--------|--|
| <code>lig</code> | table | uses the same substructure as a single <code>possu</code> item |
| <code>char</code> | string | |
| <code>components</code> | array | linear array of named components |
| <code>ccnt</code> | number | |

The `anchor` table is indexed by a string signifying the anchor type, which is one of

| key | type | explanation |
|-----------------------|-------|--|
| <code>mark</code> | table | placement mark |
| <code>basechar</code> | table | mark for attaching combining items to a base char |
| <code>baselig</code> | table | mark for attaching combining items to a ligature |
| <code>basemark</code> | table | generic mark for attaching combining items to connect to |
| <code>centry</code> | table | cursive entry point |
| <code>cexit</code> | table | cursive exit point |

The content of these is an short array of defined anchors, with the entry keys being the anchor names. For all except `baselig`, the value is a single table with this definition:

| key | type | explanation |
|---------------------------|--------|-------------------------------------|
| <code>x</code> | number | x location |
| <code>y</code> | number | y location |
| <code>ttf_pt_index</code> | number | truetype point index, only if given |

For `baselig`, the value is a small array of such anchor sets sets, one for each constituent item of the ligature.

For clarification, an anchor table could for example look like this :

```
['anchor'] = {
  ['basemark'] = {
    ['Anchor-7'] = { ['x']=170, ['y']=1080 }
  },
  ['mark'] = {
    ['Anchor-1'] = { ['x']=160, ['y']=810 },
    ['Anchor-4'] = { ['x']=160, ['y']=800 }
  },
  ['baselig'] = {
    [1] = { ['Anchor-2'] = { ['x']=160, ['y']=650 } },
    [2] = { ['Anchor-2'] = { ['x']=460, ['y']=640 } }
  }
}
```



4.15.1.3 map table

The top-level map is a list of encoding mappings. Each of those is a table itself.

| key | type | explanation |
|----------|--------|---------------------------------------|
| enccount | number | |
| encmax | number | |
| backmax | number | |
| remap | table | |
| map | array | non-linear array of mappings |
| backmap | array | non-linear array of backward mappings |
| enc | table | |

The `remap` table is very small:

| key | type | explanation |
|----------|--------|-------------|
| firstenc | number | |
| lastenc | number | |
| infont | number | |

The `enc` table is a bit more verbose:

| key | type | explanation |
|------------------|--------|-----------------------------|
| enc_name | string | |
| char_cnt | number | |
| char_max | number | |
| unicode | array | of Unicode position numbers |
| psnames | array | of PostScript glyph names |
| builtin | number | |
| hidden | number | |
| only_1byte | number | |
| has_1byte | number | |
| has_2byte | number | |
| is_unicodebmp | number | only if nonzero |
| is_unicodefull | number | only if nonzero |
| is_custom | number | only if nonzero |
| is_original | number | only if nonzero |
| is_compact | number | only if nonzero |
| is_japanese | number | only if nonzero |
| is_korean | number | only if nonzero |
| is_tradchinese | number | only if nonzero [name?] |
| is_simplechinese | number | only if nonzero |
| low_page | number | |
| high_page | number | |
| iconv_name | string | |
| iso_2022_escape | string | |



4.15.1.4 private table

This is the font's private PostScript dictionary, if any. Keys and values are both strings.

4.15.1.5 cidinfo table

| key | type | explanation |
|------------|--------|-------------|
| registry | string | |
| ordering | string | |
| supplement | number | |
| version | number | |

4.15.1.6 pfminfo table

The `pfminfo` table contains most of the OS/2 information:

| key | type | explanation |
|------------------|--------|-------------|
| pfmset | number | |
| winascent_add | number | |
| windescent_add | number | |
| hheadascent_add | number | |
| hheaddescent_add | number | |
| typoascent_add | number | |
| typodescent_add | number | |
| subsuper_set | number | |
| panose_set | number | |
| hheadset | number | |
| vheadset | number | |
| pfmfamily | number | |
| weight | number | |
| width | number | |
| avgwidth | number | |
| firstchar | number | |
| lastchar | number | |
| fstype | number | |
| linegap | number | |
| vlinegap | number | |
| hhead_ascent | number | |
| hhead_descent | number | |
| hhead_descent | number | |
| os2_typoascent | number | |
| os2_typodescent | number | |
| os2_typolinegap | number | |



| | | |
|------------------|--------|---|
| os2_winascent | number | |
| os2_windescent | number | |
| os2_subxsize | number | |
| os2_subysize | number | |
| os2_subxoff | number | |
| os2_subyoff | number | |
| os2_supxsize | number | |
| os2_supysize | number | |
| os2_supxoff | number | |
| os2_supyoff | number | |
| os2_strikeysize | number | |
| os2_strikeypos | number | |
| os2_family_class | number | |
| os2_xheight | number | |
| os2_capheight | number | |
| os2_defaultchar | number | |
| os2_breakchar | number | |
| os2_vendor | string | |
| codepages | table | A two-number array of encoded code pages |
| unicoderanges | table | A four-number array of encoded unicode ranges |
| panose | table | |

The [panose](#) subtable has exactly 10 string keys:

| key | type | explanation |
|-----------------|--------|---|
| familytype | string | Values as in the OpenType font specification: Any , No Fit , Text and Display , Script , Decorative , Pictorial |
| serifstyle | string | See the OpenType font specification for values |
| weight | string | id. |
| proportion | string | id. |
| contrast | string | id. |
| strokevariation | string | id. |
| armstyle | string | id. |
| letterform | string | id. |
| midline | string | id. |
| xheight | string | id. |

4.15.1.7 names table

Each item has two top-level keys:

| key | type | explanation |
|-------|--------|-------------------------|
| lang | string | language for this entry |
| names | table | |



The [names](#) keys are the actual TrueType name strings. The possible keys are:

| key | explanation |
|-----------------|-------------|
| copyright | |
| family | |
| subfamily | |
| uniqueid | |
| fullname | |
| version | |
| postscriptname | |
| trademark | |
| manufacturer | |
| designer | |
| descriptor | |
| venderurl | |
| designerurl | |
| license | |
| licenseurl | |
| idontknow | |
| preffamilyname | |
| prefmodifiers | |
| compatfull | |
| sampletext | |
| cidfindfontname | |
| wwsfamily | |
| wwssubfamily | |

4.15.1.8 [anchor_classes](#) table

The [anchor_classes](#) classes:

| key | type | explanation |
|--------|--------|--|
| name | string | a descriptive id of this anchor class |
| lookup | string | |
| type | string | one of mark , mkmk , curs , mklg |

4.15.1.9 [gpos](#) table

Th [gpos](#) table has one array entry for each lookup. (The [gpos_](#) prefix is somewhat redundant.)

| key | type | explanation |
|-------|--------|---|
| type | string | one of gpos_single , gpos_pair , gpos_cursive , gpos_mark2base , gpos_mark2ligature , gpos_mark2mark , gpos_context , gpos_contextchain |
| flags | table | |



| | |
|------------------|--------|
| name | string |
| features | array |
| subtables | array |

The flags table has a true value for each of the lookup flags that is actually set:

| key | type | explanation |
|-----------------------------|---------|-------------|
| r2l | boolean | |
| ignorebaseglyphs | boolean | |
| ignoreligatures | boolean | |
| ignorecombiningmarks | boolean | |

The features subtable of gpos has:

| key | type | explanation |
|----------------|--------|----------------|
| tag | string | |
| scripts | table | |
| ismac | number | (only if true) |

The scripts table within features has:

| key | type | explanation |
|---------------|------------------|-------------|
| script | string | |
| langs | array of strings | |

The subtables table has:

| key | type | explanation |
|-------------------------|--------|----------------|
| name | string | |
| suffix | string | (only if used) |
| anchor_classes | number | (only if used) |
| vertical_kerning | number | (only if used) |
| kernclass | table | (only if used) |

The kernclass with subtables table has:

| key | type | explanation |
|----------------|------------------|-------------------|
| firsts | array of strings | |
| seconds | array of strings | |
| lookup | string | associated lookup |
| offsets | array of numbers | |

4.15.1.10 gsub table

This has identical layout to the [gpos](#) table, except for the type:



| key | type | explanation |
|------|--------|--|
| type | string | one of <code>gsub_single</code> , <code>gsub_multiple</code> , <code>gsub_alternate</code> , <code>gsub_ligature</code> , <code>gsub_context</code> , <code>gsub_contextchain</code> , <code>gsub_reversecontextchain</code> |

4.15.1.11 ttf_tables and ttf_tab_saved tables

| key | type | explanation |
|--------|--------|-------------|
| tag | string | |
| len | number | |
| maxlen | number | |
| data | number | |

4.15.1.12 sm table

| key | type | explanation |
|---------|--------|--|
| type | string | one of "indic", "context", "lig", "simple", "insert", "kern" |
| lookup | string | |
| flags | table | a set of boolean values with the keys : "vert", "descending", "always" |
| classes | table | an array of named classes |
| state | table | |

The `state` table has:

| key | type | explanation |
|---------|--------|--|
| next | number | |
| flags | number | |
| context | table | A small table that has 'mark' and 'cur' as possible keys, with the values being lookup names. Only applies if the <code>sm.type = context</code> . |
| insert | table | A small table that has 'mark' and 'cur' as possible keys, with the values strings. Only applies if the <code>sm.type = insert</code> . |
| kern | table | A small array with kern data. Only applies if the <code>sm.type = kern</code> . |

4.15.1.13 features table

| key | type | explanation |
|-----------------|--------|--|
| feature | number | |
| ismutex | number | |
| default_setting | number | |
| strid | number | |
| featname | table | A set of mac names. macnames are like oftnames except that they also have an 'enc' field |
| settings | table | |



The `settings` are:

| key | type | explanation |
|-------------------|--------|--|
| setting | number | |
| strid | number | |
| initially_enabled | number | |
| setname | table | A set of mac names. macnames are like oftnames except that they also have an 'enc' field |

4.15.1.14 mm table

| key | type | explanation |
|----------------------|--------|--|
| axes | table | array of axis names |
| instance_count | number | |
| positions | table | array of instance positions (#axes * instances) |
| defweights | table | array of default weights for instances |
| cdv | string | |
| ndv | string | |
| axismaps | table | |
| named_instance_count | number | |
| named_instances | table | |
| apple | number | |

The `axismaps`:

| key | type | explanation |
|-----------|--------|---------------------------|
| blends | table | an array of blend points |
| designs | table | an array of design values |
| min | number | |
| def | number | |
| max | number | |
| axisnames | table | a set of mac names |

The `named_instances` is an array of instances:

| key | type | explanation |
|--------|-------|-------------------------|
| names | table | a set of mac names |
| coords | table | an array of coordinates |

4.15.1.15 math table

ScriptPercentScaleDown
 ScriptScriptPercentScaleDown
 DelimitedSubFormulaMinHeight



DisplayOperatorMinHeight
 MathLeading
 AxisHeight
 AccentBaseHeight
 FlattenedAccentBaseHeight
 SubscriptShiftDown
 SubscriptTopMax
 SubscriptBaselineDropMin
 SuperscriptShiftUp
 SuperscriptShiftUpCramped
 SuperscriptBottomMin
 SuperscriptBaselineDropMax
 SubSuperscriptGapMin
 SuperscriptBottomMaxWithSubscript
 SpaceAfterScript
 UpperLimitGapMin
 UpperLimitBaselineRiseMin
 LowerLimitGapMin
 LowerLimitBaselineDropMin
 StackTopShiftUp
 StackTopDisplayStyleShiftUp
 StackBottomShiftDown
 StackBottomDisplayStyleShiftDown
 StackGapMin
 StackDisplayStyleGapMin
 StretchStackTopShiftUp
 StretchStackBottomShiftDown
 StretchStackGapAboveMin
 StretchStackGapBelowMin
 FractionNumeratorShiftUp
 FractionNumeratorDisplayStyleShiftUp
 FractionDenominatorShiftDown
 FractionDenominatorDisplayStyleShiftDown
 FractionNumeratorGapMin
 FractionNumeratorDisplayStyleGapMin
 FractionRuleThickness
 FractionDenominatorGapMin
 FractionDenominatorDisplayStyleGapMin
 SkewedFractionHorizontalGap
 SkewedFractionVerticalGap
 OverbarVerticalGap
 OverbarRuleThickness
 OverbarExtraAscender
 UnderbarVerticalGap



UnderbarRuleThickness
 UnderbarExtraDescender
 RadicalVerticalGap
 RadicalDisplayStyleVerticalGap
 RadicalRuleThickness
 RadicalExtraAscender
 RadicalKernBeforeDegree
 RadicalKernAfterDegree
 RadicalDegreeBottomRaisePercent
 MinConnectorOverlap

4.15.1.16 validation_state table

| key | explanation |
|-------------------|-------------|
| bad_ps_fontname | |
| bad_glyph_table | |
| bad_cff_table | |
| bad_metrics_table | |
| bad_cmap_table | |
| bad_bitmaps_table | |
| bad_gx_table | |
| bad_ot_table | |
| bad_os2_version | |
| bad_sfnt_header | |

4.15.1.17 horiz_base and vert_base table

| key | type | explanation |
|---------|-------|------------------------------|
| tags | table | an array of script list tags |
| scripts | table | |

The `scripts` subtable:

| key | type | explanation |
|------------------|--------|-------------|
| baseline | table | |
| default_baseline | number | |
| lang | table | |

The `lang` subtable:

| key | type | explanation |
|--------|--------|--------------|
| tag | string | a script tag |
| ascent | number | |



| | |
|----------|--------|
| descent | number |
| features | table |

The **features** points to an array of tables with the same layout except that in those nested tables, the tag represents a language.

4.15.1.18 altuni table

An array of alternate unicode values. Inside that array are hashes with:

| key | type | explanation |
|---------|--------|-------------|
| unicode | number | |
| variant | number | |

4.15.1.19 vert_variants and horiz_variants table

| key | type | explanation |
|-------------------|--------|-------------|
| variants | string | |
| italic_correction | number | |
| parts | table | |

The **parts** table is an array of smaller tables:

| key | type | explanation |
|----------------------|--------|-------------|
| component | string | |
| is_extender | number | |
| startConnectorLength | number | |
| endConnectorLength | number | |
| fullAdvance | number | |

4.15.1.20 mathkern table

| key | type | explanation |
|--------------|-------|-------------|
| top_right | table | |
| bottom_right | table | |
| top_left | table | |
| bottom_left | table | |

Each of the subtables is an array of small hashes with two keys:

| key | type | explanation |
|--------|--------|-------------|
| height | number | |
| kern | number | |



4.15.1.21 kerns table

Substructure is identical to the per-glyph subtable.

4.15.1.22 vkerns table

Substructure is identical to the per-glyph subtable.

4.15.1.23 texdata table

| key | type | explanation |
|--------|--------|---|
| type | string | unset , text , math , mathext |
| params | array | 22 font numeric parameters |

4.15.1.24 lookups table

Top-level [lookups](#) is quite different from the ones at character level. The keys in this hash are strings, the values the actual lookups, represented as dictionary tables.

| key | type | explanation |
|---------------|--------|--|
| type | number | |
| format | enum | one of glyphs , class , coverage , reversecoverage |
| tag | string | |
| current_class | array | |
| before_class | array | |
| after_class | array | |
| rules | array | an array of rule items |

Rule items have one common item and one specialized item:

| key | type | explanation |
|-----------------|-------|--|
| lookups | array | a linear array of lookup names |
| glyph | array | only if the parent's format is glyph |
| class | array | only if the parent's format is glyph |
| coverage | array | only if the parent's format is glyph |
| reversecoverage | array | only if the parent's format is glyph |

A glyph table is:

| key | type | explanation |
|-------|--------|-------------|
| names | string | |
| back | string | |
| fore | string | |



A class table is:

| key | type | explanation |
|---------|-------|-------------|
| current | array | of numbers |
| before | array | of numbers |
| after | array | of numbers |

coverage:

| key | type | explanation |
|---------|-------|-------------|
| current | array | of strings |
| before | array | of strings |
| after | array | of strings |

reversecoverage:

| key | type | explanation |
|--------------|--------|-------------|
| current | array | of strings |
| before | array | of strings |
| after | array | of strings |
| replacements | string | |

4.16 The lang library

This library provides the interface to Lua_T_EX's structure representing a language, and the associated functions.

```
<language> l = lang.new()
<language> l = lang.new(<number> id)
```

This function creates a new userdata object. An object of type `<language>` is the first argument to most of the other functions in the `lang` library. These functions can also be used as if they were object methods, using the colon syntax.

Without an argument, the next available internal id number will be assigned to this object. With argument, an object will be created that links to the internal language with that id number.

```
<number> n = lang.id(<language> l)
```

returns the internal `\language` id number this object refers to.

```
<string> n = lang.hyphenation(<language> l)
lang.hyphenation(<language> l, <string> n)
```

Either returns the current hyphenation exceptions for this language, or adds new ones. The syntax of the string is explained in the next chapter, [section 5.3](#).



```
lang.clear_hyphenation(<language> l)
```

Clears the exception dictionary for this language.

```
<string> n = lang.clean(<string> o)
```

Creates a hyphenation key from the supplied hyphenation value. The syntax of the argument string is explained in the next chapter, **section 5.3**. This function is useful if you want to do something else based on the words in a dictionary file, like spell-checking.

```
<string> n = lang.patterns(<language> l)  
lang.patterns(<language> l, <string> n)
```

Adds additional patterns for this language object, or returns the current set. The syntax of this string is explained in the next chapter, **section 5.3**.

```
lang.clear_patterns(<language> l)
```

Clears the pattern dictionary for this language.

```
<number> n = lang.prehyphenchar(<language> l)  
lang.prehyphenchar(<language> l, <number> n)
```

Gets or sets the ‘pre-break’ hyphen character for implicit hyphenation in this language (initially the hyphen, decimal 45).

```
<number> n = lang.posthyphenchar(<language> l)  
lang.posthyphenchar(<language> l, <number> n)
```

Gets or sets the ‘post-break’ hyphen character for implicit hyphenation in this language (initially null, decimal 0, indicating emptiness).

```
<number> n = lang.preexhyphenchar(<language> l)  
lang.preexhyphenchar(<language> l, <number> n)
```

Gets or sets the ‘pre-break’ hyphen character for explicit hyphenation in this language (initially null, decimal 0, indicating emptiness).

```
<number> n = lang.postexhyphenchar(<language> l)  
lang.postexhyphenchar(<language> l, <number> n)
```

Gets or sets the ‘post-break’ hyphen character for explicit hyphenation in this language (initially null, decimal 0, indicating emptiness).

```
<boolean> success = lang.hyphenate(<node> head)  
<boolean> success = lang.hyphenate(<node> head, <node> tail)
```



Inserts hyphenation points (discretionary nodes) in a node list. If `tail` is given as argument, processing stops on that node. Currently, `success` is always true if `head` (and `tail`, if specified) are proper nodes, regardless of possible other errors.

Hyphenation works only on ‘characters’, a special subtype of all the glyph nodes with the node subtype having the value 1. Glyph nodes with different subtypes are not processed. See **section 5.1** for more details.





5 Languages and characters, fonts and glyphs

LuaTeX's internal handling of the characters and glyphs that eventually become typeset is quite different from the way TeX82 handles those same objects. The easiest way to explain the difference is to focus on unrestricted horizontal mode (i. e. paragraphs) and hyphenation first. Later on, it will be easy to deal with the differences that occur in horizontal and math modes.

In TeX82, the characters you type are converted into `char_node` records when they are encountered by the main control loop. TeX attaches and processes the font information while creating those records, so that the resulting 'horizontal list' contains the final forms of ligatures and implicit kerning. This packaging is needed because we may want to get the effective width of for instance a horizontal box.

When it becomes necessary to hyphenate words in a paragraph, TeX converts (one word at time) the `char_node` records into a string array by replacing ligatures with their components and ignoring the kerning. Then it runs the hyphenation algorithm on this string, and converts the hyphenated result back into a 'horizontal list' that is consecutively spliced back into the paragraph stream. Keep in mind that the paragraph may contain unboxed horizontal material, which then already contains ligatures and kerns and the words therein are part of the hyphenation process.

The `char_node` records are somewhat misnamed, as they are glyph positions in specific fonts, and therefore not really 'characters' in the linguistic sense. There is no language information inside the `char_node` records. Instead, language information is passed along using `language whatsit` records inside the horizontal list.

In LuaTeX, the situation is quite different. The characters you type are always converted into `glyph_node` records with a special subtype to identify them as being intended as linguistic characters. LuaTeX stores the needed language information in those records, but does not do any font-related processing at the time of node creation. It only stores the index of the font.

When it becomes necessary to typeset a paragraph, LuaTeX first inserts all hyphenation points right into the whole node list. Next, it processes all the font information in the whole list (creating ligatures and adjusting kerning), and finally it adjusts all the subtype identifiers so that the records are 'glyph nodes' from now on.

That was the broad overview. The rest of this chapter will deal with the minutiae of the new process.

5.1 Characters and glyphs

TeX82 (including pdfTeX) differentiated between `char_nodes` and `lig_nodes`. The former are simple items that contained nothing but a 'character' and a 'font' field, and they lived in the same memory as tokens did. The latter also contained a list of components, and a subtype indicating whether this ligature was the result of a word boundary, and it was stored in the same place as other nodes like boxes and kerns and glues.

In LuaTeX, these two types are merged into one, somewhat larger structure called a `glyph_node`. Besides having the old character, font, and component fields, and the new special fields like 'attr' (see [section 7.1.2.12](#)), these nodes also contain:



- A subtype, split into four main types:
 - `character`, for characters to be hyphenated: the lowest bit (bit 0) is set to 1.
 - `glyph`, for specific font glyphs: the lowest bit (bit 0) is not set.
 - `ligature`, for ligatures (bit 1 is set)
 - `ghost`, for ‘ghost objects’ (bit 2 is set)
 The latter two make further use of two extra fields (bits 3 and 4):
 - `left`, for ligatures created from a left word boundary and for ghosts created from `\leftghost`
 - `right`, for ligatures created from a right word boundary and for ghosts created from `\rightghost`
 For ligatures, both bits can be set at the same time (in case of a single-glyph word).
- `glyph_nodes` of type ‘character’ also contain language data, split into four items that were current when the node was created: the `\setlanguage` (15 bits), `\lefthyphenmin` (8 bits), `\righthyphenmin` (8 bits), and `\uchyph` (1 bit).

Incidentally, LuaTeX allows 32768 separate languages, and words can be 256 characters long.

Because the `\uchyph` value is saved in the actual nodes, its handling is subtly different from TeX82: changes to `\uchyph` become effective immediately, not at the end of the current partial paragraph.

Typeset boxes now always have their language information embedded in the nodes themselves, so there is no longer a possible dependancy on the surrounding language settings. In TeX82, a mid-paragraph statement like `\unhbox0` would process the box using the current paragraph language unless there was a `\setlanguage` issued inside the box. In LuaTeX, all language variables are already frozen.

5.2 The main control loop

In LuaTeX’s main loop, almost all input characters that are to be typeset are converted into `glyph_node` records with subtype ‘character’, but there are a few small exceptions.

First, the `\accent` primitives creates nodes with subtype ‘glyph’ instead of ‘character’: one for the actual accent and one for the accentee. The primary reason for this is that `\accent` in TeX82 is explicitly dependent on the current font encoding, so it would not make much sense to attach a new meaning to the primitive’s name, as that would invalidate many old documents and macro packages. A secondary reason is that in TeX82, `\accent` prohibits hyphenation of the current word. Since in LuaTeX hyphenation only takes place on ‘character’ nodes, it is possible to achieve the same effect.

This change of meaning did happen with `\char`, that now generates ‘character’ nodes, consistent with its changed meaning in XeTeX. The changed status of `\char` is not yet finalized, but if it stays as it is now, a new primitive `\glyph` should be added to directly insert a font glyph id.

Second, all the results of processing in math mode eventually become nodes with ‘glyph’ subtypes.

Third, the Aleph-derived commands `\leftghost` and `\rightghost` create nodes of a third subtype: ‘ghost’. These nodes are ignored completely by all further processing until the stage where inter-glyph kerning is added.

Fourth, automatic discretionaries are handled differently. TeX82 inserts an empty discretionary after sensing an input character that matches the `\hyphenchar` in the current font. This test is wrong, in our



opinion: whether or not hyphenation takes place should not depend on the current font, it is a language property.

In LuaTeX, it works like this: if LuaTeX senses a string of input characters that matches the value of the new integer parameter `\exhyphenchar`, it will insert an explicit discretionary after that series of nodes. Initex sets the `\exhyphenchar=\-`. Incidentally, this is a global parameter instead of a language-specific one because it may be useful to change the value depending on the document structure instead of the text language.

The exact status and meaning of `\hyphenchar` is still under consideration, it will probably become used in the character to glyph conversion stage. Currently it is simply ignored.

Fifth, `\setlanguage` no longer creates whatsits. The meaning of `\setlanguage` is changed so that it is now an integer parameter like all others. That integer parameter is used in `\glyph_node` creation to add language information to the glyph nodes. In conjunction, the `\language` primitive is extended so that it always also updates the value of `\setlanguage`.

Sixth, the `\noboundary` command (this command prohibits word boundary processing where that would normally take place) now does create whatsits. These whatsits are needed because the exact place of the `\noboundary` command in the input stream has to be retained until after the ligature and font processing stages.

Finally, there is no longer a `main_loop` label in the code. Remember that TeX82 did quite a lot of processing while adding `char_nodes` to the horizontal list? For speed reasons, it handled that processing code outside of the ‘main control’ loop, and only the first character of any ‘word’ was handled by that ‘main control’ loop. In LuaTeX, there is no longer a need for that (all hard work is done later), and the (now very small) bits of character-handling code have been moved back inline. When `\tracingcommands` is on, this is visible because the full word is reported, instead of just the initial character.

5.3 Loading patterns and exceptions

The hyphenation algorithm in LuaTeX is quite different from the one in TeX82, although it uses essentially the same user input.

After expansion, the argument for `\patterns` has to be proper UTF-8, no `\char` or `\chardef`-ed commands are allowed. (The current implementation is even more strict, and will reject all non-unicode characters, but that will be changed in the future. For now, the generated errors are a valuable tool in discovering font-encoding specific pattern files)

Likewise, the expanded argument for `\hyphenation` also has to be proper UTF-8, but here a tiny little bit of extra syntax is provided:

1. three sets of arguments in curly braces (`{ } { } { }`) indicates a desired complex discretionary, with arguments as in `\discretionary`’s command in normal document input.
2. `-` indicates a desired simple discretionary, cf. `\-` and `\discretionary-` in normal document input.



3. Internal command names are ignored. This rule is provided especially for `\discretionary`, but it also helps to deal with `\relax` commands that may sneak in.
4. `=` indicates a hyphen in the document input (but that is only useful in documents where `\exhyphenchar` is not equal to the hyphen).

The expanded argument is first converted back to a space-separated string while dropping the internal command names. This string is then converted into a dictionary by a routine that creates key–value pairs by converting the other listed items. It is important to note that the keys in an exception dictionary can always be generated from the values. Here are a few examples:

| value | implied key (input) | effect |
|------------------------------|---------------------|--|
| <code>ta-ble</code> | table | <code>ta\ble (= ta\discretionary {-}{ }ble)</code> |
| <code>ba{k-}{ }{c}ken</code> | backen | <code>ba\discretionary {k-}{ }{c}ken</code> |

The resultant patterns and exception dictionary will be stored under the language code that is the present value of `\language`.

In the last line of the table, you see there is no `\discretionary` command in the value: the command is optional in the T_EX-based input syntax. The underlying reason for that is that it is conceivable that a whole dictionary of words is stored as a plain text file and loaded into LuaT_EX using one of the functions in the Lua `lang` library. This loading method is quite a bit faster than going through the T_EX language primitives, but some (most?) of that speed gain would be lost if it had to interpret command sequences while doing so.

The motivation behind the ϵ -T_EX extension `\savingshyphcodes` was that hyphenation heavily depended on font encodings. This is no longer true in LuaT_EX, and the corresponding primitive is ignored pending complete removal. The future semantics of `\uppercase` and `\lowercase` are still under consideration, no changes have taken place yet.

5.4 Applying hyphenation

The internal structures LuaT_EX uses for the insertion of discretionaries in words is very different from the ones in T_EX82, and that means there are some noticeable differences in handling as well.

First and foremost, there is no ‘compressed trie’ involved in hyphenation. The algorithm still reads patgen-generated pattern files, but LuaT_EX uses a finite state hash to match the patterns against the word to be hyphenated. This algorithm is based on the ‘libhnj’ library used by OpenOffice, which in turn is inspired by T_EX. The memory allocation for this new implementation is completely dynamic, so the web2c setting for `trie_size` is ignored.

Differences between LuaT_EX and T_EX82 that are a direct result of that:

- LuaT_EX happily hyphenates the full Unicode character range.
- Pattern and exception dictionary size is limited by the available memory only, all allocations are done dynamically. The trie-related settings in `texmf.cnf` are ignored.



- Because there is no ‘trie preparation’ stage, language patterns never become frozen. This means that the primitive `\patterns` (and its Lua counterpart `lang.patterns`) can be used at any time, not only in `initex`.
- Only the string representation of `\patterns` and `\hyphenation` is stored in the format file. At format load time, they are simply re-evaluated. It follows that there is no real reason to preload languages in the format file. In fact, it is usually not a good idea to do so. It is much smarter to load patterns no sooner than the first time they are actually needed.
- LuaTeX uses the language-specific variables `\prehyphenchar` and `\posthyphenchar` in the creation of implicit discretionary, instead of TeX82’s `\hyphenchar`, and the values of the language-specific variables `\preexhyphenchar` and `\postexhyphenchar` for explicit discretionary (instead of TeX82’s empty discretionary).

Previously, there were problems with changing the node attributes mid-word, but that problem is now solved, as nodes in a word are not converted to and from a string any more (this was required by the old hyphenation code), they are edited in place. Inserted characters and ligatures inherit their attributes from the nearest glyph node item (usually the preceding one, but the following one for the items inserted at the left-hand side of a word).

Word boundaries are no longer implied by font switches, but by language switches. One word can have two separate fonts and still be hyphenated correctly (but it can not have two different languages, the `\setlanguage` command forces a word boundary).

All languages start out with `\prehyphenchar=-1`, `\posthyphenchar=0`, `\preexhyphenchar=0` and `\postexhyphenchar=0`. When you assign the values of one of these four parameters, you are actually changing the settings for the current `\language`, this behavior is compatible with `\patterns` and `\hyphenation`.

LuaTeX also hyphenates the first word in a paragraph.

Words can be up to 256 characters long (up from 64 in TeX82). Longer words generate an error right now, but eventually either the limitation will be removed or perhaps it will become possible to silently ignore the excess characters (this is what happens in TeX82, but there the behavior cannot be controlled).

If you are using the Lua function `lang.hyphenate`, you should be aware that this function expects to receive a list of ‘character’ nodes. It will not operate properly in the presence of ‘glyph’, ‘ligature’, or ‘ghost’ nodes, nor does it know how to deal with kerning. In the near future, it will be able to skip over ‘ghost’ nodes, and we may add a less fuzzy function you can call as well.

The hyphenation exception dictionary is maintained as key-value hash, and that is also dynamic, so the `hyph_size` setting is not used either.

A technical paper detailing the new algorithm will be released as a separate document.

5.5 Applying ligatures and kerning

After all possible hyphenation points have been inserted in the list, LuaTeX will process the list to convert the ‘character’ nodes into ‘glyph’ and ‘ligature’ nodes. This is actually done in two stages: first all ligatures are processed, then all kerning information is applied to the result list. But those two



stages are somewhat dependent on each other: If the used font makes it possible to do so, the ligaturing stage adds virtual ‘character’ nodes to the word boundaries in the list. While doing so, it removes and interprets `noboundary` nodes. The kerning stage deletes those word boundary items after it is done with them, and it does the same for ‘ghost’ nodes. Finally, at the end of the kerning stage, all remaining ‘character’ nodes are converted to ‘glyph’ nodes.

This work separation is worth mentioning because, if you overrule from Lua only one of the two callbacks related to font handling, then you have to make sure you perform the tasks normally done by LuaTeX itself in order to make sure that the other, non-overruled, routine continues to function properly.

Work in this area is not yet complete, but most of the possible cases are handled by our rewritten ligaturing engine. We are working hard to make sure all of the possible inputs will become supported soon.

For example, take the word `office`, hyphenated `of-fice`, using a ‘normal’ font with all the `f-i` ligatures:

```
Initial:           {o}{f}{f}{i}{c}{e}
After hyphenation: {o}{f}{-},{f},{f}{f}{i}{c}{e}
First ligature stage: {o}{f}{-},{f},{ff}{i}{c}{e}
Final result:       {o}{f}{-},{fi},{ffi}{c}{e}
```

That’s bad enough, but if there was a hyphenation point between the `f` and the `i`: `of-f-ice`, the final result should be:

```
{o}{f}{-},
  {f}{-},
  {i},
  {fi}},
  {{ff}{-},
  {i},
  {ffi}}}{c}{e}
```

with discretionaries in the post-break text as well as in the replacement text of the top-level discretionary that resulted from the first hyphenation point. And this is only a simple case.

5.6 Breaking paragraphs into lines

This code is still almost unchanged, but because of the above-mentioned changes with respect to discretionaries and ligatures, line breaking will potentially be different from traditional TeX. The actual line breaking code is still based on the TeX82 algorithms, and it does not expect there to be discretionaries inside of discretionaries.

But that situation is now fairly common in LuaTeX, due to the changes to the ligaturing mechanism. And also, the LuaTeX discretionary nodes are implemented slightly different from the TeX82 nodes: the `no_break` text is now embedded inside the `disc` node, where previously these nodes kept their place in the horizontal list (the discretionary node contained a counter indicating how many nodes to skip).



The combined effect of these two differences is that Lua \TeX does not always use all of the potential breakpoints in a paragraph, especially when fonts with many ligatures are used.





6 Font structure

All T_EX fonts are represented to Lua code as tables, and internally as C structures. All keys in the table below are saved in the internal font structure if they are present in the table returned by the `define_font` callback, or if they result from the normal tfm/vf reading routines if there is no `define_font` callback defined.

The column ‘from vf’ means that this key will be created by the `font.read_vf()` routine, ‘from tfm’ means that the key will be created by the `font.read_tfm()` routine, and ‘used’ means whether or not the LuaT_EX engine itself will do something with the key.

The top-level keys in the table are as follows:

| key | from vf | from tfm | used | value type | description |
|---------------|---------|----------|------|------------|--|
| name | yes | yes | yes | string | metric (file) name |
| area | no | yes | yes | string | (directory) location, typically empty |
| used | no | yes | yes | boolean | used already? (initial: false) |
| characters | yes | yes | yes | table | the defined glyphs of this font |
| checksum | yes | yes | no | number | default: 0 |
| designsize | no | yes | yes | number | expected size (default: 655360 == 10pt) |
| direction | no | yes | yes | number | default: 0 (LTR) |
| encodingbytes | no | no | yes | number | default: depends on <code>format</code> |
| encodingname | no | no | yes | string | encoding name |
| fonts | yes | no | yes | table | locally used fonts |
| fullname | no | no | yes | string | actual (PostScript) name |
| header | yes | no | no | string | header comments, if any |
| hyphenchar | no | no | yes | number | default: TeX’s <code>\hyphenchar</code> |
| parameters | no | yes | yes | hash | default: 7 parameters, all zero |
| size | no | yes | yes | number | loaded (at) size. (default: same as designsize) |
| skewchar | no | no | yes | number | default: TeX’s <code>\skewchar</code> |
| type | yes | no | yes | string | basic type of this font |
| format | no | no | yes | string | disk format type |
| embedding | no | no | yes | string | pdf inclusion |
| filename | no | no | yes | string | disk file name |
| tounicode | no | yes | yes | number | if 1, LuaT _E X assumes per-glyph touni- code entries are present in the font |
| stretch | no | no | yes | number | the ‘stretch’ value from <code>\pdffontex- pand</code> |
| shrink | no | no | yes | number | the ‘shrink’ value from <code>\pdffontex- pand</code> |
| step | no | no | yes | number | the ‘step’ value from <code>\pdffontexpand</code> |



| | | | | | |
|-------------------------------|----|----|-----|---------|--|
| <code>auto_expand</code> | no | no | yes | boolean | the 'autoexpand' keyword from <code>\pdf-fontexpand</code> |
| <code>expansion_factor</code> | no | no | no | number | the actual expansion factor of an expanded font |
| <code>attributes</code> | no | no | yes | string | the <code>\pdffontattr</code> |

The key `name` is always required. The keys `stretch`, `shrink`, `step` and optionally `auto_expand` only have meaning when used together: they can be used to replace a post-loading `\pdffontexpand` command. The `expansion_factor` is value that can be present inside a font in `font.fonts`. It is the actual expansion factor (a value between `-shrink` and `stretch`, with step `step`) of a font that was automatically generated by the font expansion algorithm. The key `attributes` can be used to replace `\pdffontattr`. The key `used` is set by the engine when a font is actively in use, this makes sure that the font's definition is written to the output file (dvi or pdf). The tfm reader sets it to false. The `direction` is a number signalling the 'normal' direction for this font. There are sixteen possibilities:

| number | meaning | number | meaning |
|--------|---------|--------|---------|
| 0 | LT | 8 | TT |
| 1 | LL | 9 | TL |
| 2 | LB | 10 | TB |
| 3 | LR | 11 | TR |
| 4 | RT | 12 | BT |
| 5 | RL | 13 | BL |
| 6 | RB | 14 | BB |
| 7 | RR | 15 | BR |

These are Omega-style direction abbreviations: the first character indicates the 'first' edge of the character glyphs (the edge that is seen first in the writing direction), the second the 'top' side.

The `parameters` is a hash with mixed key types. There are seven possible string keys, as well as a number of integer indices (these start from 8 up). The seven strings are actually used instead of the bottom seven indices, because that gives a nicer user interface.

The names and their internal remapping are:

| name | internal remapped number |
|----------------------------|--------------------------|
| <code>slant</code> | 1 |
| <code>space</code> | 2 |
| <code>space_stretch</code> | 3 |
| <code>space_shrink</code> | 4 |
| <code>x_height</code> | 5 |
| <code>quad</code> | 6 |
| <code>extra_space</code> | 7 |

The keys `type`, `format`, `embedding`, `fullname` and `filename` are used to embed OpenType fonts in the result pdf.

The `characters` table is a list of character hashes indexed by an integer number. The number is the 'internal code' T_EX knows this character by.



Two very special string indexes can be used also: `left_boundary` is a virtual character whose ligatures and kerns are used to handle word boundary processing. `right_boundary` is similar but not actually used for anything (yet!).

Other index keys are ignored.

Each character hash itself is a hash. For example, here is the character 'f' (decimal 102) in the font cmr10 at 10 points:

```
[102] = {
  ['width'] = 200250,
  ['height'] = 455111,
  ['depth'] = 0,
  ['italic'] = 50973,
  ['kerns'] = {
    [63] = 50973,
    [93] = 50973,
    [39] = 50973,
    [33] = 50973,
    [41] = 50973
  },
  ['ligatures'] = {
    [102] = {
      ['char'] = 11,
      ['type'] = 0
    },
    [108] = {
      ['char'] = 13,
      ['type'] = 0
    },
    [105] = {
      ['char'] = 12,
      ['type'] = 0
    }
  }
}
```

Of course a more compact is also possible, but keep in mind that reserved words cannot be used compact and in LuaT_EX we often have a `type` key.

```
[102] = {
  ...
  ligatures = {
    [102] = {
      char = 11,
      ['type'] = 0
    }
  }
}
```



```

    },
    ...
}
}

```

The following top-level keys can be present inside a character hash:

| key | from vf | from tfm | used | value type | description |
|------------------|---------|----------|-------|------------|---|
| width | yes | yes | yes | number | character's width, in sp (default 0) |
| height | no | yes | yes | number | character's height, in sp (default 0) |
| depth | no | yes | yes | number | character's depth, in sp (default 0) |
| italic | no | yes | yes | number | character's italic correction, in sp (default zero) |
| left_protruding | no | no | maybe | number | character's \lpcode |
| right_protruding | no | no | maybe | number | character's \rpcode |
| expansion_factor | no | no | maybe | number | character's \efcode |
| tounicode | no | no | maybe | string | character's Unicode equivalent(s), in UTF-16BE hexadecimal format |
| next | no | yes | yes | number | the 'next larger' character index |
| extensible | no | yes | yes | table | the constituent parts of an extensible recipe |
| kerns | no | yes | yes | table | kerning information |
| ligatures | no | yes | yes | table | ligaturing information |
| commands | yes | no | yes | array | virtual font commands |
| name | no | no | no | string | the character (PostScript) name |
| index | no | no | yes | number | the (OpenType or TrueType) font glyph index |
| used | no | yes | yes | boolean | typeset already (default: false)? |

The values of `left_protruding` and `right_protruding` are used only when `\pdfprotrudechars` is non-zero.

Whether or not `expansion_factor` is used depends on the font's global expansion settings, as well as on the value of `\pdfadjustspacing`.

The usage of `tounicode` is this: if this font specifies a `tounicode=1` at the top level, then LuaTeX will construct a `/ToUnicode` entry for the PDF font (or font subset) based on the character-level `tounicode` strings, where they are available. If a character does not have a sensible Unicode equivalent, do not provide a string either (no empty strings).

If the font-level `tounicode` is not set, then LuaTeX will build up `/ToUnicode` based on the TeX code points you used, and any character-level `tounicodes` will be ignored. *At the moment, the string format is exactly the format that is expected by Adobe CMAP files (UTF-16BE in hexadecimal encoding), minus the enclosing angle brackets. This may change in the future.* Small example: the `tounicode` for a `fi` ligature would be `00660069`.

The presence of `extensible` will overrule `next`, if that is also present.



The `extensible` table is very simple:

| key | type | description |
|-----|--------|------------------------------|
| top | number | 'top' character index |
| mid | number | 'middle' character index |
| bot | number | 'bottom' character index |
| rep | number | 'repeatable' character index |

The `kerns` table is a hash indexed by character index (and 'character index' is defined as either a non-negative integer or the string value `right_boundary`), with the values the kerning to be applied, in scaled points.

The `ligatures` table is a hash indexed by character index (and 'character index' is defined as either a non-negative integer or the string value `right_boundary`), with the values being yet another small hash, with two fields:

| key | type | description |
|------|--------|---|
| type | number | the type of this ligature command, default 0 |
| char | number | the character index of the resultant ligature |

The `char` field in a ligature is required.

The `type` field inside a ligature is the numerical or string value of one of the eight possible ligature types supported by T_EX. When T_EX inserts a new ligature, it puts the new glyph in the middle of the left and right glyphs. The original left and right glyphs can optionally be retained, and when at least one of them is kept, it is also possible to move the new 'insertion point' forward one or two places. The glyph that ends up to the right of the insertion point will become the next 'left'.

| textual (Knuth) | number | string | result |
|-------------------|--------|----------------------------|--------|
| $l + r =: n$ | 0 | <code>=:</code> | $ n$ |
| $l + r =: n$ | 1 | <code>=: </code> | $ nr$ |
| $l + r =: n$ | 2 | <code> =:</code> | $ ln$ |
| $l + r =: n$ | 3 | <code> =: </code> | $ lnr$ |
| $l + r =: > n$ | 5 | <code>=: ></code> | $n r$ |
| $l + r =: > n$ | 6 | <code> =: ></code> | $l n$ |
| $l + r =: > n$ | 7 | <code> =: ></code> | $l nr$ |
| $l + r =: >> n$ | 11 | <code> =: >></code> | $ln r$ |

The default value is 0, and can be left out. That signifies a 'normal' ligature where the ligature replaces both original glyphs. In this table the `|` indicates the final insertion point.

The `commands` array is explained below.



6.1 Real fonts

Whether or not a T_EX font is a ‘real’ font that should be written to the pdf document is decided by the `type` value in the top-level font structure. If the value is `real`, then this is a proper font, and the inclusion mechanism will attempt to add the needed font object definitions to the pdf.

Values for `type`:

| value | description |
|----------------------|------------------------|
| <code>real</code> | this is a base font |
| <code>virtual</code> | this is a virtual font |

The actions to be taken depend on a number of different variables:

- Whether the used font fits in an 8-bit encoding scheme or not
- The type of the disk font file
- The level of embedding requested

A font that uses anything other than an 8-bit encoding vector has to be written to the pdf in a different way.

The rule is: if the font table has `encodingbytes` set to 2, then this is a wide font, in all other cases it isn’t. The value 2 is the default for OpenType and TrueType fonts loaded via Lua. For Type1 fonts, you have to set `encodingbytes` to 2 explicitly. For pk bitmap fonts, wide font encoding is not supported at all.

If no special care is needed, LuaT_EX currently falls back to the mapfile-based solution used by pdfT_EX and dvips. This behavior will be removed in the future, when the existing code becomes integrated in the new subsystem.

But if this is a ‘wide’ font, then the new subsystem kicks in, and some extra fields have to be present in the font structure. In this case, LuaT_EX does not use a map file at all.

The extra fields are: `format`, `embedding`, `fullname`, `cidinfo` (as explained above), `filename`, and the `index` key in the separate characters.

Values for `format` are:

| value | description |
|-----------------------|--|
| <code>type1</code> | this is a PostScript Type1 font |
| <code>type3</code> | this is a bitmapped (pk) font |
| <code>truetype</code> | this is a TrueType or TrueType-based OpenType font |
| <code>opentype</code> | this is a PostScript-based OpenType font |

(`type3` fonts are provided for backward compatibility only, and do not support the new wide encoding options.)

Values for `embedding` are:



| value | description |
|--------|--|
| no | don't embed the font at all |
| subset | include and attempt to subset the font |
| full | include this font in its entirety |

It is not possible to artificially modify the transformation matrix for the font at the moment.

The other fields are used as follows: The `fullname` will be the PostScript/pdf font name. The `cidinfo` will be used as the character set (the CID `/Ordering` and `/Registry` keys). The `filename` points to the actual font file. If you include the full path in the `filename` or if the file is in the local directory, LuaTeX will run a little bit more efficient because it will not have to re-run the `find_xxx_file` callback in that case.

Be careful: when mixing old and new fonts in one document, it is possible to create PostScript name clashes that can result in printing errors. When this happens, you have to change the `fullname` of the font.

Typeset strings are written out in a wide format using 2 bytes per glyph, using the `index` key in the character information as value. The overall effect is like having an encoding based on numbers instead of traditional (PostScript) name-based reencoding. The way to get the correct `index` numbers for Type1 fonts is by loading the font via `fontforge.open`; use the table indices as `index` fields.

This type of reencoding means that there is no longer a clear connection between the text in your input file and the strings in the output pdf file. Dealing with this is high on the agenda.

6.2 Virtual fonts

You have to take the following steps if you want LuaTeX to treat the returned table from `define_font` as a virtual font:

- Set the top-level key `type` to `virtual`.
- Make sure there is at least one valid entry in `fonts` (see below).
- Give a `commands` array to every character (see below).

The presence of the toplevel `type` key with the specific value `virtual` will trigger handling of the rest of the special virtual font fields in the table, but the mere existence of 'type' is enough to prevent LuaTeX from looking for a virtual font on its own.

Therefore, this also works 'in reverse': if you are absolutely certain that a font is not a virtual font, assigning the value `base` or `real` to `type` will inhibit LuaTeX from looking for a virtual font file, thereby saving you a disk search.

The `fonts` is another Lua array. The values are one- or two-key hashes themselves, each entry indicating one of the base fonts in a virtual font. In case your font is referring to itself, you can use the `font.nextid()` function which returns the index of the next to be defined font which is probably the currently defined one.

An example makes this easy to understand



```

fonts = {
  { name = 'ptmr8a', size = 655360 },
  { name = 'psyr', size = 600000 },
  { id = 38 }
}

```

says that the first referenced font (index 1) in this virtual font is `ptrmr8a` loaded at 10pt, and the second is `psyr` loaded at a little over 9pt. The third one is previously defined font that is known to LuaTeX as fontid '38'.

The array index numbers are used by the character command definitions that are part of each character. The `commands` array is a hash where each item is another small array, with the first entry representing a command and the extra items being the parameters to that command. The allowed commands and their arguments are:

| command name | arguments | arg type | description |
|--------------|-----------|-----------|---|
| font | 1 | number | select a new font from the local <code>fonts</code> table |
| char | 1 | number | typeset this character number from the current font, and move right by the character's width |
| node | 1 | node | output this node (list), and move right by the width of this list |
| slot | 2 | number | a shortcut for the combination of a font and char command |
| push | 0 | | save current position |
| nop | 0 | | do nothing |
| pop | 0 | | pop position |
| rule | 2 | 2 numbers | output a rule $w * h$, and move right. |
| down | 1 | number | move down on the page |
| right | 1 | number | move right on the page |
| special | 1 | string | output a <code>\special</code> command |
| image | 1 | image | output an image (the argument can be either an <code><image></code> variable or an <code>image_spec</code> table) |
| comment | any | any | the arguments of this command are ignored |

Here is a rather elaborate glyph commands example:

```

...
commands = {
  {'push'},                -- remember where we are
  {'right', 5000},         -- move right about 0.08pt
  {'font', 3},             -- select the fonts[3] entry
  {'char', 97},            -- place character 97 (ASCII 'a')
  {'pop'},                -- go all the way back
  {'down', -200000},       -- move upwards by about 3pt
  {'special', 'pdf: 1 0 0 rg'} -- switch to red color
  {'rule', 500000, 20000}  -- draw a bar
  {'special', 'pdf: 0 g'}  -- back to black
}

```




```
}  
...
```

The default value for `font` is always 1 at the start of the `commands` array. Therefore, if the virtual font is essentially only a re-encoding, then you do usually not have to create an explicit ‘font’ command in the array.

Rules inside of `commands` arrays are built up using only two dimensions: they do not have depth. For correct vertical placement, an extra `down` command may be needed.

Regardless of the amount of movement you create within the `commands`, the output pointer will always move by exactly the width that was given in the `width` key of the character hash. Any movements that take place inside the `commands` array are ignored on the upper level.

6.2.1 Artificial fonts

Even in a ‘real’ font, there can be virtual characters. When LuaTeX encounters a `commands` field inside a character when it becomes time to typeset the character, it will interpret the commands, just like for a true virtual character. In this case, if you have created no ‘fonts’ array, then the default (and only) ‘base’ font is taken to be the current font itself. In practice, this means that you can create virtual duplicates of existing characters which is useful if you want to create composite characters.

Note: this feature does *not* work the other way around. There can not be ‘real’ characters in a virtual font! You cannot use this technique for font re-encoding either; you need a truly virtual font for that (because characters that are already present cannot be altered).

6.2.2 Example virtual font

Finally, here is a plain TeX input file with a virtual font demonstration:

```
\directlua0 {  
  callback.register('define_font',  
    function (name,size)  
      if name == 'cmr10-red' then  
        f = font.read_tfm('cmr10',size)  
        f.name = 'cmr10-red'  
        f.type = 'virtual'  
        f.fonts = {{ name = 'cmr10', size = size }}  
        for i,v in pairs(f.characters) do  
          if (string.char(i)):find('[taco]hanshartmut') then  
            v.commands = {  
              {'special','pdf: 1 0 0 rg'},  
              {'char',i},  
              {'special','pdf: 0 g'},  
            }  
          end  
        end  
      end  
    end  
  )  
}
```



```

        else
            v.commands = {'char',i}
        end
    end
end
else
    f = font.read_tfm(name,size)
end
return f
end
)
}

\font\myfont = cmr10-red at 10pt \myfont This is a line of text \par
\font\myfontx= cmr10 at 10pt \myfontx Here is another line of text \par

```



7 Nodes

7.1 LUA node representation

T_EX's nodes are represented in Lua as userdata object with a variable set of fields. In the following syntax tables, such the type of such a userdata object is represented as `<node>`.

The current return value of `node.types()` is: `hlist` (0), `vlist` (1), `rule` (2), `ins` (3), `mark` (4), `adjust` (5), `disc` (7), `whatsit` (8), `math` (9), `glue` (10), `kern` (11), `penalty` (12), `unset` (13), `style` (14), `choice` (15), `ord` (16), `op` (17), `bin` (18), `rel` (19), `open` (20), `close` (21), `punct` (22), `inner` (23), `radical` (24), `fraction` (25), `under` (26), `over` (27), `accent` (28), `vcenter` (29), `left` (30), `right` (31), `math_char` (32), `sub_box` (33), `sub_mlist` (34), `math_text_char` (35), `delim` (36), `margin_kern` (37), `glyph` (38), `align_record` (39), `pseudo_file` (40), `pseudo_line` (41), `page_insert` (42), `split_insert` (43), `expr_stack` (44), `nested_list` (45), `span` (46), `attribute` (47), `glue_spec` (48), `attribute_list` (49), `action` (50), `temp` (51), `align_stack` (52), `movement_stack` (53), `if_stack` (54), `unhyphenated` (55), `hyphenated` (56), `delta` (57), `passive` (58), `shape` (59), `fake` (100), but as already mentioned, the math and alignment nodes in this list are not supported at the moment. The useful list is described in the next sections.

7.1.1 Auxiliary items

A few node-typed userdata objects do not occur in the 'normal' list of nodes, but can be pointed to from within that list. They are not quite the same as regular nodes, but it is easier for the library routines to treat them as if they were.

7.1.1.1 glue_spec items

Skips are about the only type of data objects in traditional T_EX that are not a simple value. The structure that represents the glue components of a skip is called a `glue_spec`, and it has the following accessible fields:

| key | type | explanation |
|----------------------------|--------|-------------|
| <code>width</code> | number | |
| <code>stretch</code> | number | |
| <code>stretch_order</code> | number | |
| <code>shrink</code> | number | |
| <code>shrink_order</code> | number | |

These objects are reference counted, so there is actually an extra field named `ref_count` as well. This item type will likely disappear in the future, and the glue fields themselves will become part of the nodes referencing glue items.



7.1.1.2 attribute_list and attribute items

The newly introduced attribute registers are non-trivial, because the value that is attached to a node is essentially a sparse array of key-value pairs.

It is generally easiest to deal with attribute lists and attributes by using the dedicated functions in the `node` library, but for completeness, here is the low-level interface.

An `attribute_list` item is used as a head pointer for a list of attribute items. It has only one user-visible field:

| field | type | explanation |
|-------------------|---------------------------|--------------------------------|
| <code>next</code> | <code><node></code> | pointer to the first attribute |

A normal node's attribute field will point to an item of type `attribute_list`, and the `next` field in that item will point to the first defined 'attribute' item, whose `next` will point to the second 'attribute' item, etc.

Valid fields in `attribute` items:

| field | type | explanation |
|---------------------|---------------------------|-------------------------------|
| <code>next</code> | <code><node></code> | pointer to the next attribute |
| <code>number</code> | number | the attribute type id |
| <code>value</code> | number | the attribute value |

7.1.1.3 action item

Valid fields: `action_type`, `named_id`, `action_id`, `file`, `new_window`, `data`, `ref_count`

These are a special kind of item that only appears inside pdf start link objects.

| field | type | explanation |
|--------------------------|------------------|-------------|
| <code>action_type</code> | number | |
| <code>action_id</code> | number or string | |
| <code>named_id</code> | number | |
| <code>file</code> | string | |
| <code>new_window</code> | number | |
| <code>data</code> | string | |
| <code>ref_count</code> | number | |

7.1.2 Main text nodes

These are the nodes that comprise actual typesetting commands.

A few fields are present in all nodes regardless of their type, these are:

| field | type | explanation |
|-------------------|---------------------------|---------------------------------|
| <code>next</code> | <code><node></code> | The next node in a list, or nil |



| | | |
|---------|--------|--------------------------------------|
| id | number | The node's type (id) number |
| subtype | number | The node subtype identifier |

The **subtype** is sometimes just a stub entry. Not all nodes actually use the **subtype**, but this way you can be sure that all nodes accept it as a valid field name, and that is often handy in node list traversal. In the following tables **next** and **id** are not explicitly mentioned.

Besides these three fields, almost all nodes also have an **attr** field, and there is also a field called **prev**. That last field is always present, but only initialized on explicit request: when the function **node.slide()** is called, it will set up the **prev** fields to be a backwards pointer in the argument node list.

7.1.2.1 hlist nodes

Valid fields: **attr**, **width**, **depth**, **height**, **dir**, **shift**, **glue_order**, **glue_sign**, **glue_set**, **list**

| field | type | explanation |
|------------|--------|---|
| subtype | number | unused |
| attr | <node> | The head of the associated attribute list |
| width | number | |
| height | number | |
| depth | number | |
| shift | number | a displacement perpendicular to the character progression direction |
| glue_order | number | a number in the range 0–4, indicating the glue order |
| glue_set | number | the calculated glue ratio |
| glue_sign | number | |
| list | <node> | the body of this list |
| dir | string | the direction of this box. see 7.1.4.7 |

7.1.2.2 vlist nodes

Valid fields: As for hlist, except that 'shift' is a displacement perpendicular to the line progression direction.

7.1.2.3 rule nodes

Valid fields: **attr**, **width**, **depth**, **height**, **dir**

| field | type | explanation |
|---------|--------|--|
| subtype | number | unused |
| attr | <node> | |
| width | number | the width of the rule; the special value -1073741824 is used for 'running' glue dimensions |
| height | number | the height of the rule (can be negative) |



| | | |
|-------|--------|---|
| depth | number | the depth of the rule (can be negative) |
| dir | string | the direction of this rule. see 7.1.4.7 |

7.1.2.4 ins nodes

Valid fields: [attr](#), [cost](#), [depth](#), [height](#), [spec](#), [list](#)

| field | type | explanation |
|---------|------------------------------|---|
| subtype | number | the insertion class |
| attr | <node> | |
| cost | number | the penalty associated with this insert |
| height | number | |
| depth | number | |
| list | <node> | the body of this insert |
| spec | <node> | a pointer to the <code>\splittopskip</code> glue spec |

7.1.2.5 mark nodes

Valid fields: [attr](#), [class](#), [mark](#)

| field | type | explanation |
|---------|------------------------------|-----------------------------------|
| subtype | number | unused |
| attr | <node> | |
| class | number | the mark class |
| mark | table | a table representing a token list |

7.1.2.6 adjust nodes

Valid fields: [attr](#), [list](#)

| field | type | explanation |
|---------|------------------------------|-----------------------|
| subtype | number | 0 = normal, 1 = 'pre' |
| attr | <node> | |
| list | <node> | adjusted material |

7.1.2.7 disc nodes

Valid fields: [attr](#), [pre](#), [post](#), [replace](#)

| field | type | explanation |
|---------|--------|--|
| subtype | number | indicates the source of a discretionary. 0 = the <code>\discretionary</code> command, 1 = the <code>\-</code> command, 2 = added automatically following a <code>-</code> , 3 = added by the hyphenation algorithm |



| | | |
|---------|------------------------------|--------------------------------|
| attr | <node> | |
| pre | <node> | pointer to the pre-break text |
| post | <node> | pointer to the post-break text |
| replace | <node> | pointer to the no-break text |

7.1.2.8 math nodes

Valid fields: [attr](#), [surround](#)

| field | type | explanation |
|----------|------------------------------|--|
| subtype | number | 0 = 'on', 1 = 'off' |
| attr | <node> | |
| surround | number | width of the <code>\mathsurround</code> kern |

7.1.2.9 glue nodes

Valid fields: [attr](#), [spec](#), [leader](#)

| field | type | explanation |
|---------|------------------------------|---|
| subtype | number | 0 = <code>\skip</code> , 1–18 = internal glue parameters, 100 = <code>\leaders</code> , 101 = <code>\cleaders</code> , 102 = <code>\xleaders</code> |
| attr | <node> | |
| spec | <node> | pointer to a <code>glue_spec</code> item |
| leader | <node> | pointer to a box or rule for leaders |

7.1.2.10 kern nodes

Valid fields: [attr](#), [kern](#)

| field | type | explanation |
|---------|------------------------------|---|
| subtype | number | 0 = from font, 1 = from <code>\kern</code> or <code>\/</code> , 2 = from <code>\accent</code> |
| attr | <node> | |
| kern | number | |

7.1.2.11 penalty nodes

Valid fields: [attr](#), [penalty](#)

| field | type | explanation |
|---------|------------------------------|-------------|
| subtype | number | not used |
| attr | <node> | |
| penalty | number | |



7.1.2.12 glyph nodes

| bit | meaning |
|-----|-----------|
| 0 | character |
| 1 | glyph |
| 2 | ligature |
| 3 | ghost |
| 4 | left |
| 5 | right |

Valid fields: attr, char, font, lang, left, right, uchyph, components, xoffset, yoffset

| field | type | explanation |
|------------|---------|--------------------------------|
| subtype | number | bitfield, with bits: |
| attr | <node> | |
| char | number | |
| font | number | |
| lang | number | |
| left | number | |
| right | number | |
| uchyph | boolean | |
| components | <node> | pointer to ligature components |
| xoffset | number | |
| yoffset | number | |

See [section 5.1](#) for a detailed description of the [subtype](#) field.

7.1.2.13 margin_kern nodes

Valid fields: [attr](#), [width](#), [glyph](#)

| field | type | explanation |
|---------|--------|-------------------------------|
| subtype | number | 0 = left side, 1 = right side |
| attr | <node> | |
| width | number | |
| glyph | <node> | |

7.1.3 Math nodes

These are the so-called ‘noad’s and the nodes that are specifically associated with math processing. Most of these nodes contain sub-nodes so that the list of possible fields is actually quite small. First, the subnodes:

7.1.3.1 Math kernel subnodes

Many object fields in math mode are either simple characters in a specific family or math lists or node lists. There are four associated subnodes that represent these cases.



7.1.3.1.1 math_char and math_text_char subnodes

Valid fields: `attr`, `fam`, `char`

| field | type | explanation |
|-------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |
| <code>char</code> | number | |
| <code>fam</code> | number | |

7.1.3.1.2 sub_box and sub_mlist subnodes

Valid fields: `attr`, `list`

| field | type | explanation |
|-------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |
| <code>list</code> | <code><node></code> | |

7.1.3.2 Math delimiter subnode

There is a fifth subnode type that is used exclusively for delimiter fields.

7.1.3.2.1 delim subnodes

Valid fields: `attr`, `small_fam`, `small_char`, `large_fam`, `large_char`

| field | type | explanation |
|-------------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |
| <code>small_char</code> | number | |
| <code>small_fam</code> | number | |
| <code>large_char</code> | number | |
| <code>large_fam</code> | number | |

7.1.3.3 Math core nodes

First, there are the node types that are associated with the simple math objects. All of these have the same list of fields.

7.1.3.3.1 ord, op, bin, rel, open, close, punct, inner, vcenter, under, over nodes

Valid fields: `attr`, `nucleus`, `sub`, `sup`

| field | type | explanation |
|----------------------|--------|------------------------|
| <code>subtype</code> | number | unused except for 'op' |



| | |
|---------|----------|
| attr | <node> |
| nucleus | <kernel> |
| sub | <kernel> |
| sup | <kernel> |

In ‘op’ nodes, the `subtype` is a numeric value representing the ‘limits’ state. 0 = `\displaylimits`, 1 = `\limits`, 2 = `\nolimits`.

7.1.3.3.2 accent nodes

Valid fields: `attr`, `nucleus`, `sub`, `sup`, `accent`

| field | type | explanation |
|---------|----------|-------------|
| attr | <node> | |
| nucleus | <kernel> | |
| sub | <kernel> | |
| sup | <kernel> | |
| accent | <kernel> | |

7.1.3.3.3 style nodes

Valid fields: `attr`

| field | type | explanation |
|-------|--------|--------------------|
| style | string | contains the style |

There are eight possibilities for the string value: one of ‘display’, ‘text’, ‘script’, or ‘scriptscript’. Each of these can have a trailing ‘`'`’ to signify ‘cramped’ styles.

7.1.3.3.4 choice nodes

Valid fields: `attr`, `display`, `text`, `script`, `scriptscript`

| field | type | explanation |
|--------------|--------|-------------|
| attr | <node> | |
| display | <node> | |
| text | <node> | |
| script | <node> | |
| scriptscript | <node> | |

7.1.3.3.5 radical nodes

Valid fields: `attr`, `nucleus`, `sub`, `sup`, `left`



`pdf_dest` (19), `pdf_thread` (20), `pdf_start_thread` (21), `pdf_end_thread` (22), `pdf_save_pos` (23), `pdf_thread_data` (24), `pdf_link_data` (25), `open` (0), `pdf_setmatrix` (40), `pdf_restore` (42), `fake` (100), `late_lua` (35), `user_defined` (44), `pdf_colorstack` (39), `pdf_save` (41), `cancel_boundary` (43), `close_lua` (36),

7.1.4.1 open nodes

Valid fields: `attr`, `stream`, `name`, `area`, `ext`

| field | type | explanation |
|---------------------|---------------------------|--------------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>stream</code> | number | T _E X's stream id number |
| <code>name</code> | string | file name |
| <code>ext</code> | string | file extension |
| <code>area</code> | string | file area (this may become obsolete) |

7.1.4.2 write nodes

Valid fields: `attr`, `stream`, `data`

| field | type | explanation |
|---------------------|---------------------------|---|
| <code>attr</code> | <code><node></code> | |
| <code>stream</code> | number | T _E X's stream id number |
| <code>data</code> | table | a table representing the token list to be written |

7.1.4.3 close nodes

Valid fields: `attr`, `stream`

| field | type | explanation |
|---------------------|---------------------------|-------------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>stream</code> | number | T _E X's stream id number |

7.1.4.4 special nodes

Valid fields: `attr`, `data`

| field | type | explanation |
|-------------------|---------------------------|---------------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>data</code> | string | the <code>\special</code> information |



7.1.4.5 language nodes

LuaTeX does not have language whatsits any more. All language information is already present inside the glyph nodes themselves. This whatsit subtype will be removed in the next release.

7.1.4.6 local_par nodes

Valid fields: `attr`, `pen_inter`, `pen_broken`, `dir`, `box_left`, `box_left_width`, `box_right`, `box_right_width`

| field | type | explanation |
|------------------------------|---------------------------|--|
| <code>attr</code> | <code><node></code> | |
| <code>pen_inter</code> | number | interline penalty |
| <code>pen_broken</code> | number | broken penalty |
| <code>dir</code> | string | the direction of this par. see 7.1.4.7 |
| <code>box_left</code> | <code><node></code> | the <code>\localleftbox</code> |
| <code>box_left_width</code> | number | width of the <code>\localleftbox</code> |
| <code>box_right</code> | <code><node></code> | the <code>\localrightbox</code> |
| <code>box_right_width</code> | number | width of the <code>\localrightbox</code> |

7.1.4.7 dir nodes

Valid fields: `attr`, `dir`, `level`, `dvi_ptr`, `dvi_h`

| field | type | explanation |
|----------------------|---------------------------|---|
| <code>attr</code> | <code><node></code> | |
| <code>dir</code> | string | the direction (but see below) |
| <code>level</code> | number | nesting level of this direction whatsit |
| <code>dvi_ptr</code> | number | a saved dvi buffer byte offset |
| <code>dir_h</code> | number | a saved dvi position |

A note on `dir` strings. Direction specifiers are three-letter combinations of **T**, **B**, **R**, and **L**.

These are built up out of three separate items:

- the first is the direction of the ‘top’ of paragraphs.
- the second is the direction of the ‘start’ of lines.
- the third is the direction of the ‘top’ of glyphs.

Each of the three items can have 4 separate values, but the directions of the first and second items always have to be perpendicular to each other, which limits the total to 16.

Inside actual `dir` whatsit nodes, the representation of `dir` is not a three-letter but a four-letter combination. The first character in this case is always either `+` or `-`, indicating whether the value is pushed or popped from the direction stack.



7.1.4.8 pdf_literal nodes

Valid fields: `attr`, `mode`, `data`

| field | type | explanation |
|-------------------|---------------------------|--|
| <code>attr</code> | <code><node></code> | |
| <code>mode</code> | number | the 'mode' setting of this literal |
| <code>data</code> | string | the <code>\pdfliteral</code> information |

7.1.4.9 pdf_refobj nodes

Valid fields: `attr`, `objnum`

| field | type | explanation |
|---------------------|---------------------------|----------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>objnum</code> | number | the referenced pdf object number |

7.1.4.10 pdf_refxform nodes

Valid fields: `attr`, `width`, `height`, `depth`, `objnum`.

| field | type | explanation |
|---------------------|---------------------------|----------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>width</code> | number | |
| <code>height</code> | number | |
| <code>depth</code> | number | |
| <code>objnum</code> | number | the referenced pdf object number |

Be aware that `pdf_refxform` nodes have dimensions that are used by LuaTeX.

7.1.4.11 pdf_refximage nodes

Valid fields: `attr`, `width`, `height`, `depth`, `objnum`

| field | type | explanation |
|---------------------|---------------------------|----------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>width</code> | number | |
| <code>height</code> | number | |
| <code>depth</code> | number | |
| <code>objnum</code> | number | the referenced pdf object number |

Be aware that `pdf_refximage` nodes have dimensions that are used by LuaTeX.



7.1.4.12 pdf_annot nodes

Valid fields: `attr`, `width`, `height`, `depth`, `objnum`, `data`

| field | type | explanation |
|---------------------|---------------------------|----------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>width</code> | number | |
| <code>height</code> | number | |
| <code>depth</code> | number | |
| <code>objnum</code> | number | the referenced pdf object number |
| <code>data</code> | string | the annotation data |

7.1.4.13 pdf_start_link nodes

Valid fields: `attr`, `width`, `height`, `depth`, `objnum`, `link_attr`, `action`

| field | type | explanation |
|------------------------|---------------------------|----------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>width</code> | number | |
| <code>height</code> | number | |
| <code>depth</code> | number | |
| <code>objnum</code> | number | the referenced pdf object number |
| <code>link_attr</code> | table | the link attribute token list |
| <code>action</code> | <code><node></code> | the action to perform |

7.1.4.14 pdf_end_link nodes

Valid fields: `attr`

| field | type | explanation |
|-------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |

7.1.4.15 pdf_dest nodes

Valid fields: `attr`, `width`, `height`, `depth`, `named_id`, `dest_id`, `dest_type`, `xyz_zoom`, `objnum`

| field | type | explanation |
|------------------------|---------------------------|---|
| <code>attr</code> | <code><node></code> | |
| <code>width</code> | number | |
| <code>height</code> | number | |
| <code>depth</code> | number | |
| <code>named_id</code> | number | is the <code>dest_id</code> a string value? |
| <code>dest_id</code> | number or string | the destination id |
| <code>dest_type</code> | number | type of destination |



| | | |
|----------|--------|-----------------------|
| xyz_zoom | number | |
| objnum | number | the pdf object number |

7.1.4.16 pdf_thread nodes

Valid fields: `attr`, `width`, `height`, `depth`, `named_id`, `thread_id`, `thread_attr`

| field | type | explanation |
|--------------------------|---------------------------|---------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>width</code> | number | |
| <code>height</code> | number | |
| <code>depth</code> | number | |
| <code>named_id</code> | number | is the tread_id a string value? |
| <code>tread_id</code> | number or string | the thread id |
| <code>thread_attr</code> | number | extra thread information |

7.1.4.17 pdf_start_thread nodes

Valid fields: `attr`, `width`, `height`, `depth`, `named_id`, `thread_id`, `thread_attr`

| field | type | explanation |
|--------------------------|---------------------------|---------------------------------|
| <code>attr</code> | <code><node></code> | |
| <code>width</code> | number | |
| <code>height</code> | number | |
| <code>depth</code> | number | |
| <code>named_id</code> | number | is the tread_id a string value? |
| <code>tread_id</code> | number or string | the thread id |
| <code>thread_attr</code> | number | extra thread information |

7.1.4.18 pdf_end_thread nodes

Valid fields: `attr`

| field | type | explanation |
|-------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |

7.1.4.19 pdf_save_pos nodes

Valid fields: `attr`

| field | type | explanation |
|-------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |



7.1.4.20 late_lua nodes

Valid fields: `attr`, `reg`, `data`, `name`

| field | type | explanation |
|-------------------|---------------------------|---------------------|
| <code>attr</code> | <code><node></code> | |
| <code>reg</code> | number | Lua state id number |
| <code>data</code> | string | data to execute |

7.1.4.21 close_lua nodes

Valid fields: `attr`, `reg`

| field | type | explanation |
|-------------------|---------------------------|---------------------|
| <code>attr</code> | <code><node></code> | |
| <code>reg</code> | number | Lua state id number |

7.1.4.22 pdf_colorstack nodes

Valid fields: `attr`, `stack`, `cmd`, `data`

| field | type | explanation |
|--------------------|---------------------------|----------------------|
| <code>attr</code> | <code><node></code> | |
| <code>stack</code> | number | colorstack id number |
| <code>cmd</code> | number | command to execute |
| <code>data</code> | string | data |

7.1.4.23 pdf_setmatrix nodes

Valid fields: `attr`, `data`

| field | type | explanation |
|-------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |
| <code>data</code> | string | data |

7.1.4.24 pdf_save nodes

Valid fields: `attr`

| field | type | explanation |
|-------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |



7.1.4.25 pdf_restore nodes

Valid fields: `attr`

| field | type | explanation |
|-------------------|---------------------------|-------------|
| <code>attr</code> | <code><node></code> | |

7.1.4.26 user_defined nodes

User-defined whatsit nodes can only be created and handled from Lua code. In effect, they are an extension to the extension mechanism. The LuaTeX engine will simply step over such whatsits without ever looking at the contents.

Valid fields: `attr`, `user_id`, `type`, `value`

| field | type | explanation |
|----------------------|--|-------------------|
| <code>attr</code> | <code><node></code> | |
| <code>user_id</code> | number | id number |
| <code>type</code> | number | type of the value |
| <code>value</code> | number string <code><node></code> table | |

The `type` can have one of five distinct values:

| value | explanation |
|-------|---|
| 97 | the value is an attribute node list |
| 100 | the value is a number |
| 110 | the value is a node list |
| 115 | the value is a string |
| 116 | the value is a token list in Lua table form |



8 Modifications

Besides the expected changes caused by new functionality, there are a number of not-so-expected changes. These are sometimes a side-effect of a new (conflicting) feature, or, more often than not, a change necessary to clean up the internal interfaces.

8.1 Changes from T_EX 3.141592

- See **chapter 5** for many small changes related to paragraph building, language handling, and hyphenation.
- There is no pool file, all strings are embedded during compilation.
- `plus 1 filllll` does not generate an error. The extra ‘l’ is simply typeset.
- The `\endlinechar` can be either added (values 0 or more), or not (negative values). If it is added, the character is always decimal 13 a/k/a `^^M` a/k/a carriage return (This change may be temporary).
- The banner line and the statistics messages are different, as well as many warnings and error texts.

8.2 Changes from E-T_EX 2.2

- The ϵ -T_EX functionality is always present and enabled (but see below about T_EX_{Xe}T), so the prepended asterisk or `-etex` switch for `iniTEX` is not needed.
- T_EX_{Xe}T is not present, so the primitives

```
\TeXXeTstate
\beginR
\beginL
\endR
\endL
```

are missing.

- Some of the tracing information that is output by ϵ -T_EX’s `\tracingassigns` and `\tracingrestores` is not there.
- Register management in LuaT_EX uses the Aleph model, so the maximum value is 65535 and the implementation uses a flat array instead of the mixed flat&sparse model from ϵ -T_EX.
- `savinghyphcodes` is a no-op and may possibly be removed. See **chapter 5** for details.

8.3 Changes from PDFT_EX 1.40



- The (experimental) support for snap nodes has been removed, because it is much more natural to build this functionality on top of node processing and attributes. The associated primitives that are now gone are: `\pdfsnaprefpoint`, `\pdfsnapy`, and `\pdfsnapycomp`.
- The (experimental) support for specialized spacing around nodes has also been removed. The associated primitives that are now gone are: `\pdfadjustinterwordglue`, `\pdfprependkern`, and `\pdfappendkern`, as well as the five supporting primitives `\knbscode`, `\stbscode`, `\shbscode`, `\knbccode`, and `\knaccode`.
- A number of ‘utility functions’ is removed:

| | | |
|-------------------------------|------------------------------|------------------------------|
| <code>\pdfelapsedtime</code> | <code>\pdffilesize</code> | <code>\pdfstrcmp</code> |
| <code>\pdfescapehex</code> | <code>\pdflastmatch</code> | <code>\pdfunescapehex</code> |
| <code>\pdfescapeiname</code> | <code>\pdfmatch</code> | |
| <code>\pdfescapestring</code> | <code>\pdfmdfivesum</code> | |
| <code>\pdffiledump</code> | <code>\pdfresettimer</code> | |
| <code>\pdffilemoddate</code> | <code>\pdfshellescape</code> | |
- A few other experimental primitives are also provided without the extra `pdf` prefix, so they can also be called like this:

| | |
|---------------------------|------------------------|
| <code>\primitive</code> | <code>\ifabsnum</code> |
| <code>\ifprimitive</code> | <code>\ifabsdim</code> |
- The definitions for new didot and new cicero are patched.
- The `\pdfprimitive` is bugfixed.
- The `\pdftexversion` is set to 200.
- The PNG transparency fix from 1.40.6 is not applied (high-level support is pending)
- There is no syntex support yet (planned for 0.30)
- LFS (pdf Files larger than 2GiB) support is not working properly yet.

8.4 Changes from ALEPH RC4

- The input translations from Aleph are not implemented, the related primitives are not available:

| | |
|---------------------------------------|--|
| <code>\DefaultInputMode</code> | <code>\noDefaultInputTranslation</code> |
| <code>\noDefaultInputMode</code> | <code>\noInputTranslation</code> |
| <code>\noInputMode</code> | <code>\InputTranslation</code> |
| <code>\InputMode</code> | <code>\DefaultOutputTranslation</code> |
| <code>\DefaultOutputMode</code> | <code>\noDefaultOutputTranslation</code> |
| <code>\noDefaultOutputMode</code> | <code>\noOutputTranslation</code> |
| <code>\noOutputMode</code> | <code>\OutputTranslation</code> |
| <code>\OutputMode</code> | |
| <code>\DefaultInputTranslation</code> | |
- A small series of bounds checking fixes to `\ocp` and `\ocplist` has been added to prevent the system from crashing due to array indexes running out of bounds.
- The `\hoffset` bug when `\pagedir` TRT is fixed, removing the need for an explicit fix to `\hoffset`
- A bug causing `\fam` to fail for family numbers above 15 is fixed.
- Some bits of Aleph assumed `0` and `null` were identical. This resulted for instance in a bug that sometimes caused an eternal loop when trying to `\show` a box.



- A fair amount of other minor bugs are fixed as well, most of these related to `\tracingcommands` output.
- The number of possible fonts, ocps and ocplists is smaller than their maximum Aleph value (around 5000 fonts and 30000 ocps / ocplists).
- The internal function `scan_dir()` has been renamed to `scan_direction()` to prevent a naming clash.
- The `^^` notation can come in five and six item repetitions also, to insert characters that do not fit in the BMP.
- Glues *immediately after* direction change commands are not legal breakpoints.
- The `\ocp` and `\ocplist` statistics at the end of a run are only printed if OCP's are actually used.

8.5 Changes from standard WEB2C

- There is no `mltex`
- There is no `enctex`
- The following commandline switches are silently ignored, even in non-Lua mode:

```
-8bit
-translate-file=TCXNAME
-mltex
-enc
-etex
```

- `\openout` whatsits are not written to the log file.
- Some of the so-called web2c extensions are hard to set up in non-kpse mode because `texmf.cnf` is not read: `shell-escape` is off (but that is not a problem because of Lua's `os.execute`), and the paranoia checks on `openin` and `openout` do not happen (however, it is easy for a Lua script to do this itself by overloading `io.open`).
- The 'E' option does not do anything useful.





9 Implementation notes

9.1 Primitives overlap

The primitives

| | |
|-----------------------------|--------------------------|
| <code>\pdfpagewidth</code> | <code>\pagewidth</code> |
| <code>\pdfpageheight</code> | <code>\pageheight</code> |
| <code>\fontcharwd</code> | <code>\charwd</code> |
| <code>\fontcharht</code> | <code>\charht</code> |
| <code>\fontchardp</code> | <code>\chardp</code> |
| <code>\fontcharic</code> | <code>\charic</code> |

are all aliases of each other.

9.2 Memory allocation

The single internal memory heap that traditional T_EX used for tokens and nodes is split into two separate arrays. Each of these will grow dynamically when needed.

The `texmf.cnf` settings related to main memory are no longer used (these are: `main_memory`, `mem_bot`, `extra_mem_top` and `extra_mem_bot`). ‘Out of main memory’ errors can still occur, but the limiting factor is now the amount of RAM in your system, not a predefined limit.

Also, the memory (de)allocation routines for nodes are completely rewritten. The relevant code now lives in the C file `luanode.c`, and basically uses a dozen or so avail lists instead of a doubly-linked model. An extra function layer is added so that the code can ask for nodes by type instead of directly requisitioning a certain amount of memory words.

Because of the split into two arrays and the resulting differences in the data structures, some of the Pascal web macros have been duplicated. For instance, there are now `vlink` and `vinfo` as well as `link` and `info`. All access to the variable memory array is now hidden behind a macro called `vmem`.

The implementation of the growth of two arrays (via reallocation) introduces a potential pitfall: the memory arrays should never be used as the left hand side of a statement that can modify the array in question.

The input line buffer and pool size are now also reallocated when needed, and the `texmf.cnf` settings `buf_size` and `pool_size` are silently ignored.

9.3 Sparse arrays

The `\mathcode`, `\delcode`, `\catcode`, `\sfcode`, `\lccode` and `\uccode` tables are now sparse arrays that are implemented in C. They are no longer part of the T_EX ‘equivalence table’ and because



each had 1.1 million entries with a few memory words each, this makes a major difference in memory usage.

These assignments do not yet show up when using the etex tracing routines `\tracingassigns` and `\tracingrestores` (code simply not written yet).

A side-effect of the current implementation is that `\global` is now more expensive in terms of processing than non-global assignments.

See [mathcodes.c](#) and [textcodes.c](#) if you are interested in the details.

Also, the glyph ids within a font are now managed by means of a sparse array and glyph ids can go up to index $2^{21} - 1$.

9.4 Simple single-character csnames

Single-character commands are no longer treated specially in the internals, they are stored in the hash just like the multiletter csnames.

The code that displays control sequences explicitly checks if the length is one when it has to decide whether or not to add a trailing space.

9.5 Compressed format

The format is passed through zlib, allowing it to shrink to roughly half of the size it would have had in uncompressed form. This takes a bit more CPU cycles but much less disk I/O, so it should still be faster.

9.6 Binary file reading

All of the internal code is changed in such a way that if one of the `read_xxx_file` callbacks is not set, then the file is read by a C function using basically the same convention as the callback: a single read into a buffer big enough to hold the entire file contents. While this uses more memory than the previous code (that mostly used `getc` calls), it can be quite a bit faster (depending on your I/O subsystem).



10 Known bugs and limitations

The bugs below are going to be fixed eventually.

The top ones will be fixed soon, but in the later items either the actual problem is hard to find, or the code that causes the bug is going to be replaced by a new subsystem soon anyway, or it may not be worth the hassle and the limitations will eventually be documented.

- The current linebreaking implementation does not yet take all possible breakpoints into account where ligatures are involved in the process. This means that line breaks may change in future versions.
- Sometimes font loading via fontforge generates a message like this

Bad call to gww_iconv_open, neither arg is UCS4 (EUC-CN->UTF-8)

during font loading. This is a limitation of the internal iconv implementation.

- `tex.print()` and `tex.sprint()` do not work if `\directlua` is used in an otp file (in the output of an `expression` rule).
- Handling of attributes in math mode is not complete. The data structures in math mode are quite different from those in text mode, so this will take some extra effort to implement correctly.
- When used inside `\directlua`, `pdf.print()` should create a literal node instead of flushing immediately.
- Not all of Aleph's direction commands are handled properly in pdf mode, and especially the vertical scripts support is missing almost completely (only `TRT` and `TLT` are routinely tested).
- Node pointers are not always checked for validity, so if you make a mistake in the node list processing, LuaTeX may terminate itself with an assertion error or 'Emergency stop'.
- In dvi generation mode, using a `\textdir` switch inside the preamble of a `\halign` results in overprinted text in the dvi file, because the column width is not taken into account during the final placement phase (this is a bug inherited from Aleph). Also, Aleph apparently dislikes having more than one non-grouped `\textdir` command in a single lined paragraph.
- Certain constructs in math mode leak memory nodes.





11 TODO

On top of the ‘normal’ extensions that are planned, there are some more specific small feature requests. Whether these will all be included is not certain yet. New requests are welcome but should fit into our ideas, i.e. no new hard coded solutions. Beware, this is not the roadmap, which is somewhat more ambitious.

- Implement the T_EX primitive `\dimension`, cf. `\number`.
- Do something about `\withoutpt` and/or a new register type `\real`?
- Create callback for the automatic creation of missing characters in fonts.
- Do boxes with dual baselines.
- Make the number of the output box configurable.
- Complete the attributes in math and switch all the nodes to a double-linked list.
- Finish the interface from Lua to T_EX’s internals, specially the hash and equivalence table (a small subpart is implementing `\csname` lookups for `tex.box` access).
- Use of Type1C for embedded PostScript font subsets in traditional 8-bit encodings.
- Support font reencoding of 8-bit fonts via char index instead of via map files.
- Attempt to parse ofm level 0 fonts that are masquerading as level 1.



