

The l3tl-analysis package: analysing token lists*

The L^AT_EX3 Project[†]

Released 2016/03/24

1 l3tl-analysis documentation

This module mostly provides internal functions for use in the l3regex module. However, it provides as a side-effect a user debugging function, very similar to the `\ShowTokens` macro from the `ted` package.

`\tl_show_analysis:N`
`\tl_show_analysis:n`

`\tl_show_analysis:n` $\{\langle token\ list\rangle\}$

Displays to the terminal the detailed decomposition of the $\langle token\ list\rangle$ into tokens, showing the category code of each character token, the meaning of control sequences and active characters, and the value of registers.

1.1 Internal functions

`\s__tl`

The format used to store token lists internally uses the scan mark `\s__tl` as a delimiter.

`__tl_analysis_map_inline:nn`

`__tl_analysis_map_inline:nn` $\{\langle token\ list\rangle\}$ $\{\langle inline\ function\rangle\}$

Applies the $\langle inline\ function\rangle$ to each individual $\langle token\rangle$ in the $\langle token\ list\rangle$. The $\langle inline\ function\rangle$ receives three arguments:

- $\langle tokens\rangle$, which both o-expand and x-expand to the $\langle token\rangle$. The detailed form of $\langle token\rangle$ may change in later releases.
- $\langle catcode\rangle$, a capital hexadecimal digit which denotes the category code of the $\langle token\rangle$ (0: control sequence, 1: begin-group, 2: end-group, 3: math shift, 4: alignment tab, 6: parameter, 7: superscript, 8: subscript, A: space, B: letter, C:other, D:active).
- $\langle char\ code\rangle$, a decimal representation of the character code of the token, -1 if it is a control sequence (with $\langle catcode\rangle$ 0).

*This file describes v6443, last revised 2016/03/24.

[†]E-mail: latex-team@latex-project.org

For optimizations in `l3regex` (when matching control sequences), it may be useful to provide a `_t1_analysis_from_str_map_inline:nn` function, perhaps named `_str_analysis_map_inline:nn`.

1.2 Internal format

The task of the `l3tl-analysis` module is to convert token lists to an internal format which allows us to extract all the relevant information about individual tokens (category code, character code), as well as reconstruct the token list quickly. This internal format is used in `l3regex` where we need to support arbitrary tokens, and it is used in conversion functions in `l3str-convert`, where we wish to support clusters of characters instead of single tokens.

We thus need a way to encode any $\langle token \rangle$ (even begin-group and end-group character tokens) in a way amenable to manipulating tokens individually. The best we can do is to find $\langle tokens \rangle$ which both `o`-expand and `x`-expand to the given $\langle token \rangle$. Collecting more information about the category code and character code is also useful for regular expressions, since most regexes are catcode-agnostic. The internal format thus takes the form of a succession of items of the form

$\langle tokens \rangle \backslash s_t1 \langle catcode \rangle \langle char\ code \rangle \backslash s_t1$

The $\langle tokens \rangle$ `o`- and `x`-expand to the original token in the token list or to the cluster of tokens corresponding to one Unicode character in the given encoding (for `l3str-convert`). The $\langle catcode \rangle$ is given as a single hexadecimal digit, 0 for control sequences. The $\langle char\ code \rangle$ is given as a decimal number, `-1` for control sequences.

Using delimited arguments lets us build the $\langle tokens \rangle$ progressively when doing an encoding conversion in `l3str-convert`. On the other hand, the delimiter `\s_t1` may not appear unbraced in $\langle tokens \rangle$. This is not a problem because we are careful to wrap control sequences in braces (as an argument to `\exp_not:n`) when converting from a general token list to the internal format.

The current rule for converting a $\langle token \rangle$ to a balanced set of $\langle tokens \rangle$ which both `o`-expands and `x`-expands to it is the following.

- A control sequence `\cs` becomes `\exp_not:n { \cs } \s_t1 0 -1 \s_t1`.
- A begin-group character `{` becomes `\exp_after:wN { \if_false: } \fi: \s_t1 1 \langle char\ code \rangle \s_t1`.
- An end-group character `}` becomes `\if_false: { \fi: } \s_t1 2 \langle char\ code \rangle \s_t1`.
- A character with any other category code becomes `\exp_not:n { \langle character \rangle } \s_t1 \langle hex\ catcode \rangle \langle char\ code \rangle \s_t1`.

Index

The italic numbers denote the pages where the corresponding entry is described, numbers underlined point to the definition, all others indicate the places where it is used.

E		S	
exp commands:		\ShowTokens	<i>1</i>
\exp_after:wN	<i>2</i>	str commands:	
\exp_not:n	<i>2, 2</i>	__str_analysis_map_inline:nn	<i>2</i>
F		T	
fi commands:		tl commands:	
\fi:	<i>2, 2</i>	\s__tl . <i>1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2</i>	
I		__tl_analysis_from_str_map_-	
if commands:		inline:nn	<i>2</i>
\if_false:	<i>2, 2</i>	__tl_analysis_map_inline:nn ..	<i>1, 1</i>
		\tl_show_analysis:N	<i>1</i>
		\tl_show_analysis:n	<i>1, 1</i>