

The `l3regex` package: regular expressions in $\mathrm{T}_{\mathrm{E}}\mathrm{X}^*$

The $\mathrm{L}^{\mathrm{A}}\mathrm{T}_{\mathrm{E}}\mathrm{X}3$ Project[†]

Released 2013/12/14

1 `l3regex` documentation

The `l3regex` package provides regular expression testing, extraction of submatches, splitting, and replacement, all acting on token lists. The syntax of regular expressions is mostly a subset of the PCRE syntax (and very close to POSIX), with some additions due to the fact that $\mathrm{T}_{\mathrm{E}}\mathrm{X}$ manipulates tokens rather than characters. For performance reasons, only a limited set of features are implemented. Notably, back-references are not supported.

Let us give a few examples. After

```
\tl_set:Nn \l_my_tl { That~cat. }
\regex_replace_once:nnN { at } { is } \l_my_tl
```

the token list variable `\l_my_tl` holds the text “`This cat.`”, where the first occurrence of “`at`” was replaced by “`is`”. A more complicated example is a pattern to add a comma at the end of each word:

```
\regex_replace_all:nnN { \w+ } { \0 , } \l_my_tl
```

The `\w` sequence represents any “word” character, and `+` indicates that the `\w` sequence should be repeated as many times as possible (at least once), hence matching a word in the input token list. In the replacement text, `\0` denotes the full match (here, a word).

If a regular expression is to be used several times, it can be compiled once, and stored in a regex variable using `\regex_const:Nn`. For example,

```
\regex_const:Nn \c_foo_regex { \c{begin} \cB. (\c[~BE].*) \cE. }
```

stores in `\c_foo_regex` a regular expression which matches the starting marker for an environment: `\begin`, followed by a begin-group token (`\cB.`), then any number of tokens which are neither begin-group nor end-group character tokens (`\c[~BE].*`), ending with an end-group token (`\cE.`). As explained in the next section, the parentheses “capture” the result of `\c[~BE].*`, giving us access to the name of the environment when doing replacements.

*This file describes v4623, last revised 2013/12/14.

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

1.1 Syntax of regular expressions

Most characters match exactly themselves, with an arbitrary category code. Some characters are special and must be escaped with a backslash (*e.g.*, `*` matches a star character). Some escape sequences of the form backslash-letter also have a special meaning (for instance `\d` matches any digit). As a rule,

- every alphanumeric character (`A–Z`, `a–z`, `0–9`) matches exactly itself, and should not be escaped, because `\A`, `\B`, ... have special meanings;
- non-alphanumeric printable ascii characters can (and should) always be escaped: many of them have special meanings (*e.g.*, use `\(`, `\)`, `\?`, `\.`);
- spaces should always be escaped (even in character classes);
- any other character may be escaped or not, without any effect: both versions will match exactly that character.

Note that these rules play nicely with the fact that many non-alphanumeric characters are difficult to input into \TeX under normal category codes. For instance, `\\abc\%` matches the characters `\abc%` (with arbitrary category codes), but does not match the control sequence `\abc` followed by a percent character. Matching control sequences can be done using the `\c{regex}` syntax (see below).

Any special character which appears at a place where its special behaviour cannot apply matches itself instead (for instance, a quantifier appearing at the beginning of a string), after raising a warning.

Characters.

`\x{hh...}` Character with hex code `hh...`

`\xhh` Character with hex code `hh`.

`\a` Alarm (hex 07).

`\e` Escape (hex 1B).

`\f` Form-feed (hex 0C).

`\n` New line (hex 0A).

`\r` Carriage return (hex 0D).

`\t` Horizontal tab (hex 09).

Character types.

`.` A single period matches any token.

`\d` Any decimal digit.

`\h` Any horizontal space character, equivalent to `[\ \^I]`: space and tab.

`\s` Any space character, equivalent to `[\ \^I\^J\^L\^M]`.

`\v` Any vertical space character, equivalent to `[\^J\^K\^L\^M]`. Note that `\^K` is a vertical space, but not a space, for compatibility with Perl.

`\w` Any word character, *i.e.*, alpha-numerics and underscore, equivalent to `[A-Za-z0-9_]`.

`\D` Any token not matched by `\d`.

`\H` Any token not matched by `\h`.

`\N` Any token other than the `\n` character (hex 0A).

`\S` Any token not matched by `\s`.

`\V` Any token not matched by `\v`.

`\W` Any token not matched by `\w`.

Of those, `.`, `\D`, `\H`, `\N`, `\S`, `\V`, and `\W` will match arbitrary control sequences.

Character classes match exactly one token in the subject.

`[...]` Positive character class. Matches any of the specified tokens.

`[^...]` Negative character class. Matches any token other than the specified characters.

`x-y` Within a character class, this denotes a range (can be used with escaped characters).

`[:<name>:]` Within a character class (one more set of brackets), this denotes the POSIX character class *<name>*, which can be `alnum`, `alpha`, `ascii`, `blank`, `cntrl`, `digit`, `graph`, `lower`, `print`, `punct`, `space`, `upper`, `word`, or `xdigit`.

`[:^<name>:]` Negative POSIX character class.

For instance, `[a-oq-z\cC.]` matches any lowercase latin letter except `p`, as well as control sequences (see below for a description of `\c`).

Quantifiers (repetition).

`?` 0 or 1, greedy.

`??` 0 or 1, lazy.

`*` 0 or more, greedy.

`*?` 0 or more, lazy.

`+` 1 or more, greedy.

`++` 1 or more, lazy.

`{n}` Exactly *n*.

`{n,}` *n* or more, greedy.

`{n,}?` *n* or more, lazy.

$\{n, m\}$ At least n , no more than m , greedy.

$\{n, m\}?$ At least n , no more than m , lazy.

anchors and simple assertions.

`\b` Word boundary: either the previous token is matched by `\w` and the next by `\W`, or the opposite. For this purpose, the ends of the token list are considered as `\W`.

`\B` Not a word boundary: between two `\w` tokens or two `\W` tokens (including the boundary).

`^` or `\A` Start of the subject token list.

`$`, `\Z` or `\z` End of the subject token list.

`\G` Start of the current match. This is only different from `^` in the case of multiple matches: for instance `\regex_count:nnN { \G a } { aaba } \1_tmpa_int` yields 2, but replacing `\G` by `^` would result in `\1_tmpa_int` holding the value 1.

Alternation and capturing groups.

`A|B|C` Either one of A, B, or C.

`(...)` Capturing group.

`(?:...)` Non-capturing group.

`(?|...)` Non-capturing group which resets the group number for capturing groups in each alternative. The following group will be numbered with the first unused group number.

The `\c` escape sequence allows to test the category code of tokens, and match control sequences. Each character category is represented by a single uppercase letter:

- C for control sequences;
- B for begin-group tokens;
- E for end-group tokens;
- M for math shift;
- T for alignment tab tokens;
- P for macro parameter tokens;
- U for superscript tokens (up);
- D for subscript tokens (down);
- S for spaces;
- L for letters;

- 0 for others; and
- A for active characters.

The `\c` escape sequence is used as follows.

`\c{<regex>}` A control sequence whose *cname* matches the *<regex>*, anchored at the beginning and end, so that `\c{begin}` matches exactly `\begin`, and nothing else.

`\cX` Applies to the next object, which can be a character, character property, class, or group, and forces this object to only match tokens with category *X* (any of CBEMTPUDSLOA). For instance, `\cL[A-Z\d]` matches uppercase letters and digits of category code letter, `\cC.` matches any control sequence, and `\cO(abc)` matches `abc` where each character has category other.

`\c[XYZ]` Applies to the next object, and forces it to only match tokens with category *X*, *Y*, or *Z* (each being any of CBEMTPUDSLOA). For instance, `\c[LSO](..)` matches two tokens of category letter, space, or other.

`\c[~XYZ]` Applies to the next object and prevents it from matching any token with category *X*, *Y*, or *Z* (each being any of CBEMTPUDSLOA). For instance, `\c[~O]\d` matches digits which have any category different from other.

The category code tests can be used inside classes; for instance, `[\cO\d \c[LO][A-F]]` matches what $\mathrm{T}_{\mathrm{E}}\mathrm{X}$ considers as hexadecimal digits, namely digits with category other, or uppercase letters from *A* to *F* with category either letter or other. Within a group affected by a category code test, the outer test can be overridden by a nested test: for instance, `\cL(ab\cO*cd)` matches `ab*cd` where all characters are of category letter, except `*` which has category other.

The `\u` escape sequence allows to insert the contents of a token list directly into a regular expression or a replacement, avoiding the need to escape special characters. Namely, `\u{<tl var name>}` matches the exact contents of the token list *<tl var>*. Within a `\c{...}` control sequence matching, the `\u` escape sequence only expands its argument once, in effect performing `\tl_to_str:v`. Quantifiers are not supported directly: use a group.

The option `(?i)` makes the match case insensitive (identifying *A-Z* with *a-z*; no Unicode support yet). This applies until the end of the group in which it appears, and can be reverted using `(?-i)`. For instance, in `(?i)(a(?-i)b|c)d`, the letters *a* and *d* are affected by the *i* option. Characters within ranges and classes are affected individually: `(?i)[Y-\]` is equivalent to `[YZ\[\]yz]`, and `(?i)[^aeiou]` matches any character which is not a vowel. Neither character properties, nor `\c{...}` nor `\u{...}` are affected by the *i* option.

In character classes, only `[`, `^`, `-`, `]`, `\` and spaces are special, and should be escaped. Other non-alphanumeric characters can still be escaped without harm. Any escape sequence which matches a single character (`\d`, `\D`, *etc.*) is supported in character classes. If the first character is `^`, then the meaning of the character class is inverted; `^` appearing anywhere else in the range is not special. If the first character (possibly following a leading `^`) is `]` then it does not need to be escaped since ending the range there would

make it empty. Ranges of characters can be expressed using `-`, for instance, `[\D 0-5]` and `[\^6-9]` are equivalent.

Capturing groups are a means of extracting information about the match. Parenthesized groups are labelled in the order of their opening parenthesis, starting at 1. The contents of those groups corresponding to the “best” match (leftmost longest) can be extracted and stored in a sequence of token lists using for instance `\regex_extract_once:nnNTF`.

The `\K` escape sequence resets the beginning of the match to the current position in the token list. This only affects what is reported as the full match. For instance,

```
\regex_extract_all:nnN { a \K . } { a123aaxyz } \l_foo_seq
```

results in `\l_foo_seq` containing the items `{1}` and `{a}`: the true matches are `{a1}` and `{aa}`, but they are trimmed by the use of `\K`. The `\K` command does not affect capturing groups: for instance,

```
\regex_extract_once:nnN { (. \K c)+ \d } { acbc3 } \l_foo_seq
```

results in `\l_foo_seq` containing the items `{c3}` and `{bc}`: the true match is `{acbc3}`, with first submatch `{bc}`, but `\K` resets the beginning of the match to the last position where it appears.

1.2 Syntax of the replacement text

Most of the features described in regular expressions do not make sense within the replacement text. Escaped characters are supported as inside regular expressions. The whole match is accessed as `\0`, and the first 9 submatches are accessed as `\1`, ..., `\9`. Submatches with numbers higher than 9 are accessed as `\g{<number>}` instead.

For instance,

```
\tl_set:Nn \l_my_tl { Hello,~world! }
\regex_replace_all:nnN { ([er]?1|o) . } { \(\0\-\-\1\ ) } \l_my_tl
```

results in `\l_my_tl` holding `H(ell--el)(o,--o) w(or--o)(ld--l)!`

The characters inserted by the replacement have category code 12 (other) by default. The escape sequence `\c` allows to insert characters with arbitrary category codes, as well as control sequences.

`\cXY` Produces the character `Y` (which can be given as an escape sequence such as `\t` for tab) with category code `X`, which must be one of `CBEMTPUDSLOA`.

`\c{<text>}` Produces the control sequence with csname `<text>`. The `<text>` may contain references to the submatches `\0`, `\1` etc.

1.3 Pre-compiling regular expressions

If a regular expression is to be used several times, it is better to compile it once rather than doing it each time the regular expression is used. The compiled regular expression is stored in a variable. All of the `l3regex` module's functions can be given their regular expression argument either as an explicit string or as a compiled regular expression.

<code>\regex_new:N</code>	<code>\regex_new:N <regex var></code>
---------------------------	---

Creates a new *<regex var>* or raises an error if the name is already taken. The declaration is global. The *<regex var>* will initially be such that it never matches.

<code>\regex_set:Nn</code>	<code>\regex_set:Nn <regex var> {<regex>}</code>
<code>\regex_gset:Nn</code>	
<code>\regex_const:Nn</code>	

Stores a compiled version of the *<regular expression>* in the *<regex var>*. For instance, this function can be used as

```
\regex_new:N \l_my_regex
\regex_set:Nn \l_my_regex { my\ (simple\ )? reg(ex|ular\ expression) }
```

The assignment is local for `\regex_set:Nn` and global for `\regex_gset:Nn`. Use `\regex_const:Nn` for compiled expressions which will never change.

<code>\regex_show:n</code>	<code>\regex_show:n {<regex>}</code>
<code>\regex_show:N</code>	

Shows how `l3regex` interprets the *<regex>*. For instance, `\regex_show:n {\A X|Y}` shows

```
+--branch
  anchor at start (\A)
  char code 88
+--branch
  char code 89
```

indicating that the anchor `\A` only applies to the first branch: the second branch is not anchored to the beginning of the match.

1.4 Matching

All regular expression functions are available in both `:n` and `:N` variants. The former require a “standard” regular expression, while the later require a compiled expression as generated by `\regex_(g)set:Nn`.

<code>\regex_match:nnTF</code>	<code>\regex_match:nnTF {<regex>} {<token list>} {<true code>} {<false code>}</code>
<code>\regex_match:NnTF</code>	

Tests whether the *<regular expression>* matches any part of the *<token list>*. For instance,

```
\regex_match:nnTF { b [cde]* } { abecdxcx } { TRUE } { FALSE }
\regex_match:nnTF { [b-dq-w] } { example } { TRUE } { FALSE }
```

leaves TRUE then FALSE in the input stream.

`\regex_count:nnN`
`\regex_count:NnN`

`\regex_count:nnN {<regex>} {<token list>} {<int var>}`

Sets *<int var>* within the current \TeX group level equal to the number of times *<regular expression>* appears in *<token list>*. The search starts by finding the left-most longest match, respecting greedy and ungreedy operators. Then the search starts again from the character following the last character of the previous match, until reaching the end of the token list. Infinite loops are prevented in the case where the regular expression can match an empty token list: then we count one match between each pair of characters. For instance,

```
\int_new:N \l_foo_int
\regex_count:nnN { (b+|c) } { abbababcb } \l_foo_int
```

results in `\l_foo_int` taking the value 5.

1.5 Submatch extraction

`\regex_extract_once:nnNTF`
`\regex_extract_once:NnNTF`

`\regex_extract_once:nnN {<regex>} {<token list>} {<seq var>}`
`\regex_extract_once:nnNTF {<regex>} {<token list>} {<seq var>} {<true code>} {<false code>}`

Finds the first match of the *<regular expression>* in the *<token list>*. If it exists, the match is stored as the zeroth item of the *<seq var>*, and further items are the contents of capturing groups, in the order of their opening parenthesis. The *<seq var>* is assigned locally. If there is no match, the *<seq var>* is cleared. The testing versions insert the *<true code>* into the input stream if a match was found, and the *<false code>* otherwise. For instance, assume that you type

```
\regex_extract_once:nnNTF { \A(La)?TeX(!*)\Z } { LaTeX!!! } \l_foo_seq
{ true } { false }
```

Then the regular expression (anchored at the start with `\A` and at the end with `\Z`) will match the whole token list. The first capturing group, `(La)?`, matches `La`, and the second capturing group, `(!*)`, matches `!!!`. Thus, `\l_foo_seq` will contain the items `{LaTeX!!!}`, `{La}`, and `{!!!}`, and the `true` branch is left in the input stream.

<code>\regex_extract_all:nnNTF</code> <code>\regex_extract_all:NnNTF</code>	<code>\regex_extract_all:nnN {<regex>} {<token list>} <seq var></code> <code>\regex_extract_all:nnNTF {<regex>} {<token list>} <seq var> {<true code>} {<false code>}</code>
--	---

Finds all matches of the *<regular expression>* in the *<token list>*, and stores all the sub-match information in a single sequence (concatenating the results of multiple `\regex_extract_once:nnN` calls). The *<seq var>* is assigned locally. If there is no match, the *<seq var>* is cleared. The testing versions insert the *<true code>* into the input stream if a match was found, and the *<false code>* otherwise. For instance, assume that you type

```
\regex_extract_all:nnNTF { \w+ } { Hello,~world! } \l_foo_seq
{ true } { false }
```

Then the regular expression will match twice, and the resulting sequence contains the two items `{Hello}` and `{world}`, and the `true` branch is left in the input stream.

<code>\regex_split:nnNTF</code> <code>\regex_split:NnNTF</code>	<code>\regex_split:nnN {<regular expression>} {<token list>} <seq var></code> <code>\regex_split:nnNTF {<regular expression>} {<token list>} <seq var> {<true code>} {<false code>}</code>
--	---

Splits the *<token list>* into a sequence of parts, delimited by matches of the *<regular expression>*. If the *<regular expression>* has capturing groups, then the token lists that they match are stored as items of the sequence as well. The assignment to *<seq var>* is local. If no match is found the resulting *<seq var>* has the *<token list>* as its sole item. If the *<regular expression>* matches the empty token list, then the *<token list>* is split into single tokens. The testing versions insert the *<true code>* into the input stream if a match was found, and the *<false code>* otherwise. For example, after

```
\seq_new:N \l_path_seq
\regex_split:nnNTF { / } { the/path/for/this/file.tex } \l_path_seq
{ true } { false }
```

the sequence `\l_path_seq` contains the items `{the}`, `{path}`, `{for}`, `{this}`, and `{file.tex}`, and the `true` branch is left in the input stream.

1.6 Replacement

<code>\regex_replace_once:nnNTF</code> <code>\regex_replace_once:NnNTF</code>	<code>\regex_replace_once:nnN {<regular expression>} {<replacement>} <tl var></code> <code>\regex_replace_once:nnNTF {<regular expression>} {<replacement>} <tl var> {<true code>} {<false code>}</code>
--	---

Searches for the *<regular expression>* in the *<token list>* and replaces the first match with the *<replacement>*. The result is assigned locally to *<tl var>*. In the *<replacement>*, `\0` represents the full match, `\1` represent the contents of the first capturing group, `\2` of the second, *etc.*

<code>\regex_replace_all:nnNTF</code> <code>\regex_replace_all:NnNTF</code>	<code>\regex_replace_all:nnN {<regular expression>} {<replacement>} <tl var></code> <code>\regex_replace_all:nnNTF {<regular expression>} {<replacement>} <tl var> {<true code>} {<false code>}</code>
--	---

Replaces all occurrences of the `\regular expression` in the `<token list>` by the `<replacement>`, where `\0` represents the full match, `\1` represent the contents of the first capturing group, `\2` of the second, *etc.* Every match is treated independently, and matches cannot overlap. The result is assigned locally to `<tl var>`.

1.7 Bugs, misfeatures, future work, and other possibilities

The following need to be done now.

- Change user function names!
- Clean up the use of messages.
- Rewrite the documentation in a more ordered way, perhaps add a BNF?

Additional error-checking to come.

- Detect that a trailing `\c<category>` is an invalid regex.
- Currently, `a{\x34}` is recognized as `a{4}`.
- Cleaner error reporting in the replacement phase.
- Add tracing information.
- Detect attempts to use back-references.
- Test for the maximum register `\c_max_register_int`.
- Find out whether the fact that `\W` and friends match the end-marker leads to bugs. Possibly update `__regex_item_reverse:n`.
- Enforce that `\cC` can only be followed by a match-all dot.

Code improvements to come.

- Change `\skip` to `\dimen` for the array of active threads, and shift the array of submatch informations so that it starts at `\skip0`.
- Optimize `\c{abc}` for matching a specific control sequence.
- Only build `.,` once.
- Use `\skip` for the left and right state stacks when compiling a regex.
- Should `__regex_action_free_group:n` only be used for greedy `{n,}` quantifier? (I think not.)
- Quantifiers for `\u` and assertions.

- Improve digit grabbing for the `\g` escape in replacement. Allow arbitrary integer expressions for all those numbers?
- When matching, keep track of an explicit stack of `current_state` and `current_submatches`.
- If possible, when a state is reused by the same thread, kill other subthreads.
- Use `\dimen` registers rather than `\l__regex_balance_tl` to build `__regex_replacement_balance_one_match:n`.
- Reduce the number of epsilon-transitions in alternatives.
- Optimize simple strings: use less states (`abcade` should give two states, for `abc` and `ade`). [Does that really make sense?]
- Optimize groups with no alternative.
- Optimize states with a single `__regex_action_free:n`.
- Optimize the use of `__regex_action_success:` by inserting it in state 2 directly instead of having an extra transition.
- Optimize the use of `\int_step...` functions.
- Groups don't capture within regexes for csnames; optimize and document.
- Decide and document what `\c{\c{...}}` should do in the replacement text, similar questions for `\u`.
- Better “show” for anchors, properties, and catcode tests.
- Does `\K` really need a new state for itself?
- When compiling, use a boolean `in_cs` and less magic numbers.
- Instead of checking whether the character is special or alphanumeric using its character code, check if it is special in regexes with `\cs_if_exist` tests.

The following features are likely to be implemented at some point in the future.

- Allow `\cL(abc)` in replacement text.
- General look-ahead/behind assertions.
- Regex matching on external files.
- Conditional subpatterns with look ahead/behind: “if what follows is [...], then [...]”.
- `(*...)` and `(?...)` sequences to set some options.
- UTF-8 mode for pdfTeX.

- Newline conventions are not done. In particular, we should have an option for `.` not to match newlines. Also, `\A` should differ from `^`, and `\Z`, `\z` and `$` should differ.
- Unicode properties: `\p{...}` and `\P{...}`; `\X` which should match any “extended” Unicode sequence. This requires to manipulate a lot of data, probably using tree-boxes.

The following features of PCRE or Perl will probably not be implemented.

- `\ddd`, matching the character with octal code `ddd`;
- Callout with `(?C...)`, we cannot run arbitrary user code during the matching, because the regex code uses registers in an unsafe way;
- Conditional subpatterns (other than with a look-ahead or look-behind condition): this is non-regular, isn’t it?
- Named subpatterns: \TeX programmers have lived so far without any need for named macro parameters.

The following features of PCRE or Perl will definitely not be implemented.

- `\cx`, similar to \TeX ’s own `\^^x`;
- Comments: \TeX already has its own system for comments.
- `\Q... \E` escaping: this would require to read the argument verbatim, which is not in the scope of this module.
- Atomic grouping, possessive quantifiers: those tools, mostly meant to fix catastrophic backtracking, are unnecessary in a non-backtracking algorithm, and difficult to implement.
- Subroutine calls: this syntactic sugar is difficult to include in a non-backtracking algorithm, in particular because the corresponding group should be treated as atomic. Also, we cannot afford to run user code within the regular expression matching, because of our “misuse” of registers.
- Recursion: this is a non-regular feature.
- Back-references: non-regular feature, this requires backtracking, which is prohibitively slow.
- Backtracking control verbs: intrinsically tied to backtracking.
- `\C` single byte in UTF-8 mode: \XeTeX and \LuaTeX serve us characters directly, and splitting those into bytes is tricky, encoding dependent, and most likely not useful anyways.

2 l3regex implementation

```

1 <*initex | package>
2 <@@=regex>
3 <*package>
4 \ProvidesExplPackage
5   {\ExplFileName}{\ExplFileDate}{\ExplFileVersion}{\ExplFileDescription}
6 \RequirePackage{l3tl-build, l3tl-analysis, l3flag, l3str, l3str-convert}
7 </package>

```

2.1 Plan of attack

Most regex engines use backtracking. This allows to provide very powerful features (back-references come to mind first), but it is costly, and raises the problem of catastrophic backtracking. Since T_EX is not first and foremost a programming language, complicated code tends to run slowly, and we must use faster, albeit slightly more restrictive, techniques, coming from automata theory.

Given a regular expression of n characters, we do the following:

- (Compiling.) Analyse the regex, finding invalid input, and convert it to an internal representation.
- (Building.) Convert the compiled regex to a non-deterministic finite automaton (NFA) with roughly n states which accepts precisely token lists matching that regex.
- (Matching.) Loop through the query token list one token (one “position”) at a time, exploring in parallel every possible path (“active thread”) through the NFA, considering active threads in an order determined by the quantifiers’ greediness.

We use the following vocabulary in the code comments (and in variable names).

- *Group*: index of the capturing group, -1 for non-capturing groups.
- *Position*: each token in the query is labelled by an integer $\langle position \rangle$, with $\text{min_pos} - 1 \leq \langle position \rangle \leq \text{max_pos}$. The lowest and highest positions correspond to imaginary begin and end markers (with inaccessible category code and character code).
- *Query*: the token list to which we apply the regular expression.
- *State*: each state of the NFA is labelled by an integer $\langle state \rangle$ with $\text{min_state} \leq \langle state \rangle < \text{max_state}$.
- *Active thread*: state of the NFA that is reached when reading the query token list for the matching. Those threads are ordered according to the greediness of quantifiers.
- *Step*: used when matching, starts at 0, incremented every time a character is read, and is not reset when searching for repeated matches. The integer `\l__regex_step_int` is a unique id for all the steps of the matching algorithm.

To achieve a good performance, we abuse TeX's registers in two ways. We access registers directly by number rather than tying them to control sequence using `\int_new:N` and other allocation functions. And we store integers in `\dimen` registers in scaled points (`sp`), using TeX's implicit conversion from dimensions to integers in some contexts. Specifically, the registers are used as follows. When compiling, `\toks` registers are used under the hood by functions from the `l3tl-build` module. When building,

- `\toks<state>` holds the tests and actions to perform in the `<state>` of the NFA.
- (Not implemented yet.) `\skipi` has the form `<group id> plus <left state> minus <right state>`.

When matching,

- `\dimen<state>` is equal to the last `<step>` in which the `<state>` was active.
- (Currently, we use `\skip` instead of `\dimen`.) `\dimen<thread>`, with `min_active ≤ <thread> <max_active>`, is equal to the `<state>` in which the `<thread>` currently is. The `<threads>` are ordered starting from the best to the least preferred.
- `\toks<thread>` holds the submatch information for the `<thread>`, as the contents of a property list.
- `\muskip<position>` holds as its main and stretch components the character and category code of the token at this `<position>` in the query.
- `\toks<position>` holds `<tokens>` which o- and x-expand to the `<position>`-th token in the query.
- `\skip` registers hold the value of end-points of all submatches as would be extracted by the `\regex_extract` functions. Since smaller `\skip` registers are used, the minimum index is twice `max_state`, and the used registers go up to `\l__regex_submatch_int`. They are organized in blocks of `capturing_group`, each block corresponding to one match with all its submatches stored in consecutive `\skips`.

`\count` registers are not abused, which means that we can safely use named integers in this module. Note that `\box` registers are not abused either; maybe we could leverage those for some purpose.

The code is structured as follows. Variables are introduced in the relevant section. First we present some generic helper functions. Then comes the code for compiling a regular expression, and for showing the result of the compilation. The building phase converts a compiled regex to NFA states, and the automaton is run by the code in the following section. The only remaining brick is parsing the replacement text and performing the replacement. We are then ready for all the user functions. Finally, messages, and a little bit of tracing code.

2.2 Helpers

`\tl_to_str:V` A variant we need for the `\u` escape in the replacement text.

```
⋮ \cs_generate_variant:Nn \tl_to_str:n { V }
(End definition for \tl_to_str:V.)
```

2.2.1 Constants and variables

`__regex_tmp:w` Temporary function used for various short-term purposes.

```
9 \cs_new:Npn \__regex_tmp:w { }
(End definition for \__regex_tmp:w.)
```

`\l__regex_internal_a_tl` Temporary variables used for various purposes.

```
\l__regex_internal_b_tl 10 \tl_new:N \l__regex_internal_a_tl
\l__regex_internal_a_int 11 \tl_new:N \l__regex_internal_b_tl
\l__regex_internal_b_int 12 \int_new:N \l__regex_internal_a_int
\l__regex_internal_c_int 13 \int_new:N \l__regex_internal_b_int
\l__regex_internal_bool 14 \int_new:N \l__regex_internal_c_int
\l__regex_internal_seq 15 \bool_new:N \l__regex_internal_bool
\g__regex_internal_tl 16 \seq_new:N \l__regex_internal_seq
17 \tl_new:N \g__regex_internal_tl
```

(End definition for `\l__regex_internal_a_tl` and others. These variables are documented on page ??.)

`\c__regex_no_match_regex` This regular expression matches nothing, but is still a valid regular expression. We could use a failing assertion, but I went for an empty class. It is used as the initial value for regular expressions declared using `\regex_new:N`.

```
18 \tl_const:Nn \c__regex_no_match_regex
19 {
20   \__regex_branch:n
21   { \__regex_class:NnnnN \c_true_bool { } { 1 } { 0 } \c_true_bool }
22 }
```

(End definition for `\c__regex_no_match_regex`. This variable is documented on page ??.)

`\l__regex_balance_int` The first thing we do when matching is to go once through the query token list and store the information for each token as `\muskip` and `\toks` registers. During this phase, `\l__regex_balance_int` counts the balance of begin-group and end-group character tokens which appear before a given point in the token list, and we store it as the shrink component of each `\muskip` register. This variable is also used to keep track of the balance in the replacement text.

```
23 \int_new:N \l__regex_balance_int
```

(End definition for `\l__regex_balance_int`. This variable is documented on page ??.)

2.2.2 Testing characters

`__regex_break_point:TF` When testing whether a character of the query token list matches a given character class in the regular expression, we often have to test it against several ranges of characters, checking if any one of those matches. This is done with a structure like

```
<test1> ... <test_n>
\__regex_break_point:TF {<true code>} {<false code>}
```

If any of the tests succeeds, it calls `__regex_break_true:w`, which cleans up and leaves *<true code>* in the input stream. Otherwise, `__regex_break_point:TF` leaves the *<false code>* in the input stream.

```

24 \cs_new_protected:Npn \__regex_break_true:w
25   #1 \__regex_break_point:TF #2 #3 {#2}
26 \cs_new_protected:Npn \__regex_break_point:TF #1 #2 { #2 }
(End definition for \__regex_break_point:TF.)

```

`__regex_item_reverse:n` This function makes showing regular expressions easier, and lets us define `\D` in terms of `\d` for instance. There is a subtlety: the end of the query is marked by `-2`, and will thus match `\D` and other negated properties; this case is caught by another part of the code.

```

27 \cs_new_protected:Npn \__regex_item_reverse:n #1
28   {
29     #1
30     \__regex_break_point:TF { } \__regex_break_true:w
31   }
(End definition for \__regex_item_reverse:n.)

```

`_regex_item_caseful_equal:n` Simple comparisons triggering `__regex_break_true:w` when true.

```

\_regex_item_caseful_range:nn
32 \cs_new_protected:Npn \__regex_item_caseful_equal:n #1
33   {
34     \if_int_compare:w #1 = \l__regex_current_char_int
35       \exp_after:wN \__regex_break_true:w
36     \fi:
37   }
38 \cs_new_protected:Npn \__regex_item_caseful_range:nn #1 #2
39   {
40     \reverse_if:N \if_int_compare:w #1 > \l__regex_current_char_int
41     \reverse_if:N \if_int_compare:w #2 < \l__regex_current_char_int
42     \exp_after:wN \exp_after:wN \exp_after:wN \__regex_break_true:w
43     \fi:
44     \fi:
45   }
(End definition for \__regex_item_caseful_equal:n and \__regex_item_caseful_range:nn.)

```

`_regex_item_caseless_equal:n` For caseless matching, we perform the test both on the `current_char` and on the `case_`
`_regex_item_caseless_range:nn` `changed_char`. Before doing the second set of tests, we make sure that `case_changed_`
`char` has been computed.

```

46 \cs_new_protected:Npn \__regex_item_caseless_equal:n #1
47   {
48     \if_int_compare:w #1 = \l__regex_current_char_int
49       \exp_after:wN \__regex_break_true:w
50     \fi:
51     \if_int_compare:w \l__regex_case_changed_char_int = \c_max_int
52       \__regex_compute_case_changed_char:
53     \fi:
54     \if_int_compare:w #1 = \l__regex_case_changed_char_int
55       \exp_after:wN \__regex_break_true:w

```



```

56   \fi:
57 }
58 \cs_new_protected:Npn \__regex_item_caseless_range:nn #1 #2
59 {
60   \reverse_if:N \if_int_compare:w #1 > \l__regex_current_char_int
61   \reverse_if:N \if_int_compare:w #2 < \l__regex_current_char_int
62   \exp_after:wN \exp_after:wN \exp_after:wN \__regex_break_true:w
63   \fi:
64   \fi:
65   \if_int_compare:w \l__regex_case_changed_char_int = \c_max_int
66   \__regex_compute_case_changed_char:
67   \fi:
68   \reverse_if:N \if_int_compare:w #1 > \l__regex_case_changed_char_int
69   \reverse_if:N \if_int_compare:w #2 < \l__regex_case_changed_char_int
70   \exp_after:wN \exp_after:wN \exp_after:wN \__regex_break_true:w
71   \fi:
72   \fi:
73 }

```

(End definition for __regex_item_caseless_equal:n and __regex_item_caseless_range:nn.)

__regex_compute_case_changed_char: This function is called when \l__regex_case_changed_char_int has not yet been computed (or rather, when it is set to the marker value \c_max_int). If the current character code is in the range [65,90] (upper-case), then add 32, making it lowercase. If it is in the lower-case letter range [97,122], subtract 32.

```

74 \cs_new_protected_nopar:Npn \__regex_compute_case_changed_char:
75 {
76   \int_set_eq:NN \l__regex_case_changed_char_int \l__regex_current_char_int
77   \if_int_compare:w \l__regex_current_char_int < \c_ninety_one
78   \if_int_compare:w \l__regex_current_char_int < \c_sixty_five
79   \else:
80     \int_add:Nn \l__regex_case_changed_char_int { \c_thirty_two }
81   \fi:
82   \else:
83     \if_int_compare:w \l__regex_current_char_int < \c_one_hundred_twenty_three
84     \if_int_compare:w \l__regex_current_char_int < \c_ninety_seven
85     \else:
86       \int_sub:Nn \l__regex_case_changed_char_int { \c_thirty_two }
87     \fi:
88   \fi:
89   \fi:
90 }

```

(End definition for __regex_compute_case_changed_char:.)

__regex_item_equal:n Those must always be defined to expand to a `caseful` (default) or `caseless` version, and not be protected: they must expand when compiling, to hard-code which tests are caseless or caseful.

```

91 \cs_new_eq:NN \__regex_item_equal:n ?
92 \cs_new_eq:NN \__regex_item_range:nn ?

```

(End definition for __regex_item_equal:n and __regex_item_range:nn.)

`_regex_item_catcode:nT` The argument is a sum of powers of 4 with exponents given by the allowed category codes (between 0 and 13). Dividing by a given power of 4 gives an odd result if and only if that category code is allowed. If the catcode does not match, then skip the character code tests which follow.

```

93 \cs_new_protected:Npn \_regex_item_catcode:
94 {
95   "
96   \if_case:w \l__regex_current_catcode_int
97     1      \or: 4      \or: 10      \or: 40
98     \or: 100  \or:      \or: 1000   \or: 4000
99     \or: 10000 \or:      \or: 100000 \or: 400000
100    \or: 1000000 \or: 4000000 \else: 1*\c_zero
101    \fi:
102  }
103 \cs_new_protected:Npn \_regex_item_catcode:nT #1
104 {
105   \if_int_odd:w \_int_eval:w #1 / \_regex_item_catcode: \_int_eval_end:
106   \exp_after:wN \use:n
107   \else:
108   \exp_after:wN \use_none:n
109   \fi:
110 }
111 \cs_new_protected:Npn \_regex_item_catcode_reverse:nT #1#2
112 { \_regex_item_catcode:nT {#1} { \_regex_item_reverse:n {#2} } }
(End definition for \_regex_item_catcode:nT and \_regex_item_catcode_reverse:nT.)

```

`_regex_item_exact:nn` This matches an exact *category*-*character code* pair, or an exact control sequence.

```

\_regex_item_exact_cs:c
113 \cs_new_protected:Npn \_regex_item_exact:nn #1#2
114 {
115   \if_int_compare:w #1 = \l__regex_current_catcode_int
116   \if_int_compare:w #2 = \l__regex_current_char_int
117   \exp_after:wN \exp_after:wN \exp_after:wN \_regex_break_true:w
118   \fi:
119   \fi:
120 }
121 \cs_new_protected:Npn \_regex_item_exact_cs:c #1
122 {
123   \int_compare:nNnTF \l__regex_current_catcode_int = \c_zero
124   {
125     \str_if_eq_x:nnTF
126     {
127       \exp_after:wN \exp_after:wN \exp_after:wN \cs_to_str:N
128       \tex_the:D \tex_toks:D \l__regex_current_pos_int
129     }
130     { #1 }
131     { \_regex_break_true:w } { }
132   }
133   { }
134 }

```

(End definition for `_regex_item_exact:nn` and `_regex_item_exact_cs:c`.)

`_regex_item_cs:n` Match a control sequence (the argument is a compiled regex). First test the catcode of the current token to be zero. Then perform the matching test, and break if the csname indeed matches. The three `\exp_after:wN` expand the contents of the `\toks<current position>` (of the form `\exp_not:n {<control sequence>}`) to `<control sequence>`.

```

135 \cs_new_protected:Npn \_regex_item_cs:n #1
136 {
137   \int_compare:nNnT \l__regex_current_catcode_int = \c_zero
138   {
139     \group_begin:
140       \_regex_single_match:
141       \_regex_disable_submatches:
142       \_regex_build_for_cs:n {#1}
143       \bool_set_eq:NN \l__regex_saved_success_bool \g__regex_success_bool
144       \exp_args:Nx \_regex_match:n
145       {
146         \exp_after:wN \exp_after:wN
147         \exp_after:wN \cs_to_str:N
148         \tex_the:D \tex_toks:D \l__regex_current_pos_int
149       }
150       \if_meaning:w \c_true_bool \g__regex_success_bool
151       \group_insert_after:N \_regex_break_true:w
152       \fi:
153       \bool_gset_eq:NN \g__regex_success_bool \l__regex_saved_success_bool
154     \group_end:
155   }
156 }

```

(End definition for `_regex_item_cs:n`.)

2.2.3 Character property tests

`_regex_prop_d:` Character property tests for `\d`, `\W`, *etc.* These character properties are not affected by the `(?i)` option. The characters recognized by each one are as follows: `\d=[0-9]`, `\w=[0-9A-Z_a-z]`, `\s=[_\^\^I\^\^J\^\^L\^\^M]`, `\h=[_\^\^I]`, `\v=[\^\^J-\^\^M]`, and the upper case counterparts match anything that the lower case does not match. The order in which the various tests appear is optimized for usual mostly lower case letter text.

```

157 \cs_new_protected_nopar:Npn \_regex_prop_d:
158 { \_regex_item_caseful_range:nn \c_forty_eight { 57 } } % 0--9
159 \cs_new_protected_nopar:Npn \_regex_prop_h:
160 {
161   \_regex_item_caseful_equal:n \c_thirty_two % space
162   \_regex_item_caseful_equal:n \c_nine % tab
163 }
164 \cs_new_protected_nopar:Npn \_regex_prop_s:
165 {
166   \_regex_item_caseful_equal:n \c_thirty_two % space
167   \_regex_item_caseful_equal:n \c_nine % tab
168   \_regex_item_caseful_equal:n \c_ten % lf

```

```

169     \_regex_item_caseful_equal:n \c_twelve      % ff
170     \_regex_item_caseful_equal:n \c_thirteen   % cr
171 }
172 \cs_new_protected_nopar:Npn \_regex_prop_v:
173 { \_regex_item_caseful_range:nn \c_ten \c_thirteen } % lf, vtab, ff, cr
174 \cs_new_protected_nopar:Npn \_regex_prop_w:
175 {
176     \_regex_item_caseful_range:nn \c_ninety_seven { 122 } % a--z
177     \_regex_item_caseful_range:nn \c_sixty_five { 90 } % A--Z
178     \_regex_item_caseful_range:nn \c_forty_eight { 57 } % 0--9
179     \_regex_item_caseful_equal:n { 95 } % _
180 }
181 \cs_new_protected_nopar:Npn \_regex_prop_N:
182 { \_regex_item_reverse:n { \_regex_item_caseful_equal:n \c_ten } }

```

(End definition for _regex_prop_d: and others.)

```

\_regex_posix_alnum: POSIX properties. No surprise.
\_regex_posix_alpha: 183 \cs_new_protected_nopar:Npn \_regex_posix_alnum:
\_regex_posix_ascii: 184 { \_regex_posix_alpha: \_regex_posix_digit: }
\_regex_posix_blank: 185 \cs_new_protected_nopar:Npn \_regex_posix_alpha:
\_regex_posix_cntrl: 186 { \_regex_posix_lower: \_regex_posix_upper: }
\_regex_posix_digit: 187 \cs_new_protected_nopar:Npn \_regex_posix_ascii:
\_regex_posix_graph: 188 { \_regex_item_caseful_range:nn \c_zero \c_one_hundred_twenty_seven }
\_regex_posix_lower: 189 \cs_new_eq:NN \_regex_posix_blank: \_regex_prop_h:
\_regex_posix_print: 190 \cs_new_protected_nopar:Npn \_regex_posix_cntrl:
\_regex_posix_punct: 191 {
\_regex_posix_space: 192     \_regex_item_caseful_range:nn \c_zero { 31 }
\_regex_posix_upper: 193     \_regex_item_caseful_equal:n \c_one_hundred_twenty_seven
194 }
\_regex_posix_word: 195 \cs_new_eq:NN \_regex_posix_digit: \_regex_prop_d:
\_regex_posix_xdigit: 196 \cs_new_protected_nopar:Npn \_regex_posix_graph:
197 { \_regex_item_caseful_range:nn { 33 } { 126 } }
198 \cs_new_protected_nopar:Npn \_regex_posix_lower:
199 { \_regex_item_caseful_range:nn \c_ninety_seven { 122 } }
200 \cs_new_protected_nopar:Npn \_regex_posix_print:
201 { \_regex_item_caseful_range:nn \c_thirty_two { 126 } }
202 \cs_new_protected_nopar:Npn \_regex_posix_punct:
203 {
204     \_regex_item_caseful_range:nn { 33 } { 47 }
205     \_regex_item_caseful_range:nn { 58 } { 64 }
206     \_regex_item_caseful_range:nn { 91 } { 96 }
207     \_regex_item_caseful_range:nn { 123 } { 126 }
208 }
209 \cs_new_protected_nopar:Npn \_regex_posix_space:
210 {
211     \_regex_item_caseful_equal:n \c_thirty_two
212     \_regex_item_caseful_range:nn \c_nine \c_thirteen
213 }
214 \cs_new_protected_nopar:Npn \_regex_posix_upper:
215 { \_regex_item_caseful_range:nn \c_sixty_five { 90 } }

```

```

216 \cs_new_eq:NN \__regex_posix_word: \__regex_prop_w:
217 \cs_new_protected_nopar:Npn \__regex_posix_xdigit:
218 {
219   \__regex_posix_digit:
220   \__regex_item_caseful_range:nn \c_sixty_five { 70 }
221   \__regex_item_caseful_range:nn \c_ninety_seven { 102 }
222 }

```

(End definition for __regex_posix_alnum: and others.)

2.2.4 Simple character escape

Before actually parsing the regular expression or the replacement text, we go through them once, converting `\n` to the character 10, *etc.* In this pass, we also convert any special character (`*`, `?`, `{`, *etc.*) or escaped alphanumeric character into a marker indicating that this was a special sequence, and replace escaped special characters and non-escaped alphanumeric characters by markers indicating that those were “raw” characters. The rest of the code can then avoid caring about escaping issues (those can become quite complex to handle in combination with ranges in character classes).

Usage: `__regex_escape_use:nnnn` *<inline 1>* *<inline 2>* *<inline 3>* *{<token list>}*
The *<token list>* is converted to a string, then read from left to right, interpreting backslashes as escaping the next character. Unescaped characters are fed to the function *<inline 1>*, and escaped characters are fed to the function *<inline 2>* within an `x`-expansion context (typically those functions perform some tests on their argument to decide how to output them). The escape sequences `\a`, `\e`, `\f`, `\n`, `\r`, `\t` and `\x` are recognized, and those are replaced by the corresponding character, then fed to *<inline 3>*. The result is then left in the input stream. Spaces are ignored unless escaped.

The conversion is mostly done within an `x`-expanding assignment, except for the `\x` escape sequence, which is not amenable to that in general. For this, we use the general framework of `__tl_build:Nw`.

`__regex_escape_use:nnnn` The result is built in `\l__regex_internal_a_tl`, which is then left in the input stream. Go through `#4` once, applying `#1`, `#2`, or `#3` as relevant to each character (after de-escaping it). Note that we cannot replace `\tl_set:Nx` and `__tl_build_one:o` by a single call to `__tl_build_one:x`, because the `x`-expanding assignment may be interrupted by `\x`.

```

223 \cs_new_protected:Npn \__regex_escape_use:nnnn #1#2#3#4
224 {
225   <trace> \trace_push:nnn { regex } { 1 } { \__regex_escape_use:nnnn }
226   \__tl_build:Nw \l__regex_internal_a_tl
227   \cs_set_nopar:Npn \__regex_escape_unescaped:N ##1 { #1 }
228   \cs_set_nopar:Npn \__regex_escape_escaped:N ##1 { #2 }
229   \cs_set_nopar:Npn \__regex_escape_raw:N ##1 { #3 }
230   \int_set:Nn \tex_escapechar:D { 92 }
231   \__str_gset_other:Nn \g__regex_internal_tl { #4 }
232   \tl_set:Nx \l__regex_internal_b_tl
233   {
234     \exp_after:wN \__regex_escape_loop:N \g__regex_internal_tl
235     { break } \__prg_break_point:
236   }

```

```

237     \_tl_build_one:o \l__regex_internal_b_tl
238     \_tl_build_end:
239     <trace> \trace_pop:nnn { regex } { 1 } { __regex_escape_use:nnnn }
240     \l__regex_internal_a_tl
241 }
(End definition for \_regex_escape_use:nnnn.)

```

_regex_escape_loop:N _regex_escape_loop:N reads one character: if it is special (space, backslash, or end-marker), perform the associated action, otherwise it is simply an unescaped character. After a backslash, the same is done, but unknown characters are “escaped”.

```

242 \cs_new:Npn \_regex_escape_loop:N #1
243 {
244     \cs_if_exist_use:cF { __regex_escape\_token_to_str:N #1:w }
245     { \_regex_escape_unescaped:N #1 }
246     \_regex_escape_loop:N
247 }
248 \cs_new_nopar:cpn { __regex_escape_ \c_backslash_str :w }
249     \_regex_escape_loop:N #1
250 {
251     \cs_if_exist_use:cF { __regex_escape\_token_to_str:N #1:w }
252     { \_regex_escape_escaped:N #1 }
253     \_regex_escape_loop:N
254 }
(End definition for \_regex_escape_loop:N.)

```

_regex_escape_unescaped:N Those functions are never called before being given a new meaning, so their definitions here don’t matter.
_regex_escape_escaped:N
_regex_escape_raw:N

```

255 \cs_new_eq:NN \_regex_escape_unescaped:N ?
256 \cs_new_eq:NN \_regex_escape_escaped:N ?
257 \cs_new_eq:NN \_regex_escape_raw:N ?
(End definition for \_regex_escape_unescaped:N, \_regex_escape_escaped:N, and \_regex_escape_raw:N.)

```

_regex_escape_break:w The loop is ended upon seeing the end-marker “break”, with an error if the string ended in a backslash. Spaces are ignored, and \a, \e, \f, \n, \r, \t take their meaning here.
_regex_escape_/break:w

```

258 \cs_new_eq:NN \_regex_escape_break:w \_prg_break:
259 \cs_new_nopar:cpn { __regex_escape_/break:w }
260 {
261     \if_false: { \fi: }
262     \_msg_kernel_error:nn { regex } { trailing-backslash }
263     \exp_after:wN \use_none:n \exp_after:wN { \if_false: } \fi:
264 }
265 \cs_new_nopar:cpn { __regex_escape_~:w } { }
266 \cs_new_nopar:cpx { __regex_escape_/a:w }
267     { \exp_not:N \_regex_escape_raw:N \iow_char:N ^^G }
268 \cs_new_nopar:cpx { __regex_escape_/t:w }
269     { \exp_not:N \_regex_escape_raw:N \iow_char:N ^^I }
270 \cs_new_nopar:cpx { __regex_escape_/n:w }
271     { \exp_not:N \_regex_escape_raw:N \iow_char:N ^^J }
272 \cs_new_nopar:cpx { __regex_escape_/f:w }

```

```

273 { \exp_not:N \__regex_escape_raw:N \iow_char:N \^^L }
274 \cs_new_nopar:cpx { __regex_escape_/r:w }
275 { \exp_not:N \__regex_escape_raw:N \iow_char:N \^^M }
276 \cs_new_nopar:cpx { __regex_escape_/e:w }
277 { \exp_not:N \__regex_escape_raw:N \iow_char:N \^^[ }
(End definition for \__regex_escape_break:w and others.)

```

```

\__regex_escape_/x:w
\__regex_escape_x_end:w
\__regex_escape_x_large:n

```

When `\x` is encountered, `__regex_escape_x_test:N` is responsible for grabbing some hexadecimal digits, and feeding the result to `__regex_escape_x_end:w`. If the number is < 256 , then it is turned into a byte and fed to `__regex_escape_raw:N`. Otherwise, interrupt the assignment, and either produce an error, or use a standard `\lowercase` trick depending on the precise value.

```

278 \cs_new:cpn { __regex_escape_/x:w } \__regex_escape_loop:N
279 {
280   \exp_after:wN \__regex_escape_x_end:w
281   \__int_value:w "0 \__regex_escape_x_test:N
282 }
283 \cs_new:Npn \__regex_escape_x_end:w #1 ;
284 {
285   \int_compare:nNnTF {#1} < \c_two_hundred_fifty_six
286   {
287     \exp_last_unbraced:Nf \__regex_escape_raw:N
288     { \__str_output_byte:n {#1} }
289   }
290   { \__regex_escape_x_large:n {#1} }
291 }
292 \group_begin:
293 \char_set_catcode_other:n { 0 }
294 \cs_new:Npn \__regex_escape_x_large:n #1
295 {
296   \if_false: { \fi: }
297   \__tl_build_one:o \l__regex_internal_b_tl
298   \int_compare:nNnTF {#1} > \c_max_char_int
299   {
300     \__msg_kernel_error:nnx { regex } { x-overflow } {#1}
301     \tl_set:Nx \l__regex_internal_b_tl
302     { \if_false: } \fi:
303   }
304   {
305     \char_set_lccode:nn { \c_zero } {#1}
306     \tl_to_lowercase:n
307     {
308       \tl_set:Nx \l__regex_internal_b_tl
309       { \if_false: } \fi:
310       \__regex_escape_raw:N ^^@
311     }
312   }
313 }
314 \group_end:

```

(End definition for `_regex_escape_/x:w`.)

`_regex_escape_x_test:N` Find out whether the first character is a left brace (allowing any number of hexadecimal digits), or not (allowing up to two hexadecimal digits). We need to check for the end-of-string marker. Eventually, call either `_regex_escape_x_loop:N` or `_regex_escape_x:N`.

```
315 \cs_new:Npn \_regex_escape_x_test:N #1
316 {
317   \str_if_eq_x:nnTF {#1} { break } { ; }
318   {
319     \if_charcode:w \c_space_token #1
320     \exp_after:wN \_regex_escape_x_test:N
321     \else:
322       \exp_after:wN \_regex_escape_x_test_two:N
323       \exp_after:wN #1
324     \fi:
325   }
326 }
327 \cs_new:Npn \_regex_escape_x_test_two:N #1
328 {
329   \if_charcode:w \c_left_brace_str #1
330   \exp_after:wN \_regex_escape_x_loop:N
331   \else:
332     \_str_hexadecimal_use:NTF #1
333     { \exp_after:wN \_regex_escape_x:N }
334     { ; \exp_after:wN \_regex_escape_loop:N \exp_after:wN #1 }
335   \fi:
336 }
```

(End definition for `_regex_escape_x_test:N`.)

`_regex_escape_x:N` This looks for the second digit in the unbraced case.

```
337 \cs_new:Npn \_regex_escape_x:N #1
338 {
339   \str_if_eq_x:nnTF {#1} { break } { ; }
340   {
341     \_str_hexadecimal_use:NTF #1
342     { ; \_regex_escape_loop:N }
343     { ; \_regex_escape_loop:N #1 }
344   }
345 }
```

(End definition for `_regex_escape_x:N`.)

`_regex_escape_x_loop:N` Grab hexadecimal digits, skip spaces, and at the end, check that there is a right brace, otherwise raise an error outside the assignment.

```
346 \cs_new:Npn \_regex_escape_x_loop:N #1
347 {
348   \_str_hexadecimal_use:NTF #1
349   { \_regex_escape_x_loop:N }
350   {
```



```

351     \token_if_eq_charcode:NNTF \c_space_token #1
352     { \__regex_escape_x_loop:N }
353     {
354         ;
355         \exp_after:wN \token_if_eq_charcode:NNTF \c_right_brace_str #1
356         { \__regex_escape_loop:N }
357         {
358             \if_false: { \fi: }
359             \__tl_build_one:o \l__regex_internal_b_tl
360             \__msg_kernel_error:nn { regex } { x-missing-rbrace } {#1}
361             \tl_set:Nx \l__regex_internal_b_tl
362             { \if_false: } \fi: \__regex_escape_loop:N #1
363         }
364     }
365 }
366 }

```

(End definition for __regex_escape_x_loop:N.)

__regex_char_if_alphanumeric:NNTF
 __regex_char_if_special:NNTF

These two tests are used in the first pass when parsing a regular expression. That pass is responsible for finding escaped and non-escaped characters, and recognizing which ones have special meanings and which should be interpreted as “raw” characters. Namely,

- alphanumerics are “raw” if they are not escaped, and may have a special meaning when escaped;
- non-alphanumeric printable ascii characters are “raw” if they are escaped, and may have a special meaning when not escaped;
- characters other than printable ascii are always “raw”.

The code is ugly, and highly based on magic numbers and the ascii codes of characters. This is mostly unavoidable for performance reasons. Maybe the tests can be optimized a little bit more. Here, “alphanumeric” means 0–9, A–Z, a–z; “special” character means non-alphanumeric but printable ascii, from space (hex 20) to del (hex 7E).

```

367 \prg_new_conditional:Npnn \__regex_char_if_special:N #1 { TF }
368 {
369     \if_int_compare:w '#1 < \c_ninety_one
370     \if_int_compare:w '#1 < \c_fifty_eight
371     \if_int_compare:w '#1 < \c_forty_eight
372     \if_int_compare:w '#1 < \c_thirty_two
373     \prg_return_false: \else: \prg_return_true: \fi:
374     \else: \prg_return_false: \fi:
375     \else:
376     \if_int_compare:w '#1 < \c_sixty_five
377     \prg_return_true: \else: \prg_return_false: \fi:
378     \fi:
379     \else:
380     \if_int_compare:w '#1 < \c_one_hundred_twenty_three
381     \if_int_compare:w '#1 < \c_ninety_seven
382     \prg_return_true: \else: \prg_return_false: \fi:

```

```

383     \else:
384         \if_int_compare:w '#1 < \c_one_hundred_twenty_seven
385         \prg_return_true: \else: \prg_return_false: \fi:
386     \fi:
387 \fi:
388 }
389 \prg_new_conditional:Npnn \__regex_char_if_alphanumeric:N #1 { TF }
390 {
391     \if_int_compare:w '#1 < \c_ninety_one
392     \if_int_compare:w '#1 < \c_fifty_eight
393     \if_int_compare:w '#1 < \c_forty_eight
394     \prg_return_false: \else: \prg_return_true: \fi:
395 \else:
396     \if_int_compare:w '#1 < \c_sixty_five
397     \prg_return_false: \else: \prg_return_true: \fi:
398 \fi:
399 \else:
400     \if_int_compare:w '#1 < \c_one_hundred_twenty_three
401     \if_int_compare:w '#1 < \c_ninety_seven
402     \prg_return_false: \else: \prg_return_true: \fi:
403 \else:
404     \prg_return_false:
405 \fi:
406 \fi:
407 }

```

(End definition for __regex_char_if_alphanumeric:N and __regex_char_if_special:N.)

2.3 Compiling

A regular expression starts its life as a string of characters. In this section, we convert it to internal instructions, resulting in a “compiled” regular expression. This compiled expression is then turned into states of an automaton in the building phase. Compiled regular expressions consist of the following:

- __regex_class:NnnN *<boolean>* {*<tests>*} {*<min>*} {*<more>*} *<lazyness>*
- __regex_group:nnnN {*<branches>*} {*<min>*} {*<more>*} *<lazyness>*, also __regex_group_no_capture:nnnN and __regex_group_resetting:nnnN with the same syntax.
- __regex_branch:n {*<contents>*}
- __regex_command_K:
- __regex_assertion:Nn *<boolean>* {*<assertion test>*}, where the *<assertion test>* is __regex_b_test: or __regex_anchor:N *<integer>*}

Tests can be the following:

- __regex_item_caseful_equal:n {*<char code>*}

- `__regex_item_caseless_equal:n` $\{\langle char\ code\rangle\}$
- `__regex_item_caseful_range:nn` $\{\langle min\rangle\}\{\langle max\rangle\}$
- `__regex_item_caseless_range:nn` $\{\langle min\rangle\}\{\langle max\rangle\}$
- `__regex_item_catcode:nT` $\{\langle catcode\ bitmap\rangle\}\{\langle tests\rangle\}$
- `__regex_item_catcode_reverse:nT` $\{\langle catcode\ bitmap\rangle\}\{\langle tests\rangle\}$
- `__regex_item_reverse:n` $\{\langle tests\rangle\}$
- `__regex_item_exact:nn` $\{\langle catcode\rangle\}\{\langle char\ code\rangle\}$
- `__regex_item_exact_cs:c` $\{\langle csname\rangle\}$
- `__regex_item_cs:n` $\{\langle compiled\ regex\rangle\}$

2.3.1 Variables used when compiling

`\l__regex_group_level_int` We make sure to open the same number of groups as we close.

```
408 \int_new:N \l__regex_group_level_int
```

(End definition for `\l__regex_group_level_int`. This variable is documented on page ??.)

`\l__regex_mode_int` While compiling, ten modes are recognized, labelled $-63, -23, -6, -2, 0, 2, 3, 6, 23, 63$. See section 2.3.3.

```
409 \int_new:N \l__regex_mode_int
```

(End definition for `\l__regex_mode_int`. This variable is documented on page ??.)

`\l__regex_catcodes_int` We wish to allow constructions such as `\c[^BE](. . \cL[a-z] . .)`, where the outer catcode test applies to the whole group, but is superseded by the inner catcode test. For this to work, we need to keep track of lists of allowed category codes: `\l__regex_catcodes_int` and `\l__regex_default_catcodes_int` are bitmaps, sums of 4^c , for all allowed catcodes c . The latter is local to each capturing group, and we reset `\l__regex_catcodes_int` to that value after each character or class, changing it only when encountering a `\c` escape. The boolean records whether the list of categories of a catcode test has to be inverted: compare `\c[^BE]` and `\c[BE]`.

```
410 \int_new:N \l__regex_catcodes_int
```

```
411 \int_new:N \l__regex_default_catcodes_int
```

```
412 \bool_new:N \l__regex_catcodes_bool
```

(End definition for `\l__regex_catcodes_int` and `\l__regex_default_catcodes_int`. These variables are documented on page ??.)

`\c__regex_catcode_C_int` Constants: 4^c for each category, and the sum of all powers of 4.

```
413 \int_const:Nn \c__regex_catcode_C_int { "1 }
```

```
414 \int_const:Nn \c__regex_catcode_B_int { "4 }
```

```
415 \int_const:Nn \c__regex_catcode_E_int { "10 }
```

```
416 \int_const:Nn \c__regex_catcode_M_int { "40 }
```

```
417 \int_const:Nn \c__regex_catcode_T_int { "100 }
```

```
418 \int_const:Nn \c__regex_catcode_P_int { "1000 }
```

`\c__regex_catcode_D_int`

`\c__regex_catcode_S_int`

`\c__regex_catcode_L_int`

`\c__regex_catcode_O_int`

`\c__regex_catcode_A_int`

`\c__regex_all_catcodes_int`

```

419 \int_const:Nn \c__regex_catcode_U_int { "4000 }
420 \int_const:Nn \c__regex_catcode_D_int { "10000 }
421 \int_const:Nn \c__regex_catcode_S_int { "100000 }
422 \int_const:Nn \c__regex_catcode_L_int { "400000 }
423 \int_const:Nn \c__regex_catcode_O_int { "1000000 }
424 \int_const:Nn \c__regex_catcode_A_int { "4000000 }
425 \int_const:Nn \c__regex_all_catcodes_int { "5515155 }

```

(End definition for \c__regex_catcode_C_int and others. These variables are documented on page ??.)

`\l__regex_internal_regex` The compilation step stores its result in this variable.

```

426 \cs_new_eq:NN \l__regex_internal_regex \c__regex_no_match_regex

```

(End definition for \l__regex_internal_regex. This variable is documented on page ??.)

`\l__regex_show_prefix_seq` This sequence holds the prefix that makes up the line displayed to the user. The various items must be removed from the right, which is tricky with a token list, hence we use a sequence.

```

427 \seq_new:N \l__regex_show_prefix_seq

```

(End definition for \l__regex_show_prefix_seq. This variable is documented on page ??.)

`\l__regex_show_lines_int` A hack. To know whether a given class has a single item in it or not, we count the number of lines when showing the class.

```

428 \int_new:N \l__regex_show_lines_int

```

(End definition for \l__regex_show_lines_int. This variable is documented on page ??.)

2.3.2 Generic helpers used when compiling

`__regex_get_digits:NTFw` If followed by some raw digits, collect them one by one in the integer variable #1, and
`__regex_get_digits_loop:w` take the true branch. Otherwise, take the false branch.

```

429 \cs_new_protected:Npn \__regex_get_digits:NTFw #1#2#3#4#5
430 {
431   \__regex_if_raw_digit:NNTF #4 #5
432   { #1 = #5 \__regex_get_digits_loop:nw {#2} }
433   { #3 #4 #5 }
434 }
435 \cs_new:Npn \__regex_get_digits_loop:nw #1#2#3
436 {
437   \__regex_if_raw_digit:NNTF #2 #3
438   { #3 \__regex_get_digits_loop:nw {#1} }
439   { \scan_stop: #1 #2 #3 }
440 }

```

(End definition for __regex_get_digits:NTFw.)

`__regex_if_raw_digit:NNTF` Test used when grabbing digits for the {m,n} quantifier. It only accepts non-escaped digits.

```

441 \prg_new_conditional:Npnn \__regex_if_raw_digit:NN #1#2 { TF }
442 {
443   \if_meaning:w \__regex_compile_raw:N #1
444   \if_int_compare:w \c_one < 1 #2 \exp_stop_f:

```

```

445         \prg_return_true:
446     \else:
447         \prg_return_false:
448     \fi:
449 \else:
450     \prg_return_false:
451 \fi:
452 }

```

(End definition for `_regex_if_raw_digit:NNTF`.)

2.3.3 Mode

When compiling the NFA corresponding to a given regex string, we can be in ten distinct modes, which we label by some magic numbers:

- 6 `[\c{...}]` control sequence in a class,
- 2 `\c{...}` control sequence,
- 0 ... outer,
- 2 `\c...` catcode test,
- 6 `[\c...]` catcode test in a class,
- 63 `[\c{[...]}]` class inside mode -6,
- 23 `\c{[...]}` class inside mode -2,
- 3 `[...]` class inside mode -3,
- 23 `\c[...]` class inside mode 2,
- 63 `[\c[...]]` class inside mode 6.

This list is exhaustive, because `\c` escape sequences cannot be nested, and character classes cannot be nested directly. The choice of numbers is such as to optimize the most useful tests, and make transitions from one mode to another as simple as possible.

- Even modes mean that we are not directly in a character class. In this case, a left bracket appends 3 to the mode. In a character class, a right bracket changes the mode as $m \rightarrow (m - 15)/13$, truncated.
- Grouping, assertion, and anchors are allowed in non-positive even modes (0, -2, -6), and do not change the mode. Otherwise, they trigger an error.
- A left bracket is special in even modes, appending 3 to the mode; in those modes, quantifiers and the dot are recognized, and the right bracket is normal. In odd modes (within classes), the left bracket is normal, but the right bracket ends the class, changing the mode from m to $(m - 15)/13$, truncated; also, ranges are recognized.

- In non-negative modes, left and right braces are normal. In negative modes, however, left braces trigger a warning; right braces end the control sequence, going from -2 to 0 or -6 to 3 , with error recovery for odd modes.
- Properties (such as the `\d` character class) can appear in any mode.

`_regex_if_in_class:TF` Test whether we are directly in a character class (at the innermost level of nesting). There, many escape sequences are not recognized, and special characters are normal. Also, for every raw character, we must look ahead for a possible raw dash.

```

453 \cs_new_nopar:Npn \_regex_if_in_class:TF
454 {
455   \if_int_odd:w \l__regex_mode_int
456   \exp_after:wN \use_i:nn
457   \else:
458   \exp_after:wN \use_ii:nn
459   \fi:
460 }

```

(End definition for _regex_if_in_class:TF.)

`_regex_if_in_cs:TF` Right braces are special only directly inside control sequences (at the inner-most level of nesting, not counting groups).

```

461 \cs_new_nopar:Npn \_regex_if_in_cs:TF
462 {
463   \if_int_odd:w \l__regex_mode_int
464   \exp_after:wN \use_ii:nn
465   \else:
466   \if_int_compare:w \l__regex_mode_int < \c_zero
467   \exp_after:wN \exp_after:wN \exp_after:wN \use_i:nn
468   \else:
469   \exp_after:wN \exp_after:wN \exp_after:wN \use_ii:nn
470   \fi:
471   \fi:
472 }

```

(End definition for _regex_if_in_cs:TF.)

`_regex_if_in_class_or_catcode:TF` Assertions are only allowed in modes 0 , -2 , and -6 , *i.e.*, even, non-positive modes.

```

473 \cs_new_nopar:Npn \_regex_if_in_class_or_catcode:TF
474 {
475   \if_int_odd:w \l__regex_mode_int
476   \exp_after:wN \use_i:nn
477   \else:
478   \if_int_compare:w \l__regex_mode_int > \c_zero
479   \exp_after:wN \exp_after:wN \exp_after:wN \use_i:nn
480   \else:
481   \exp_after:wN \exp_after:wN \exp_after:wN \use_ii:nn
482   \fi:
483   \fi:
484 }

```

(End definition for _regex_if_in_class_or_catcode:TF.)

`_regex_if_within_catcode:TF` This test takes the true branch if we are in a catcode test, either immediately following it (modes 2 and 6) or in a class on which it applies (modes 23 and 63). This is used to tweak how left brackets behave in modes 2 and 6.

```

485 \cs_new_nopar:Npn \_regex_if_within_catcode:TF
486 {
487   \if_int_compare:w \l__regex_mode_int > \c_zero
488     \exp_after:wN \use_i:nn
489   \else:
490     \exp_after:wN \use_ii:nn
491   \fi:
492 }

```

(End definition for _regex_if_within_catcode:TF.)

`_regex_chk_c_allowed:T` The `\c` escape sequence is only allowed in modes 0 and 3, *i.e.*, not within any other `\c` escape sequence.

```

493 \cs_new_protected:Npn \_regex_chk_c_allowed:T
494 {
495   \if_int_compare:w \l__regex_mode_int = \c_zero
496     \exp_after:wN \use:n
497   \else:
498     \if_int_compare:w \l__regex_mode_int = \c_three
499       \exp_after:wN \exp_after:wN \exp_after:wN \use:n
500     \else:
501       \__msg_kernel_error:nn { regex } { c-bad-mode }
502       \exp_after:wN \exp_after:wN \exp_after:wN \use_none:n
503     \fi:
504   \fi:
505 }

```

(End definition for _regex_chk_c_allowed:T.)

`_regex_mode_quit:c:` This function changes the mode as it is needed just after a catcode test.

```

506 \cs_new_protected:Npn \_regex_mode_quit:c:
507 {
508   \if_int_compare:w \l__regex_mode_int = \c_two
509     \l__regex_mode_int = \c_zero
510   \else:
511     \if_int_compare:w \l__regex_mode_int = \c_six
512       \l__regex_mode_int = \c_three
513     \fi:
514   \fi:
515 }

```

(End definition for _regex_mode_quit:c:.)

2.3.4 Framework

`_regex_compile:w` Used when compiling a user regex or a regex for the `\c{...}` escape sequence within
`_regex_compile_end:` another regex. Start building a token list within a group (with x-expansion at the outset), and set a few variables (group level, catcodes), then start the first branch. At the end,

make sure there are no dangling classes nor groups, close the last branch: we are done building \l__regex_internal_regex.

```

516 \cs_new_protected_nopar:Npn \__regex_compile:w
517 {
518   \__tl_build_x:Nw \l__regex_internal_regex
519   \int_zero:N \l__regex_group_level_int
520   \int_set_eq:NN \l__regex_default_catcodes_int \c__regex_all_catcodes_int
521   \int_set_eq:NN \l__regex_catcodes_int \l__regex_default_catcodes_int
522   \cs_set_nopar:Npn \__regex_item_equal:n { \__regex_item_caseful_equal:n }
523   \cs_set_nopar:Npn \__regex_item_range:nn { \__regex_item_caseful_range:nn }
524   \__tl_build_one:n { \__regex_branch:n { \if_false: } \fi: }
525 }
526 \cs_new_protected_nopar:Npn \__regex_compile_end:
527 {
528   \__regex_if_in_class:TF
529   {
530     \__msg_kernel_error:nn { regex } { missing-rbrack }
531     \use:c { __regex_compile_]: }
532     \prg_do_nothing: \prg_do_nothing:
533   }
534   { }
535   \if_int_compare:w \l__regex_group_level_int > \c_zero
536     \__msg_kernel_error:nnx { regex } { missing-rparen }
537     { \int_use:N \l__regex_group_level_int }
538     \prg_replicate:nn
539     { \l__regex_group_level_int }
540     {
541       \__tl_build_one:n
542       {
543         \if_false: { \fi: }
544         \if_false: { \fi: } { 1 } { 0 } \c_true_bool
545       }
546       \__tl_build_end:
547       \__tl_build_one:o \l__regex_internal_regex
548     }
549     \fi:
550     \__tl_build_one:n { \if_false: { \fi: } }
551     \__tl_build_end:
552   }

```

(End definition for __regex_compile:w and __regex_compile_end:.)

__regex_compile:n The compilation is done between __regex_compile:w and __regex_compile_end:, starting in mode 0. Then __regex_escape_use:nnnn distinguishes special characters, escaped alphanumerics, and raw characters, interpreting \a, \x and other sequences. The 4 trailing \prg_do_nothing: are needed because some functions defined later look up to 4 tokens ahead. Before ending, make sure that any \c{...} is properly closed.

```

553 \cs_new_protected:Npn \__regex_compile:n #1
554 {
555   \__regex_compile:w

```



```

556 \int_set:Nn \tex_escapechar:D { 92 }
557 \int_set_eq:NN \l__regex_mode_int \c_zero
558 \__regex_escape_use:nnnn
559 {
560   \__regex_char_if_special:NTF ##1
561   \__regex_compile_special:N \__regex_compile_raw:N ##1
562 }
563 {
564   \__regex_char_if_alphanumeric:NTF ##1
565   \__regex_compile_escaped:N \__regex_compile_raw:N ##1
566 }
567 { \__regex_compile_raw:N ##1 }
568 { #1 }
569 \prg_do_nothing: \prg_do_nothing:
570 \prg_do_nothing: \prg_do_nothing:
571 \int_compare:nNnT \l__regex_mode_int < \c_zero
572 {
573   \msg_kernel_error:nn { regex } { c-missing-rbrace }
574   \__regex_compile_end:
575   \__regex_compile_one:x
576   { \__regex_item_cs:n { \exp_not:o \l__regex_internal_regex } }
577   \prg_do_nothing: \prg_do_nothing:
578   \prg_do_nothing: \prg_do_nothing:
579 }
580 \__regex_compile_end:
581 }

```

(End definition for __regex_compile:n.)

__regex_compile_escaped:N
 __regex_compile_special:N

If the special character or escaped alphanumeric has a particular meaning in regexes, the corresponding function is used. Otherwise, it is interpreted as a raw character. We distinguish special characters from escaped alphanumeric characters because they behave differently when appearing as an end-point of a range.

```

582 \cs_new_protected:Npn \__regex_compile_special:N #1
583 {
584   \cs_if_exist_use:cF { __regex_compile_#1: }
585   { \__regex_compile_raw:N #1 }
586 }
587 \cs_new_protected:Npn \__regex_compile_escaped:N #1
588 {
589   \cs_if_exist_use:cF { __regex_compile_/#1: }
590   { \__regex_compile_raw:N #1 }
591 }

```

(End definition for __regex_compile_escaped:N and __regex_compile_special:N.)

__regex_compile_one:x

This is used after finding one “test”, such as \d, or a raw character. If that followed a catcode test (e.g., \cL), then restore the mode. If we are not in a class, then the test is “standalone”, and we need to add __regex_class:NnnnN and search for quantifiers. In any case, insert the test, possibly together with a catcode test if appropriate.

```

592 \cs_new_protected:Npn \__regex_compile_one:x #1

```

```

593 {
594   \_regex_mode_quit_c:
595   \_regex_if_in_class:TF { }
596   {
597     \_tl_build_one:n
598     { \_regex_class:NnnnN \c_true_bool { \if_false: } \fi: }
599   }
600   \_tl_build_one:x
601   {
602     \if_int_compare:w \l__regex_catcodes_int < \c__regex_all_catcodes_int
603     \_regex_item_catcode:nT { \int_use:N \l__regex_catcodes_int }
604     { \exp_not:N \exp_not:n {#1} }
605     \else:
606     \exp_not:N \exp_not:n {#1}
607     \fi:
608   }
609   \int_set_eq:NN \l__regex_catcodes_int \l__regex_default_catcodes_int
610   \_regex_if_in_class:TF { } { \_regex_compile_quantifier:w }
611 }

```

(End definition for _regex_compile_one:x.)

_regex_compile_abort_tokens:n This function places the collected tokens back in the input stream, each as a raw character.
_regex_compile_abort_tokens:x Spaces are not preserved.

```

612 \cs_new_protected:Npn \_regex_compile_abort_tokens:n #1
613 {
614   \use:x
615   {
616     \exp_args:No \tl_map_function:nN { \tl_to_str:n {#1} }
617     \_regex_compile_raw:N
618   }
619 }
620 \cs_generate_variant:Nn \_regex_compile_abort_tokens:n { x }

```

(End definition for _regex_compile_abort_tokens:n and _regex_compile_abort_tokens:x.)

2.3.5 Quantifiers

_regex_compile_quantifier:w This looks ahead and finds any quantifier (special character equal to either of ?+*{).

```

621 \cs_new_protected:Npn \_regex_compile_quantifier:w #1#2
622 {
623   \token_if_eq_meaning:NNTF #1 \_regex_compile_special:N
624   {
625     \cs_if_exist_use:cF { \_regex_compile_quantifier_#2:w }
626     { \_regex_compile_quantifier_none: #1 #2 }
627   }
628   { \_regex_compile_quantifier_none: #1 #2 }
629 }

```

(End definition for _regex_compile_quantifier:w.)

`_regex_compile_quantifier_none:`
`_regex_compile_quantifier_abort:xNN`

Those functions are called whenever there is no quantifier, or a braced construction is invalid (equivalent to no quantifier, and whatever characters were grabbed are left raw).

```
630 \cs_new_protected:Npn \__regex_compile_quantifier_none:
631   { \__tl_build_one:n { \if_false: { \fi: } { 1 } { 0 } \c_false_bool } }
632 \cs_new_protected:Npn \__regex_compile_quantifier_abort:xNN #1#2#3
633   {
634     \__regex_compile_quantifier_none:
635     \__msg_kernel_warning:nnxx { regex } { invalid-quantifier } {#1} {#3}
636     \__regex_compile_abort_tokens:x {#1}
637     #2 #3
638   }
```

(End definition for `__regex_compile_quantifier_none:.`)

`_regex_compile_quantifier_lazy:nnNN`

Once the “main” quantifier (`?`, `*`, `+` or a braced construction) is found, we check whether it is lazy (followed by a question mark). We then add to the compiled regex a closing brace (ending `__regex_class:NnnnN` and friends), the start-point of the range, its end-point, and a boolean, true for lazy and false for greedy operators.

```
639 \cs_new_protected:Npn \__regex_compile_quantifier_lazy:nnNN #1#2#3#4
640   {
641     \str_if_eq:nnTF { #3 #4 } { \__regex_compile_special:N ? }
642     { \__tl_build_one:n { \if_false: { \fi: } { #1 } { #2 } \c_true_bool } }
643     {
644       \__tl_build_one:n { \if_false: { \fi: } { #1 } { #2 } \c_false_bool }
645       #3 #4
646     }
647   }
```

(End definition for `__regex_compile_quantifier_lazy:nnNN.`)

`_regex_compile_quantifier_?:w`
`_regex_compile_quantifier_*:w`
`_regex_compile_quantifier_+:w`

For each “basic” quantifier, `?`, `*`, `+`, feed the correct arguments to `__regex_compile_quantifier_lazy:nnNN`, `-1` means that there is no upper bound on the number of repetitions.

```
648 \cs_new_protected_nopar:cpn { \__regex_compile_quantifier_?:w }
649   { \__regex_compile_quantifier_lazy:nnNN { 0 } { 1 } }
650 \cs_new_protected_nopar:cpn { \__regex_compile_quantifier_*:w }
651   { \__regex_compile_quantifier_lazy:nnNN { 0 } { -1 } }
652 \cs_new_protected_nopar:cpn { \__regex_compile_quantifier_+:w }
653   { \__regex_compile_quantifier_lazy:nnNN { 1 } { -1 } }
```

(End definition for `_regex_compile_quantifier_?:w`, `_regex_compile_quantifier_*:w`, and `_regex_compile_quantifier_+:w`.)

`_regex_compile_quantifier_{:w`
`_regex_compile_quantifier_braced_auxi:w`
`_regex_compile_quantifier_braced_auxii:w`
`_regex_compile_quantifier_braced_auxiii:w`

Three possible syntaxes: `{⟨int⟩}`, `{⟨int⟩,}`, or `{⟨int⟩,⟨int⟩}`. Any other syntax causes us to abort and put whatever we collected back in the input stream, as raw characters, including the opening brace. Grab a number into `\l__regex_internal_a_int`. If the number is followed by a right brace, the range is `[a, a]`. If followed by a comma, grab one more number, and call the `_ii` or `_iii` auxiliary. Those auxiliaries check for a closing brace, leading to the range `[a, ∞]` or `[a, b]`, encoded as `{a}{-1}` and `{a}{b-a}`.

```
654 \cs_new_protected:cpn { \__regex_compile_quantifier_ \c_left_brace_str :w }
655   {
656     \__regex_get_digits:NTFw \l__regex_internal_a_int
```

```

657     { \_regex_compile_quantifier_braced_auxi:w }
658     { \_regex_compile_quantifier_abort:xNN { \c_left_brace_str } }
659 }
660 \cs_new_protected:Npn \_regex_compile_quantifier_braced_auxi:w #1#2
661 {
662     \str_case_x:nnn { #1 #2 }
663     {
664         { \_regex_compile_special:N \c_right_brace_str }
665         {
666             \exp_args:No \_regex_compile_quantifier_lazyness:nnNN
667             { \int_use:N \l__regex_internal_a_int } { 0 }
668         }
669         { \_regex_compile_special:N , }
670         {
671             \_regex_get_digits:NTFw \l__regex_internal_b_int
672             { \_regex_compile_quantifier_braced_auxiii:w }
673             { \_regex_compile_quantifier_braced_auxii:w }
674         }
675     }
676     {
677         \_regex_compile_quantifier_abort:xNN
678         { \c_left_brace_str \int_use:N \l__regex_internal_a_int }
679         #1 #2
680     }
681 }
682 \cs_new_protected:Npn \_regex_compile_quantifier_braced_auxii:w #1#2
683 {
684     \str_if_eq_x:nnTF
685     { #1 #2 } { \_regex_compile_special:N \c_right_brace_str }
686     {
687         \exp_args:No \_regex_compile_quantifier_lazyness:nnNN
688         { \int_use:N \l__regex_internal_a_int } { -1 }
689     }
690     {
691         \_regex_compile_quantifier_abort:xNN
692         { \c_left_brace_str \int_use:N \l__regex_internal_a_int , }
693         #1 #2
694     }
695 }
696 \cs_new_protected:Npn \_regex_compile_quantifier_braced_auxiii:w #1#2
697 {
698     \str_if_eq_x:nnTF
699     { #1 #2 } { \_regex_compile_special:N \c_right_brace_str }
700     {
701         \if_int_compare:w \l__regex_internal_a_int > \l__regex_internal_b_int
702         \_msg_kernel_error:nnxx { regex } { backwards-quantifier }
703         { \int_use:N \l__regex_internal_a_int }
704         { \int_use:N \l__regex_internal_b_int }
705         \int_zero:N \l__regex_internal_b_int
706         \else:

```

```

707         \int_sub:Nn \l__regex_internal_b_int \l__regex_internal_a_int
708     \fi:
709     \exp_args:Noo \__regex_compile_quantifier_lazy:nnNN
710     { \int_use:N \l__regex_internal_a_int }
711     { \int_use:N \l__regex_internal_b_int }
712 }
713 {
714     \__regex_compile_quantifier_abort:xNN
715     {
716         \c_left_brace_str
717         \int_use:N \l__regex_internal_a_int ,
718         \int_use:N \l__regex_internal_b_int
719     }
720     #1 #2
721 }
722 }

```

(End definition for __regex_compile_quantifier_{:w}.)

2.3.6 Raw characters

__regex_compile_raw_error:N Within character classes, and following catcode tests, some escaped alphanumeric sequences such as `\b` do not have any meaning. They are replaced by a raw character, after spitting out an error.

```

723 \cs_new_protected:Npn \__regex_compile_raw_error:N #1
724 {
725     \__msg_kernel_error:nnx { regex } { bad-escape } {#1}
726     \__regex_compile_raw:N #1
727 }

```

(End definition for __regex_compile_raw_error:N.)

__regex_compile_raw:N If we are in a character class and the next character is an unescaped dash, this denotes a range. Otherwise, the current character `#1` matches itself.

```

728 \cs_new_protected:Npn \__regex_compile_raw:N #1#2#3
729 {
730     \__regex_if_in_class:TF
731     {
732         \str_if_eq:nnTF {#2#3} { \__regex_compile_special:N - }
733         { \__regex_compile_range:Nw #1 }
734         {
735             \__regex_compile_one:x
736             { \__regex_item_equal:n { \__int_value:w '#1 ~ } }
737             #2 #3
738         }
739     }
740     {
741         \__regex_compile_one:x
742         { \__regex_item_equal:n { \__int_value:w '#1 ~ } }
743         #2 #3
744     }

```

```

745     }
(End definition for \_regex_compile_raw:N.)

\_regex_compile_range:Nw We have just read a raw character followed by a dash; this should be followed by an
\_regex_if_end_range:NNTF end-point for the range. Valid end-points are: any raw character; any special character,
except a right bracket. In particular, escaped characters are forbidden.

746 \prg_new_protected_conditional:Npnn \_regex_if_end_range:NN #1#2 { TF }
747 {
748   \if_meaning:w \_regex_compile_raw:N #1
749     \prg_return_true:
750   \else:
751     \if_meaning:w \_regex_compile_special:N #1
752       \if_charcode:w ] #2
753         \prg_return_false:
754       \else:
755         \prg_return_true:
756       \fi:
757     \else:
758       \prg_return_false:
759     \fi:
760   \fi:
761 }
762 \cs_new_protected:Npn \_regex_compile_range:Nw #1#2#3
763 {
764   \_regex_if_end_range:NNTF #2 #3
765   {
766     \if_int_compare:w '#1 > '#3 \exp_stop_f:
767       \_msg_kernel_error:nxxx { regex } { range-backwards } {#1} {#3}
768     \else:
769       \_tl_build_one:x
770       {
771         \if_int_compare:w '#1 = '#3 \exp_stop_f:
772           \_regex_item_equal:n
773         \else:
774           \_regex_item_range:nn { \_int_value:w '#1 ~ }
775         \fi:
776         { \_int_value:w '#3 ~ }
777       }
778     \fi:
779   }
780   {
781     \_msg_kernel_warning:nxxx { regex } { range-missing-end }
782     {#1} { \c_backslash_str #3 }
783     \_tl_build_one:x
784     {
785       \_regex_item_equal:n { \_int_value:w '#1 ~ }
786       \_regex_item_equal:n { \_int_value:w '- ~ }
787     }
788     #2#3

```

```

789     }
790 }
(End definition for \_regex_compile_range:Nw and \_regex_if_end_range:NNTF.)

```

2.3.7 Character properties

_regex_compile_.: In a class, the dot has no special meaning. Outside, insert _regex_prop_., which matches any character or control sequence, and refuses -2 (end-marker).

```

791 \cs_new_protected_nopar:cpx { \_regex_compile_.: }
792 {
793   \exp_not:N \_regex_if_in_class:TF
794   { \_regex_compile_raw:N . }
795   { \_regex_compile_one:x \exp_not:c { \_regex_prop_.: } }
796 }
797 \cs_new_protected_nopar:cpn { \_regex_prop_.: }
798 {
799   \if_int_compare:w \_regex_current_char_int > - \c_two
800   \exp_after:wN \_regex_break_true:w
801   \fi:
802 }

```

(End definition for _regex_compile_.: and _regex_prop_..)

_regex_compile_/d: The constants _regex_prop_d:, etc. hold a list of tests which match the corresponding character class, and jump to the _regex_break_point:TF marker. As for a normal character, we check for quantifiers.

```

803 \cs_set_protected:Npn \_regex_tmp:w #1#2
804 {
805   \cs_new_protected_nopar:cpx { \_regex_compile_/#1: }
806   { \_regex_compile_one:x \exp_not:c { \_regex_prop_#1: } }
807   \cs_new_protected_nopar:cpx { \_regex_compile_/#2: }
808   {
809     \_regex_compile_one:x
810     { \_regex_item_reverse:n \exp_not:c { \_regex_prop_#1: } }
811   }
812 }
813 \_regex_tmp:w d D
814 \_regex_tmp:w h H
815 \_regex_tmp:w s S
816 \_regex_tmp:w v V
817 \_regex_tmp:w w W
818 \cs_new_protected_nopar:cpn { \_regex_compile_/N: }
819 { \_regex_compile_one:x \_regex_prop_N: }

```

(End definition for _regex_compile_/d: and others.)

2.3.8 Anchoring and simple assertions

_regex_compile_anchor:NF In modes where assertions are allowed, anchor to the start of the query, the start of the match, or the end of the query, depending on the integer #1. In other modes, #2 treats

```

\_regex_compile_~:
\_regex_compile_/A:
\_regex_compile_/G:
\_regex_compile_$:
\_regex_compile_/Z:
\_regex_compile_/z:

```

the character as raw, with an error for escaped letters (\$ is valid in a class, but \A is definitely a mistake on the user's part).

```

820 \cs_new_protected:Npn \__regex_compile_anchor:NF #1#2
821 {
822   \__regex_if_in_class_or_catcode:TF {#2}
823   {
824     \__tl_build_one:n
825     { \__regex_assertion:Nn \c_true_bool { \__regex_anchor:N #1 } }
826   }
827 }
828 \cs_set_protected:Npn \__regex_tmp:w #1#2
829 {
830   \cs_new_protected_nopar:cpn { __regex_compile_/#1: }
831   { \__regex_compile_anchor:NF #2 { \__regex_compile_raw_error:N #1 } }
832 }
833 \__regex_tmp:w A \l__regex_min_pos_int
834 \__regex_tmp:w G \l__regex_start_pos_int
835 \__regex_tmp:w Z \l__regex_max_pos_int
836 \__regex_tmp:w z \l__regex_max_pos_int
837 \cs_set_protected:Npn \__regex_tmp:w #1#2
838 {
839   \cs_new_protected_nopar:cpn { __regex_compile_#1: }
840   { \__regex_compile_anchor:NF #2 { \__regex_compile_raw:N #1 } }
841 }
842 \exp_args:Nx \__regex_tmp:w { \iow_char:N ^ } \l__regex_min_pos_int
843 \exp_args:Nx \__regex_tmp:w { \iow_char:N $ } \l__regex_max_pos_int
(End definition for \__regex_compile_anchor:NF.)

```

__regex_compile_/b: Contrarily to ^ and \$, which could be implemented without really knowing what precedes in the token list, this requires more information, namely, the knowledge of the last character code.

__regex_compile_/B:

```

844 \cs_new_protected_nopar:cpn { __regex_compile_/b: }
845 {
846   \__regex_if_in_class_or_catcode:TF
847   { \__regex_compile_raw_error:N b }
848   {
849     \__tl_build_one:n
850     { \__regex_assertion:Nn \c_true_bool { \__regex_b_test: } }
851   }
852 }
853 \cs_new_protected_nopar:cpn { __regex_compile_/B: }
854 {
855   \__regex_if_in_class_or_catcode:TF
856   { \__regex_compile_raw_error:N B }
857   {
858     \__tl_build_one:n
859     { \__regex_assertion:Nn \c_false_bool { \__regex_b_test: } }
860   }
861 }

```


(End definition for `_regex_compile_/b:` and `_regex_compile_/B:.`)

2.3.9 Character classes

`_regex_compile_:` Outside a class, right brackets have no meaning. In a class, change the mode ($m \rightarrow (m - 15)/13$, truncated) to reflect the fact that we are leaving the class. Look for quantifiers, unless we are still in a class after leaving one (the case of `[... \cL[...] ...]`). quantifiers.

```

862 \cs_new_protected:cpn { \_regex_compile_}: }
863 {
864   \_regex_if_in_class:TF
865   {
866     \if_int_compare:w \l\_regex_mode_int > \c_sixteen
867     \_tl_build_one:n { \if_false: { \fi: } }
868     \fi:
869     \tex_advance:D \l\_regex_mode_int - \c_fifteen
870     \tex_divide:D \l\_regex_mode_int \c_thirteen
871     \if_int_odd:w \l\_regex_mode_int \else:
872       \exp_after:wN \_regex_compile_quantifier:w
873     \fi:
874   }
875   { \_regex_compile_raw:N ] }
876 }

```

(End definition for `_regex_compile_:.`)

`_regex_compile_[:` In a class, left brackets might introduce a POSIX character class, or mean nothing. Immediately following `\c<category>`, we must insert the appropriate catcode test, then parse the class; we pre-expand the catcode as an optimization. Otherwise (modes 0, -2 and -6) just parse the class. The mode is updated later.

```

877 \cs_new_protected_nopar:cpn { \_regex_compile_[: }
878 {
879   \_regex_if_in_class:TF
880   { \_regex_compile_class_posix_test:w }
881   {
882     \_regex_if_within_catcode:TF
883     {
884       \exp_after:wN \_regex_compile_class_catcode:w
885       \int_use:N \l\_regex_catcodes_int ;
886     }
887     { \_regex_compile_class_normal:w }
888   }
889 }

```

(End definition for `_regex_compile_[:.`)

`_regex_compile_class_normal:w` In the “normal” case, we will insert `_regex_class:NnnnN <boolean>` in the compiled code. The `<boolean>` is true for positive classes, and false for negative classes, characterized by a leading `^`. The auxiliary `_regex_compile_class:TFNN` also checks for a leading `]` which has a special meaning.

```

890 \cs_new_protected_nopar:Npn \_regex_compile_class_normal:w

```

```

891 {
892   \_regex_compile_class:TFNN
893   { \_regex_class:NnnnN \c_true_bool }
894   { \_regex_class:NnnnN \c_false_bool }
895 }

```

(End definition for _regex_compile_class_normal:w.)

_regex_compile_class_catcode:w This function is called for a left bracket in modes 2 or 6 (catcode test, and catcode test within a class). In mode 2 the whole construction needs to be put in a class (like single character). Then determine if the class is positive or negative, inserting _regex_item_catcode:nT or the reverse variant as appropriate, each with the current catcodes bitmap #1 as an argument, and reset the catcodes.

```

896 \cs_new_protected:Npn \_regex_compile_class_catcode:w #1;
897 {
898   \if_int_compare:w \l__regex_mode_int = \c_two
899     \_tl_build_one:n
900     { \_regex_class:NnnnN \c_true_bool { \if_false: } \fi: }
901   \fi:
902   \int_set_eq:NN \l__regex_catcodes_int \l__regex_default_catcodes_int
903   \_regex_compile_class:TFNN
904   { \_regex_item_catcode:nT {#1} }
905   { \_regex_item_catcode_reverse:nT {#1} }
906 }

```

(End definition for _regex_compile_class_catcode:w.)

_regex_compile_class:TFNN If the first character is ^, then the class is negative (use #2), otherwise it is positive (use #1). If the next character is a right bracket, then it should be changed to a raw one.

```

907 \cs_new_protected:Npn \_regex_compile_class:TFNN #1#2#3#4
908 {
909   \l__regex_mode_int = \__int_value:w \l__regex_mode_int 3 \exp_stop_f:
910   \str_if_eq:nnTF { #3 #4 } { \_regex_compile_special:N ^ }
911   {
912     \_tl_build_one:n { #2 { \if_false: } \fi: }
913     \_regex_compile_class:NN
914   }
915   {
916     \_tl_build_one:n { #1 { \if_false: } \fi: }
917     \_regex_compile_class:NN #3 #4
918   }
919 }
920 \cs_new_protected:Npn \_regex_compile_class:NN #1#2
921 {
922   \token_if_eq_charcode:NNTF #2 ]
923   { \_regex_compile_raw:N #2 }
924   { #1 #2 }
925 }

```

(End definition for _regex_compile_class:TFNN and _regex_compile_class:NN.)

```

\__regex_compile_class_posix_test:w
\__regex_compile_class_posix:NNNNw
\__regex_compile_class_posix_loop:w
\__regex_compile_class_posix_end:w

```

Here we check for a syntax such as [:alpha:]. We also detect [= and [. which have a meaning in POSIX regular expressions, but are not implemented in l3regex. In case we see [:, grab raw characters until hopefully reaching :]. If that's missing, or the POSIX class is unknown, abort. If all is right, add the test to the current class, with an extra __regex_item_reverse:n for negative classes.

```

926 \cs_new_protected:Npn \__regex_compile_class_posix_test:w #1#2
927 {
928   \token_if_eq_meaning:NNT \__regex_compile_special:N #1
929   {
930     \str_case:nn { #2 }
931     {
932       : { \__regex_compile_class_posix:NNNNw }
933       = { \__msg_kernel_warning:nxx { regex } { posix-unsupported } { = } }
934       . { \__msg_kernel_warning:nxx { regex } { posix-unsupported } { . } }
935     }
936   }
937   \__regex_compile_raw:N [ #1 #2
938 }
939 \cs_new_protected:Npn \__regex_compile_class_posix:NNNNw #1#2#3#4#5#6
940 {
941   \str_if_eq:nnTF { #5 #6 } { \__regex_compile_special:N ^ }
942   {
943     \bool_set_false:N \l__regex_internal_bool
944     \tl_set:Nx \l__regex_internal_a_tl { \if_false: } \fi:
945     \__regex_compile_class_posix_loop:w
946   }
947   {
948     \bool_set_true:N \l__regex_internal_bool
949     \tl_set:Nx \l__regex_internal_a_tl { \if_false: } \fi:
950     \__regex_compile_class_posix_loop:w #5 #6
951   }
952 }
953 \cs_new:Npn \__regex_compile_class_posix_loop:w #1#2
954 {
955   \token_if_eq_meaning:NNTF \__regex_compile_raw:N #1
956   { #2 \__regex_compile_class_posix_loop:w }
957   { \if_false: { \fi: } \__regex_compile_class_posix_end:w #1 #2 }
958 }
959 \cs_new_protected:Npn \__regex_compile_class_posix_end:w #1#2#3#4
960 {
961   \str_if_eq:nnTF { #1 #2 #3 #4 }
962   { \__regex_compile_special:N : \__regex_compile_special:N ] }
963   {
964     \cs_if_exist:cTF { __regex_posix_ \l__regex_internal_a_tl : }
965     {
966       \__regex_compile_one:x
967       {
968         \bool_if:NF \l__regex_internal_bool \__regex_item_reverse:n
969         \exp_not:c { __regex_posix_ \l__regex_internal_a_tl : }

```

```

970     }
971   }
972   {
973     \_msg_kernel_warning:nxx { regex } { posix-unknown }
974     { \l__regex_internal_a_tl }
975     \_regex_compile_abort_tokens:x
976     {
977       [: \bool_if:NF \l__regex_internal_bool { ^ }
978       \l__regex_internal_a_tl :]
979     }
980   }
981 }
982 {
983   \_msg_kernel_error:nxxx { regex } { posix-missing-close }
984   { [: \l__regex_internal_a_tl ] { #2 #4 }
985   \_regex_compile_abort_tokens:x { [: \l__regex_internal_a_tl ]
986   #1 #2 #3 #4
987   }
988 }

```

(End definition for `_regex_compile_class_posix_test:w` and others.)

2.3.10 Groups and alternations

`_regex_compile_group_begin:N` The contents of a regex group are turned into compiled code in `\l__regex_internal_regex`, which ends up with items of the form `_regex_branch:n {⟨concatenation⟩}`. This construction is done using `l3tl-build` within a `TeX` group, which automatically makes sure that options (case-sensitivity and default catcode) are reset at the end of the group. The argument `#1` is `_regex_group:nnnN` or a variant thereof. A small subtlety to support `\cL(abc)` as a shorthand for `(\cLa\cLb\cLc)`: exit any pending catcode test, save the category code at the start of the group as the default catcode for that group, and make sure that the catcode is restored to the default outside the group.

```

989 \cs_new_protected:Npn \_regex_compile_group_begin:N #1
990 {
991   \_tl_build_one:n { #1 { \if_false: } \fi: }
992   \_regex_mode_quit_c:
993   \_tl_build:Nw \l__regex_internal_regex
994   \int_set_eq:NN \l__regex_default_catcodes_int \l__regex_catcodes_int
995   \int_incr:N \l__regex_group_level_int
996   \_tl_build_one:n { \_regex_branch:n { \if_false: } \fi: }
997 }
998 \cs_new_protected:Npn \_regex_compile_group_end:
999 {
1000   \if_int_compare:w \l__regex_group_level_int > \c_zero
1001     \_tl_build_one:n { \if_false: { \fi: } }
1002     \_tl_build_end:
1003     \int_set_eq:NN \l__regex_catcodes_int \l__regex_default_catcodes_int
1004     \_tl_build_one:o \l__regex_internal_regex
1005     \exp_after:wN \_regex_compile_quantifier:w
1006   \else:

```

```

1007     \_msg_kernel_warning:nn { regex } { extra-rparen }
1008     \exp_after:wN \_regex_compile_raw:N \exp_after:wN )
1009     \fi:
1010 }
(End definition for \_regex_compile_group_begin:N and \_regex_compile_group_end:.)

```

`_regex_compile_(:` In a class, parentheses are not special. Outside, check for a `?`, denoting special groups, and run the code for the corresponding special group.

```

1011 \cs_new_protected_nopar:cpn { \_regex_compile_(: }
1012 {
1013     \_regex_if_in_class:TF { \_regex_compile_raw:N ( }
1014     { \_regex_compile_lparen:w }
1015 }
1016 \cs_new_protected:Npn \_regex_compile_lparen:w #1#2#3#4
1017 {
1018     \str_if_eq:nnTF { #1 #2 } { \_regex_compile_special:N ? }
1019     {
1020         \cs_if_exist_use:cF
1021         { \_regex_compile_special_group\_token_to_str:N #4 :w }
1022         {
1023             \_msg_kernel_warning:nnx { regex } { special-group-unknown }
1024             { (? \token_to_str:N #4 }
1025             \_regex_compile_group_begin:N \_regex_group:nnnN
1026             \_regex_compile_raw:N ? #3 #4
1027         }
1028     }
1029     {
1030         \_regex_compile_group_begin:N \_regex_group:nnnN
1031         #1 #2 #3 #4
1032     }
1033 }
(End definition for \_regex_compile_(.:)

```

`_regex_compile_|:` In a class, the pipe is not special. Otherwise, end the current branch and open another one.

```

1034 \cs_new_protected_nopar:cpn { \_regex_compile_|: }
1035 {
1036     \_regex_if_in_class:TF { \_regex_compile_raw:N | }
1037     {
1038         \_tl_build_one:n
1039         { \if_false: { \fi: } \_regex_branch:n { \if_false: } \fi: }
1040     }
1041 }
(End definition for \_regex_compile_|.:)

```

`_regex_compile_):` Within a class, parentheses are not special. Outside, close a group.

```

1042 \cs_new_protected_nopar:cpn { \_regex_compile_): }
1043 {
1044     \_regex_if_in_class:TF { \_regex_compile_raw:N ) }

```

```

1045     { \_regex_compile_group_end: }
1046   }

```

(End definition for _regex_compile_.:.)

_regex_compile_special_group::w Non-capturing, and resetting groups are easy to take care of during compilation; for those
_regex_compile_special_group|:w groups, the harder parts will come when building.

```

1047 \cs_new_protected_nopar:cpn { \_regex_compile_special_group::w }
1048   { \_regex_compile_group_begin:N \_regex_group_no_capture:nnnN }
1049 \cs_new_protected_nopar:cpn { \_regex_compile_special_group|:w }
1050   { \_regex_compile_group_begin:N \_regex_group_resetting:nnnN }

```

(End definition for _regex_compile_special_group_:w.)

_regex_compile_special_group_i:w The match can be made case-insensitive by setting the option with (?i); the original
_regex_compile_special_group_:w behaviour is restored by (?-i). This is the only supported option.

```

1051 \cs_new_protected:Npn \_regex_compile_special_group_i:w #1#2
1052   {
1053     \str_if_eq:nnTF { #1 #2 } { \_regex_compile_special:N } {
1054       {
1055         \cs_set_nopar:Npn \_regex_item_equal:n { \_regex_item_caseless_equal:n }
1056         \cs_set_nopar:Npn \_regex_item_range:nn { \_regex_item_caseless_range:nn }
1057       }
1058       {
1059         \__msg_kernel_warning:nnx { regex } { unknown-option } { (?i #2 }
1060         \_regex_compile_raw:N (
1061           \_regex_compile_raw:N ?
1062           \_regex_compile_raw:N i
1063           #1 #2
1064         )
1065       }
1066 \cs_new_protected_nopar:cpn { \_regex_compile_special_group_:w } #1#2#3#4
1067   {
1068     \str_if_eq:nnTF { #1 #2 #3 #4 }
1069     { \_regex_compile_raw:N i \_regex_compile_special:N } {
1070       {
1071         \cs_set_nopar:Npn \_regex_item_equal:n { \_regex_item_caseful_equal:n }
1072         \cs_set_nopar:Npn \_regex_item_range:nn { \_regex_item_caseful_range:nn }
1073       }
1074       {
1075         \__msg_kernel_warning:nnx { regex } { unknown-option } { (?-#2#4 }
1076         \_regex_compile_raw:N (
1077           \_regex_compile_raw:N ?
1078           \_regex_compile_raw:N -
1079           #1 #2 #3 #4
1080         )
1081       }
1082     }

```

(End definition for _regex_compile_special_group_i:w and _regex_compile_special_group_:w.)

2.3.11 Catcodes and csnames

`__regex_compile_/c:`
`__regex_compile_c_test:NN`

The `\c` escape sequence can be followed by a capital letter representing a character category, by a left bracket which starts a list of categories, or by a brace group holding a regular expression for a control sequence name. Otherwise, raise an error.

```

1082 \cs_new_protected:cpn { __regex_compile_/c: }
1083   { __regex_chk_c_allowed:T { __regex_compile_c_test:NN } }
1084 \cs_new_protected:Npn __regex_compile_c_test:NN #1#2
1085   {
1086     \token_if_eq_meaning:NNTF #1 __regex_compile_raw:N
1087     {
1088       \int_if_exist:cTF { c__regex_catcode_#2_int }
1089       {
1090         \int_set_eq:Nc \l__regex_catcodes_int { c__regex_catcode_#2_int }
1091         \l__regex_mode_int
1092         = \if_case:w \l__regex_mode_int \c_two \else: \c_six \fi:
1093       }
1094     }
1095     { \cs_if_exist_use:cF { __regex_compile_c_#2:w } }
1096     {
1097       \msg_kernel_error:nxx { regex } { c-missing-category } {#2}
1098       #1 #2
1099     }
1100   }

```

(End definition for `__regex_compile_/c:` and `__regex_compile_c_test:NN`.)

`__regex_compile_c_[:w`
`__regex_compile_c_lbrack_loop:NN`
`__regex_compile_c_lbrack_add:N`
`__regex_compile_c_lbrack_end:`

When encountering `\c[`, the task is to collect uppercase letters representing character categories. First check for `^` which negates the list of category codes.

```

1101 \cs_new_protected:cpn { __regex_compile_c_[:w } #1#2
1102   {
1103     \l__regex_mode_int
1104     = \if_case:w \l__regex_mode_int \c_two \else: \c_six \fi:
1105     \int_zero:N \l__regex_catcodes_int
1106     \str_if_eq:nnTF { #1 #2 } { __regex_compile_special:N ^ }
1107     {
1108       \bool_set_false:N \l__regex_catcodes_bool
1109       __regex_compile_c_lbrack_loop:NN
1110     }
1111     {
1112       \bool_set_true:N \l__regex_catcodes_bool
1113       __regex_compile_c_lbrack_loop:NN
1114       #1 #2
1115     }
1116   }
1117 \cs_new_protected:Npn __regex_compile_c_lbrack_loop:NN #1#2
1118   {
1119     \token_if_eq_meaning:NNTF #1 __regex_compile_raw:N
1120     {
1121       \int_if_exist:cTF { c__regex_catcode_#2_int }

```

```

1122     {
1123       \exp_args:Nc \__regex_compile_c_lbrack_add:N
1124       { c__regex_catcode_#2_int }
1125       \__regex_compile_c_lbrack_loop:NN
1126     }
1127   }
1128   {
1129     \token_if_eq_charcode:NNTF #2 ]
1130     { \__regex_compile_c_lbrack_end: }
1131   }
1132   {
1133     \__msg_kernel_error:nmx { regex } { c-missing-rbrack } {#2}
1134     \__regex_compile_c_lbrack_end:
1135     #1 #2
1136   }
1137 }
1138 \cs_new_protected:Npn \__regex_compile_c_lbrack_add:N #1
1139 {
1140   \if_int_odd:w \__int_eval:w \l__regex_catcodes_int / #1 \__int_eval_end:
1141   \else:
1142     \tex_advance:D \l__regex_catcodes_int #1
1143   \fi:
1144 }
1145 \cs_new_protected_nopar:Npn \__regex_compile_c_lbrack_end:
1146 {
1147   \if_meaning:w \c_false_bool \l__regex_catcodes_bool
1148   \int_set:Nn \l__regex_catcodes_int
1149   { \c__regex_all_catcodes_int - \l__regex_catcodes_int }
1150   \fi:
1151 }

```

(End definition for __regex_compile_c[:w and others.]

__regex_compile_c_{: The case of a left brace is easy, based on what we have done so far: in a group, compile the regular expression, after changing the mode to forbid nesting `\c`. Additionally, disable submatch tracking since groups don't escape the scope of `\c{...}`.

```

1152 \cs_new_protected_nopar:cpn { __regex_compile_c_ \c_left_brace_str :w }
1153 {
1154   \__regex_compile:w
1155   \__regex_disable_submatches:
1156   \l__regex_mode_int
1157   = - \if_case:w \l__regex_mode_int \c_two \else: \c_six \fi:
1158 }

```

(End definition for __regex_compile_c_{:}. This function is documented on page ??.)

__regex_compile_} Non-escaped right braces are only special if they appear when compiling the regular expression for a csname, but not within a class: `\c{[{}]}` matches the control sequences `\}` and `\{...`. Admittedly, that would be better done as `\c{[{}]}`. So, end compiling the inner regex (this closes any dangling class or group). Then insert the corresponding test in the outer regex.


```

1159 \cs_new_protected:cpn { __regex_compile_ \c_right_brace_str : }
1160 {
1161   \__regex_if_in_cs:TF
1162   {
1163     \__regex_compile_end:
1164     \__regex_compile_one:x
1165     { \__regex_item_cs:n { \exp_not:o \l__regex_internal_regex } }
1166   }
1167   { \exp_after:wN \__regex_compile_raw:N \c_right_brace_str }
1168 }

```

(End definition for `__regex_compile_`:. This function is documented on page ??.)

2.3.12 Raw token lists with `\u`

```

\__regex_compile_/u:
\__regex_compile_u_loop:NN

```

The `\u` escape is invalid in classes and directly following a catcode test. Otherwise, it must be followed by a left brace. We then collect the characters for the argument of `\u` within an x-expanding assignment. In principle we could just wait to encounter a right brace, but this is unsafe: if the right brace is missing, then we will reach the end-markers of the regex, and continue, leading to obscure fatal errors. Instead, we only allow raw and special characters, and stop when encountering a special right brace, any escaped character, or the end-marker.

```

1169 \cs_new_protected:cpn { __regex_compile_/u: } #1#2
1170 {
1171   \__regex_if_in_class_or_catcode:TF
1172   { \__regex_compile_raw_error:N u #1 #2 }
1173   {
1174     \str_if_eq_x:nnTF {#1#2} { \__regex_compile_special:N \c_left_brace_str }
1175     {
1176       \tl_set:Nx \l__regex_internal_a_tl { \if_false: } \fi:
1177       \__regex_compile_u_loop:NN
1178     }
1179     {
1180       \__msg_kernel_error:nn { regex } { u-missing-lbrace }
1181       \__regex_compile_raw:N u #1 #2
1182     }
1183   }
1184 }
1185 \cs_new:Npn \__regex_compile_u_loop:NN #1#2
1186 {
1187   \token_if_eq_meaning:NNTF #1 \__regex_compile_raw:N
1188   { #2 \__regex_compile_u_loop:NN }
1189   {
1190     \token_if_eq_meaning:NNTF #1 \__regex_compile_special:N
1191     {
1192       \exp_after:wN \token_if_eq_charcode:NNTF \c_right_brace_str #2
1193       { \if_false: { \fi: } \__regex_compile_u_end: }
1194       { #2 \__regex_compile_u_loop:NN }
1195     }
1196     {

```

```

1197         \if_false: { \fi: }
1198         \_msg_kernel_error:nxx { regex } { u-missing-rbrace } {#2}
1199         \_regex_compile_u_end:
1200         #1 #2
1201     }
1202 }
1203 }

```

(End definition for _regex_compile_u:.)

_regex_compile_u_end: Once we have extracted the variable's name, we store the contents of that variable in \l_regex_internal_a_tl. The behaviour of \u then depends on whether we are within a \c{...} escape (in this case, the variable is turned to a string), or not.

```

1204 \cs_new_protected:Npn \_regex_compile_u_end:
1205 {
1206     \tl_set:Nv \l\_regex_internal_a_tl { \l\_regex_internal_a_tl }
1207     \if_int_compare:w \l\_regex_mode_int = \c_zero
1208         \_regex_compile_u_not_cs:
1209     \else:
1210         \_regex_compile_u_in_cs:
1211     \fi:
1212 }

```

(End definition for _regex_compile_u_end:.)

_regex_compile_u_in_cs: When \u appears within a control sequence, we convert the variable to a string with escaped spaces. Then for each character insert a class matching exactly that character, once.

```

1213 \cs_new_protected:Npn \_regex_compile_u_in_cs:
1214 {
1215     \exp_args:NNo \_str_gset_other:Nn \g\_regex_internal_tl
1216     { \l\_regex_internal_a_tl }
1217     \_tl_build_one:x
1218     {
1219         \tl_map_function:NN \g\_regex_internal_tl
1220         \_regex_compile_u_in_cs_aux:n
1221     }
1222 }
1223 \cs_new:Npn \_regex_compile_u_in_cs_aux:n #1
1224 {
1225     \_regex_class:NnnnN \c_true_bool
1226     { \_regex_item_caseful_equal:n { \_int_value:w '#1 } }
1227     { 1 } { 0 } \c_false_bool
1228 }

```

(End definition for _regex_compile_u_in_cs:.)

_regex_compile_u_not_cs: In mode 0, the \u escape adds one state to the NFA for each token in \l_regex_internal_a_tl. If a given *<token>* is a control sequence, then insert a string comparison test, otherwise, _regex_item_exact:nn which compares catcode and character code.

```

1229 \cs_new_protected:Npn \_regex_compile_u_not_cs:

```

```

1230 {
1231   \exp_args:No \__tl_analysis_map_inline:nn { \l__regex_internal_a_tl }
1232   {
1233     \__tl_build_one:n
1234     {
1235       \__regex_class:NnnnN \c_true_bool
1236       {
1237         \if_int_compare:w "##2 = \c_zero
1238         \__regex_item_exact_cs:c { \exp_after:wN \cs_to_str:N ##1 }
1239         \else:
1240         \__regex_item_exact:nn { \__int_value:w "##2 } { ##3 }
1241         \fi:
1242       }
1243       { 1 } { 0 } \c_false_bool
1244     }
1245   }
1246 }

```

(End definition for __regex_compile_u_not_cs:.)

2.3.13 Other

`__regex_compile_/K:` The `\K` control sequence is currently the only “command”, which performs some action, rather than matching something. It is allowed in the same contexts as `\b`. At the compilation stage, we leave it as a single control sequence, defined later.

```

1247 \cs_new_protected_nopar:cpn { \__regex_compile_/K: }
1248 {
1249   \int_compare:nNnTF \l__regex_mode_int = \c_zero
1250   { \__tl_build_one:n { \__regex_command_K: } }
1251   { \__regex_compile_raw_error:N K }
1252 }

```

(End definition for __regex_compile_/K:.)

2.3.14 Showing regexes

`__regex_show:Nx` Within a `__tl_build:Nw ... __tl_build_end:` group, we redefine all the function that can appear in a compiled regex, then run the regex. The result is then shown.

```

1253 \cs_new_protected:Npn \__regex_show:Nx #1#2
1254 {
1255   \__tl_build:Nw \l__regex_internal_a_tl
1256   \cs_set_protected_nopar:Npn \__regex_branch:n
1257   {
1258     \seq_pop_right:NN \l__regex_show_prefix_seq \l__regex_internal_a_tl
1259     \__regex_show_one:n { +-branch }
1260     \seq_put_right:No \l__regex_show_prefix_seq \l__regex_internal_a_tl
1261     \use:n
1262   }
1263   \cs_set_protected_nopar:Npn \__regex_group:nnnN
1264   { \__regex_show_group_aux:nnnnN { } }
1265   \cs_set_protected_nopar:Npn \__regex_group_no_capture:nnnN

```

```

1266     { \_regex_show_group_aux:nnnnN { ~(no~capture) } }
1267 \cs_set_protected_nopar:Npn \_regex_group_resetting:nnnN
1268   { \_regex_show_group_aux:nnnnN { ~(resetting) } }
1269 \cs_set_eq:NN \_regex_class:NnnnN \_regex_show_class:NnnnN
1270 \cs_set_protected_nopar:Npn \_regex_command_K:
1271   { \_regex_show_one:n { reset~match~start~(\iow_char:N\K) } }
1272 \cs_set_protected:Npn \_regex_assertion:Nn ##1##2
1273   { \_regex_show_one:n { \bool_if:NF ##1 { negative~ } assertion::~##2 } }
1274 \cs_set_nopar:Npn \_regex_b_test: { word~boundary }
1275 \cs_set_eq:NN \_regex_anchor:N \_regex_show_anchor_to_str:N
1276 \cs_set_protected:Npn \_regex_item_caseful_equal:n ##1
1277   { \_regex_show_one:n { char~code~\int_eval:n{##1} } }
1278 \cs_set_protected:Npn \_regex_item_caseful_range:nn ##1##2
1279   { \_regex_show_one:n { range~[\int_eval:n{##1}, \int_eval:n{##2}] } }
1280 \cs_set_protected:Npn \_regex_item_caseless_equal:n ##1
1281   { \_regex_show_one:n { char~code~\int_eval:n{##1}~(caseless) } }
1282 \cs_set_protected:Npn \_regex_item_caseless_range:nn ##1##2
1283   {
1284     \_regex_show_one:n
1285     { Range~[\int_eval:n{##1}, \int_eval:n{##2}]~(caseless) }
1286   }
1287 \cs_set_protected:Npn \_regex_item_catcode:nT
1288   { \_regex_show_item_catcode:NnT \c_true_bool }
1289 \cs_set_protected:Npn \_regex_item_catcode_reverse:nT
1290   { \_regex_show_item_catcode:NnT \c_false_bool }
1291 \cs_set_protected:Npn \_regex_item_reverse:n
1292   { \_regex_show_scope:nn { Reversed~match } }
1293 \cs_set_protected:Npn \_regex_item_exact:nn ##1##2
1294   { \_regex_show_one:n { char~##2,~catcode~##1 } }
1295 \cs_set_protected:Npn \_regex_item_exact_cs:c ##1
1296   { \_regex_show_one:n { control~sequence~\iow_char:N\\##1 } }
1297 \cs_set_protected:Npn \_regex_item_cs:n
1298   { \_regex_show_scope:nn { control~sequence } }
1299 \cs_set:cpn { \_regex_prop_.: } { \_regex_show_one:n { any~token } }
1300 \seq_clear:N \l_regex_show_prefix_seq
1301 \_regex_show_push:n { ~ }
1302 #1
1303 \__tl_build_end:
1304 \__msg_show_variable:n { >~Compiled~regex~#2: \l_regex_internal_a_tl }
1305 }

```

(End definition for _regex_show:Nx.)

_regex_show_one:n Every part of the final message go through this function, which adds one line to the output, with the appropriate prefix.

```

1306 \cs_new_protected:Npn \_regex_show_one:n #1
1307   {
1308     \int_incr:N \l_regex_show_lines_int
1309     \__tl_build_one:x
1310     {

```

```

1311         \exp_not:N \
1312         \seq_map_function:NN \l__regex_show_prefix_seq \use:n
1313         #1
1314     }
1315 }
(End definition for \__regex_show_one:n.)

```

__regex_show_push:n Enter and exit levels of nesting. The `scope` function prints its first argument as an “introduction”, then performs its second argument in a deeper level of nesting.

```

\__regex_show_pop:
\__regex_show_scope:nn
1316 \cs_new_protected:Npn \__regex_show_push:n #1
1317 { \seq_put_right:Nx \l__regex_show_prefix_seq { #1 ~ } }
1318 \cs_new_protected:Npn \__regex_show_pop:
1319 { \seq_pop_right:NN \l__regex_show_prefix_seq \l__regex_internal_a_tl }
1320 \cs_new_protected:Npn \__regex_show_scope:nn #1#2
1321 {
1322     \__regex_show_one:n {#1}
1323     \__regex_show_push:n { ~ }
1324     #2
1325     \__regex_show_pop:
1326 }
(End definition for \__regex_show_push:n, \__regex_show_pop:, and \__regex_show_scope:nn.)

```

_regex_show_group_aux:nnnnN We display all groups in the same way, simply adding a message, (no capture) or (resetting), to special groups. The odd `\use_ii:nn` avoids printing a spurious `+-branch` for the first branch.

```

1327 \cs_new_protected:Npn \__regex_show_group_aux:nnnnN #1#2#3#4#5
1328 {
1329     \__regex_show_one:n { , -group-begin #1 }
1330     \__regex_show_push:n { | }
1331     \use_ii:nn #2
1332     \__regex_show_pop:
1333     \__regex_show_one:n
1334     { ‘-group-end \__regex_msg_repeated:nnN {#3} {#4} #5 }
1335 }
(End definition for \__regex_show_group_aux:nnnnN.)

```

__regex_show_class:NnnnN I’m entirely unhappy about this function: I couldn’t find a way to test if a class is a single test. Instead, collect the representation of the tests in the class. If that had more than one line, write `Match` or `Don’t match` on its own line, with the repeating information if any. Then the various tests on lines of their own, and finally a line. Otherwise, we need to evaluate the representation of the tests again (since the prefix is incorrect). That’s clunky, but not too expensive, since it’s only one test.

```

1336 \cs_set:Npn \__regex_show_class:NnnnN #1#2#3#4#5
1337 {
1338     \__tl_build:Nw \l__regex_internal_a_tl
1339     \int_zero:N \l__regex_show_lines_int
1340     \__regex_show_push:n {~}
1341     #2

```

```

1342     \exp_last_unbraced:Nf
1343     \int_case:nnF { \l__regex_show_lines_int }
1344     {
1345         {0}
1346         {
1347             \__tl_build_end:
1348             \__regex_show_one:n { \bool_if:NTF #1 { Fail } { Pass } }
1349         }
1350         {1}
1351         {
1352             \__tl_build_end:
1353             \bool_if:NTF #1
1354             {
1355                 #2
1356                 \__tl_build_one:n { \__regex_msg_repeated:nnN {#3} {#4} #5 }
1357             }
1358             {
1359                 \__regex_show_one:n
1360                 { Don't~match~\__regex_msg_repeated:nnN {#3} {#4} #5 }
1361                 \__tl_build_one:o \l__regex_internal_a_tl
1362             }
1363         }
1364     }
1365     {
1366         \__tl_build_end:
1367         \__regex_show_one:n
1368         {
1369             \bool_if:NTF #1 { M } { Don't~m } atch
1370             \__regex_msg_repeated:nnN {#3} {#4} #5
1371         }
1372         \__tl_build_one:o \l__regex_internal_a_tl
1373     }
1374 }

```

(End definition for __regex_show_class:NnnnN.)

__regex_show_anchor_to_str:N

The argument is an integer telling us where the anchor is. We convert that to the relevant info.

```

1375 \cs_new:Npn \__regex_show_anchor_to_str:N #1
1376 {
1377     anchor~at~
1378     \str_case:nnF { #1 }
1379     {
1380         { \l__regex_min_pos_int } { start~(\iow_char:N\\A) }
1381         { \l__regex_start_pos_int } { start~of~match~(\iow_char:N\\G) }
1382         { \l__regex_max_pos_int } { end~(\iow_char:N\\Z) }
1383     }
1384     { <error:~'~'#1'~not~recognized> }
1385 }

```

(End definition for __regex_show_anchor_to_str:N.)

`__regex_show_item_catcode:NnT` Produce a sequence of categories which the catcode bitmap #2 contains, and show it, indenting the tests on which this catcode constraint applies.

```

1386 \cs_new_protected:Npn \__regex_show_item_catcode:NnT #1#2
1387 {
1388   \seq_set_split:Nnn \l__regex_internal_seq { } { CBEMTPUDSLOA }
1389   \seq_set_filter:NNn \l__regex_internal_seq \l__regex_internal_seq
1390     { \int_if_odd_p:n { #2 / \int_use:c { c__regex_catcode_##1_int } } }
1391   \__regex_show_scope:nn
1392     {
1393       categories~
1394       \seq_map_function:NN \l__regex_internal_seq \use:n
1395       , ~
1396       \bool_if:NF #1 { negative~ } class
1397     }
1398 }

```

(End definition for `__regex_show_item_catcode:NnT`.)

2.4 Building

2.4.1 Variables used while building

`\l__regex_min_state_int` The last state that was allocated is `\l__regex_max_state_int - 1`, so that `\l__regex_max_state_int` always points to a free state. The `min_state` variable is always 0, but is included to avoid hard-coding this value.

```

1399 \int_new:N \l__regex_min_state_int
1400 \int_new:N \l__regex_max_state_int

```

(End definition for `\l__regex_min_state_int` and `\l__regex_max_state_int`. These variables are documented on page ??.)

`\l__regex_left_state_int` Alternatives are implemented by branching from a `left` state into the various choices, then merging those into a `right` state. We store information about those states in two sequences. Those states are also used to implement group quantifiers. Most often, the `\l__regex_right_state_int` left and right pointers only differ by 1.

```

1401 \int_new:N \l__regex_left_state_int
1402 \int_new:N \l__regex_right_state_int
1403 \seq_new:N \l__regex_left_state_seq
1404 \seq_new:N \l__regex_right_state_seq

```

(End definition for `\l__regex_left_state_int` and `\l__regex_right_state_int`. These variables are documented on page ??.)

`\l__regex_capturing_group_int` `\l__regex_capturing_group_int` is the ID number that will be assigned to a capturing group if one was opened now. This starts at 0 for the group enclosing the full regular expression, and groups are counted in the order of their left parenthesis, except when encountering `resetting` groups.

```

1405 \int_new:N \l__regex_capturing_group_int

```

(End definition for `\l__regex_capturing_group_int`. This variable is documented on page ??.)

2.4.2 Framework

This phase is about going from a compiled regex to an NFA. Each state of the NFA is stored in a `\toks`. The operations which can appear in the `\toks` are

- `__regex_action_start_wildcard`: inserted at the start of the regular expression to make it unanchored.
- `__regex_action_success`: marks the exit state of the NFA.
- `__regex_action_cost:n {\langle shift \rangle}` is a transition from the current $\langle state \rangle$ to $\langle state \rangle + \langle shift \rangle$, which consumes the current character: the target state is saved and will be considered again when matching at the next position.
- `__regex_action_free:n {\langle shift \rangle}`, and `__regex_action_free_group:n {\langle shift \rangle}` are free transitions, which immediately perform the actions for the state $\langle state \rangle + \langle shift \rangle$ of the NFA. They differ in how they detect and avoid infinite loops. For now, we just need to know that the `group` variant must be used for transitions back to the start of a group.
- `__regex_action_submatch:n {\langle key \rangle}` where the $\langle key \rangle$ is a group number followed by `<` or `>` for the beginning or end of group. This causes the current position in the query to be stored as the $\langle key \rangle$ submatch boundary.

We strive to preserve the following properties while building.

- The current capturing group is `capturing_group - 1`, and if a group is opened now, it will be labelled `capturing_group`.
- The last allocated state is `max_state - 1`, so `max_state` is a free state.
- The `left_state` points to a state to the left of the current group or of the last class.
- The `right_state` points to a newly created, empty state, with some transitions leading to it.
- The `left/right` sequences hold a list of the corresponding end-points of nested groups.

`__regex_build:n` The `n`-type function first compiles its argument. Reset some variables. Allocate two states, and put a wildcard in state 0 (transitions to state 1 and 0 state). Then build the regex within a (capturing) group, which will be numbered 0 (current value of `capturing_group`). Finally, if the match reaches the last state, it is successful.

```
1406 \cs_new_protected:Npn \__regex_build:n #1
1407 {
1408   \__regex_compile:n {#1}
1409   \__regex_build:N \l__regex_internal_regex
1410 }
1411 \cs_new_protected:Npn \__regex_build:N #1
1412 {
```



```

1413 <trace> \trace_push:nnn { regex } { 1 } { __regex_build }
1414 \int_set:Nn \tex_escapechar:D { 92 }
1415 \int_zero:N \l__regex_capturing_group_int
1416 \int_set_eq:NN \l__regex_max_state_int \l__regex_min_state_int
1417 \__regex_build_new_state:
1418 \__regex_build_new_state:
1419 \__regex_toks_put_right:Nn \l__regex_left_state_int
1420 { \__regex_action_start_wildcard: }
1421 \__regex_group:nnnN {#1} { 1 } { 0 } \c_false_bool
1422 \__regex_toks_put_right:Nn \l__regex_right_state_int
1423 { \__regex_action_success: }
1424 <trace> \__regex_trace_states:n { 2 }
1425 <trace> \trace_pop:nnn { regex } { 1 } { __regex_build }
1426 }

```

(End definition for `__regex_build:n` and `__regex_build:N`.)

`__regex_build_for_cs:n` When using a regex to match a cs, we don't insert a wildcard, we anchor at the end, and since we ignore submatches, there is no need to surround the expression with a group. However, for branches to work properly at the outer level, we need to put the appropriate `left` and `right` states in their sequence.

```

1427 \cs_new_protected:Npn \__regex_build_for_cs:n #1
1428 {
1429 <trace> \trace_push:nnn { regex } { 1 } { __regex_build_for_cs }
1430 \int_set_eq:NN \l__regex_max_state_int \l__regex_min_state_int
1431 \__regex_build_new_state:
1432 \__regex_build_new_state:
1433 \__regex_push_lr_states:
1434 #1
1435 \__regex_pop_lr_states:
1436 \__regex_toks_put_right:Nn \l__regex_right_state_int
1437 {
1438 \if_int_compare:w \l__regex_current_pos_int = \l__regex_max_pos_int
1439 \exp_after:wN \__regex_action_success:
1440 \fi:
1441 }
1442 <trace> \__regex_trace_states:n { 2 }
1443 <trace> \trace_pop:nnn { regex } { 1 } { __regex_build_for_cs }
1444 }

```

(End definition for `__regex_build_for_cs:n`.)

2.4.3 Helpers for building an nfa

`__regex_push_lr_states:` When building the regular expression, we keep track of pointers to the left-end and right-end of each group without help from TeX's grouping.

`__regex_pop_lr_states:`

```

1445 \cs_new_protected_nopar:Npn \__regex_push_lr_states:
1446 {
1447 \seq_push:No \l__regex_left_state_seq
1448 { \int_use:N \l__regex_left_state_int }
1449 \seq_push:No \l__regex_right_state_seq

```

```

1450     { \int_use:N \l__regex_right_state_int }
1451   }
1452 \cs_new_protected_nopar:Npn \__regex_pop_lr_states:
1453 {
1454   \seq_pop:Nn \l__regex_left_state_seq \l__regex_internal_a_tl
1455   \int_set:Nn \l__regex_left_state_int \l__regex_internal_a_tl
1456   \seq_pop:Nn \l__regex_right_state_seq \l__regex_internal_a_tl
1457   \int_set:Nn \l__regex_right_state_int \l__regex_internal_a_tl
1458 }

```

(End definition for __regex_push_lr_states: and __regex_pop_lr_states:.)

__regex_toks_put_left:Nx During the building phase we wish to add x-expanded material to \toks, either to the left or to the right. The expansion is done “by hand” for optimization (these operations are used quite a lot). The Nn version of __regex_toks_put_right:Nx is provided because it is more efficient than x-expanding with \exp_not:n.

```

1459 \cs_new_protected:Npn \__regex_toks_put_left:Nx #1#2
1460 {
1461   \cs_set_nopar:Npx \__regex_tmp:w { #2 }
1462   \tex_toks:D #1 \exp_after:wN \exp_after:wN \exp_after:wN
1463   { \exp_after:wN \__regex_tmp:w \tex_the:D \tex_toks:D #1 }
1464 }
1465 \cs_new_protected:Npn \__regex_toks_put_right:Nx #1#2
1466 {
1467   \cs_set_nopar:Npx \__regex_tmp:w {#2}
1468   \tex_toks:D #1 \exp_after:wN
1469   { \tex_the:D \tex_toks:D \exp_after:wN #1 \__regex_tmp:w }
1470 }
1471 \cs_new_protected:Npn \__regex_toks_put_right:Nn #1#2
1472 { \tex_toks:D #1 \exp_after:wN { \tex_the:D \tex_toks:D #1 #2 } }

```

(End definition for __regex_toks_put_left:Nx.)

__regex_build_transition_left:NNN Add a transition from #2 to #3 using the function #1. The left function is used for higher priority transitions, and the right function for lower priority transitions (which should be performed later). The signatures differ to reflect the differing usage later on. Both functions could be optimized.

```

1473 \cs_new_protected:Npn \__regex_build_transition_left:NNN #1#2#3
1474 { \__regex_toks_put_left:Nx #2 { #1 { \int_eval:n { #3 - #2 } } } }
1475 \cs_new_protected:Npn \__regex_build_transition_right:nNn #1#2#3
1476 { \__regex_toks_put_right:Nx #2 { #1 { \int_eval:n { #3 - #2 } } } }

```

(End definition for __regex_build_transition_left:NNN and __regex_build_transition_right:nNn.)

__regex_build_new_state: Add a new empty state to the NFA. Then update the left, right, and max states, so that the right state is the new empty state, and the left state points to the previously “current” state.

```

1477 \cs_new_protected_nopar:Npn \__regex_build_new_state:
1478 {
1479   <*trace>
1480   \trace:nxx { regex } { 2 }

```

```

1481     {
1482       regex~new~state~
1483       L=\int_use:N \l__regex_left_state_int ~ -> ~
1484       R=\int_use:N \l__regex_right_state_int ~ -> ~
1485       M=\int_use:N \l__regex_max_state_int ~ -> ~
1486       \int_eval:n { \l__regex_max_state_int + \c_one }
1487     }
1488   </trace>
1489   \tex_toks:D \l__regex_max_state_int { }
1490   \int_set_eq:NN \l__regex_left_state_int \l__regex_right_state_int
1491   \int_set_eq:NN \l__regex_right_state_int \l__regex_max_state_int
1492   \int_incr:N \l__regex_max_state_int
1493 }
(End definition for \__regex_build_new_state:.)

```

`__regex_build_transitions_lazyness:NNNNN` This function creates a new state, and puts two transitions starting from the old current state. The order of the transitions is controlled by #1, true for lazy quantifiers, and false for greedy quantifiers.

```

1494 \cs_new_protected:Npn \__regex_build_transitions_lazyness:NNNNN #1#2#3#4#5
1495 {
1496   \__regex_build_new_state:
1497   \__regex_toks_put_right:Nx \l__regex_left_state_int
1498   {
1499     \if_meaning:w \c_true_bool #1
1500       #2 { \int_eval:n { #3 - \l__regex_left_state_int } }
1501       #4 { \int_eval:n { #5 - \l__regex_left_state_int } }
1502     \else:
1503       #4 { \int_eval:n { #5 - \l__regex_left_state_int } }
1504       #2 { \int_eval:n { #3 - \l__regex_left_state_int } }
1505     \fi:
1506   }
1507 }
(End definition for \__regex_build_transitions_lazyness:NNNNN.)

```

2.4.4 Building classes

`__regex_class:NnnnN` The arguments are: $\langle\text{boolean}\rangle$ $\{\langle\text{tests}\rangle\}$ $\{\langle\text{min}\rangle\}$ $\{\langle\text{more}\rangle\}$ $\langle\text{lazyness}\rangle$. First store the tests with a trailing `__regex_action_cost:n`, in the true branch of `__regex_break_point:TF` for positive classes, or the false branch for negative classes. The integer $\langle\text{more}\rangle$ is 0 for fixed repetitions, -1 for unbounded repetitions, and $\langle\text{max}\rangle - \langle\text{min}\rangle$ for a range of repetitions.

```

1508 \cs_new_protected:Npn \__regex_class:NnnnN #1#2#3#4#5
1509 {
1510   \cs_set_nopar:Npx \__regex_tests_action_cost:n ##1
1511   {
1512     \exp_not:n { \exp_not:n {#2} }
1513     \bool_if:NTF #1
1514       { \__regex_break_point:TF { \__regex_action_cost:n {##1} } { } }

```

```

1515         { \_regex_break_point:TF { } { \_regex_action_cost:n {##1} } }
1516     }
1517     \if_case:w - #4 \exp_stop_f:
1518         \_regex_class_repeat:n {#3}
1519     \or: \_regex_class_repeat:nN {#3} #5
1520     \else: \_regex_class_repeat:nnN {#3} {#4} #5
1521     \fi:
1522 }
1523 \cs_new:Npn \_regex_tests_action_cost:n { \_regex_action_cost:n }
(End definition for \_regex_class:NnnN.)

```

`_regex_class_repeat:n` This is used for a fixed number of repetitions. Build one state for each repetition, with a transition controlled by the tests that we have collected. That works just fine for `#1 = 0` repetitions: nothing is built.

```

1524 \cs_new_protected:Npn \_regex_class_repeat:n #1
1525 {
1526     \prg_replicate:nn {#1}
1527     {
1528         \_regex_build_new_state:
1529         \_regex_build_transition_right:nNn \_regex_tests_action_cost:n
1530         \l__regex_left_state_int \l__regex_right_state_int
1531     }
1532 }
(End definition for \_regex_class_repeat:n.)

```

`_regex_class_repeat:nN` This implements unbounded repetitions of a single class (*e.g.* the `*` and `+` quantifiers). If the minimum number `#1` of repetitions is 0, then build a transition from the current state to itself governed by the tests, and a free transition to a new state (hence skipping the tests). Otherwise, call `_regex_class_repeat:n` for the code to match `#1` repetitions, and add free transitions from the last state to the previous one, and to a new one. In both cases, the order of transitions is controlled by the laziness boolean `#2`.

```

1533 \cs_new_protected:Npn \_regex_class_repeat:nN #1#2
1534 {
1535     \if_int_compare:w #1 = \c_zero
1536         \_regex_build_transitions_lazyness:NNNN #2
1537         \_regex_action_free:n \l__regex_right_state_int
1538         \_regex_tests_action_cost:n \l__regex_left_state_int
1539     \else:
1540         \_regex_class_repeat:n {#1}
1541         \int_set_eq:NN \l__regex_internal_a_int \l__regex_left_state_int
1542         \_regex_build_transitions_lazyness:NNNN #2
1543         \_regex_action_free:n \l__regex_right_state_int
1544         \_regex_action_free:n \l__regex_internal_a_int
1545     \fi:
1546 }
(End definition for \_regex_class_repeat:nN.)

```

`_regex_class_repeat:nnN` We want to build the code to match from `#1` to `#1+#2` repetitions. Match `#1` repetitions (can be 0). Compute the final state of the next construction as `a`. Build `#2 > 0` states, each with a transition to the next state governed by the tests, and a transition to the final state `a`. The computation of `a` is safe because states are allocated in order, starting from `max_state`.

```

1547 \\cs_new_protected:Npn \\_regex_class_repeat:nnN #1#2#3
1548 {
1549   \\_regex_class_repeat:n {#1}
1550   \\int_set:Nn \\l__regex_internal_a_int
1551     { \\l__regex_max_state_int + #2 - \\c_one }
1552   \\prg_replicate:nn { #2 }
1553   {
1554     \\_regex_build_transitions_lazyness:NNNN #3
1555     \\_regex_action_free:n      \\l__regex_internal_a_int
1556     \\_regex_tests_action_cost:n \\l__regex_right_state_int
1557   }
1558 }

```

(End definition for `_regex_class_repeat:nnN`.)

2.4.5 Building groups

`_regex_group_aux:nnnnN` Arguments: $\{\langle label \rangle\} \{\langle contents \rangle\} \{\langle min \rangle\} \{\langle more \rangle\} \langle lazyness \rangle$. If $\langle min \rangle$ is 0, we need to add a state before building the group, so that the thread which skips the group does not also set the start-point of the submatch. After adding one more state, the `left_state` is the left end of the group, from which all branches will stem, and the `right_state` is the right end of the group, and all branches end their course in that state. We store those two integers to be queried for each branch, we build the NFA states for the contents `#2` of the group, and we forget about the two integers. Once this is done, perform the repetition: either exactly `#3` times, or `#3` or more times, or between `#3` and `#3 + #4` times, with lazyness `#5`. The $\langle label \rangle$ `#1` is used for submatch tracking. Each of the three auxiliaries expects `left_state` and `right_state` to be set properly.

```

1559 \\cs_new_protected:Npn \\_regex_group_aux:nnnnN #1#2#3#4#5
1560 {
1561   <trace>      \\trace_push:nnn { regex } { 1 } { \\_regex_group }
1562               \\if_int_compare:w #3 = \\c_zero
1563               \\_regex_build_new_state:
1564   <assert> \\assert_int:n { \\l__regex_max_state_int = \\l__regex_right_state_int + 1 }
1565           \\_regex_build_transition_right:nNn \\_regex_action_free_group:n
1566           \\l__regex_left_state_int \\l__regex_right_state_int
1567   \\fi:
1568   \\_regex_build_new_state:
1569   \\_regex_push_lr_states:
1570   #2
1571   \\_regex_pop_lr_states:
1572   \\if_case:w - #4 \\exp_stop_f:
1573     \\_regex_group_repeat:nn {#1} {#3}
1574   \\or:  \\_regex_group_repeat:nnN {#1} {#3}      #5
1575   \\else: \\_regex_group_repeat:nnnN {#1} {#3} {#4} #5

```

```

1576      \fi:
1577      <trace>      \trace_pop:nnn { regex } { 1 } { __regex_group }
1578      }
(End definition for \__regex_group_aux:nnnnN.)

```

`__regex_group:nnnN` Hand to `__regex_group_aux:nnnnN` the label of that group (expanded), and the group itself, with some extra commands to perform.

```

\__regex_group_no_capture:nnnN
1579 \cs_new_protected:Npn \__regex_group:nnnN #1
1580 {
1581     \exp_args:No \__regex_group_aux:nnnnN
1582     { \int_use:N \l__regex_capturing_group_int }
1583     {
1584         \int_incr:N \l__regex_capturing_group_int
1585         #1
1586     }
1587 }
1588 \cs_new_protected_nopar:Npn \__regex_group_no_capture:nnnN
1589 { \__regex_group_aux:nnnnN { -1 } }
(End definition for \__regex_group:nnnN and \__regex_group_no_capture:nnnN.)

```

`__regex_group_resetting:nnnN` Again, hand the label `-1` to `__regex_group_aux:nnnnN`, but this time we work a little bit harder to keep track of the maximum group label at the end of any branch, and to reset the group number at each branch. This relies on the fact that a compiled regex always is a sequence of items of the form `__regex_branch:n {<branch>}`.

```

\__regex_group_resetting_loop:nnNn
1590 \cs_new_protected:Npn \__regex_group_resetting:nnnN #1
1591 {
1592     \__regex_group_aux:nnnnN { -1 }
1593     {
1594         \exp_args:Noo \__regex_group_resetting_loop:nnNn
1595         { \int_use:N \l__regex_capturing_group_int }
1596         { \int_use:N \l__regex_capturing_group_int }
1597         #1
1598         { ?? \__prg_break:n } { }
1599         \__prg_break_point:
1600     }
1601 }
1602 \cs_new_protected:Npn \__regex_group_resetting_loop:nnNn #1#2#3#4
1603 {
1604     \use_none:nn #3 { \int_set:Nn \l__regex_capturing_group_int {#1} }
1605     \int_set:Nn \l__regex_capturing_group_int {#2}
1606     #3 {#4}
1607     \exp_args:Nf \__regex_group_resetting_loop:nnNn
1608     { \int_max:nn {#1} { \l__regex_capturing_group_int } }
1609     {#2}
1610 }
(End definition for \__regex_group_resetting:nnnN.)

```

`__regex_branch:n` Add a free transition from the left state of the current group to a brand new state, starting point of this branch. Once the branch is built, add a transition from its last

state to the right state of the group. The left and right states of the group are extracted from the relevant sequences.

```

1611 \cs_new_protected:Npn \__regex_branch:n #1
1612 {
1613   \trace \trace_push:nnn { regex } { 1 } { __regex_branch }
1614   \__regex_build_new_state:
1615   \seq_get:NN \l__regex_left_state_seq \l__regex_internal_a_tl
1616   \int_set:Nn \l__regex_left_state_int \l__regex_internal_a_tl
1617   \__regex_build_transition_right:nNn \__regex_action_free:n
1618   \l__regex_left_state_int \l__regex_right_state_int
1619   #1
1620   \seq_get:NN \l__regex_right_state_seq \l__regex_internal_a_tl
1621   \__regex_build_transition_right:nNn \__regex_action_free:n
1622   \l__regex_right_state_int \l__regex_internal_a_tl
1623   \trace \trace_pop:nnn { regex } { 1 } { __regex_branch }
1624 }

```

(End definition for __regex_branch:n.)

`__regex_group_repeat:nn` This function is called to repeat a group a fixed number of times #2; if this is 0 we remove the group altogether (but don't reset the `capturing_group` label). Otherwise, the auxiliary `__regex_group_repeat_aux:n` copies #2 times the `\toks` for the group, and leaves `internal_a` pointing to the left end of the last repetition. We only record the submatch information at the last repetition. Finally, add a state at the end (the transition to it has been taken care of by the replicating auxiliary).

```

1625 \cs_new_protected:Npn \__regex_group_repeat:nn #1#2
1626 {
1627   \if_int_compare:w #2 = \c_zero
1628     \int_set:Nn \l__regex_max_state_int
1629     { \l__regex_left_state_int - \c_one }
1630     \__regex_build_new_state:
1631   \else:
1632     \__regex_group_repeat_aux:n {#2}
1633     \__regex_group_submatches:nNN {#1}
1634     \l__regex_internal_a_int \l__regex_right_state_int
1635     \__regex_build_new_state:
1636   \fi:
1637 }

```

(End definition for __regex_group_repeat:nn.)

`__regex_group_submatches:nNN` This inserts in states #2 and #3 the code for tracking submatches of the group #1, unless inhibited by a label of -1.

```

1638 \cs_new_protected:Npn \__regex_group_submatches:nNN #1#2#3
1639 {
1640   \if_int_compare:w #1 > \c_minus_one
1641     \__regex_toks_put_left:Nx #2 { \__regex_action_submatch:n { #1 < } }
1642     \__regex_toks_put_left:Nx #3 { \__regex_action_submatch:n { #1 > } }
1643   \fi:
1644 }

```

(End definition for `_regex_group_submatches:nNN`.)

`_regex_group_repeat_aux:n` Here we repeat `\toks` ranging from `left_state` to `max_state`, `#1 > 0` times. First add a transition so that the copies will “chain” properly. Compute the shift `c` between the original copy and the last copy we want. Shift the `right_state` and `max_state` to their final values. We then want to perform `c` copy operations. At the end, `b` is equal to the `max_state`, and `a` points to the left of the last copy of the group.

```

1645 \cs_new_protected:Npn \_regex_group_repeat_aux:n #1
1646 {
1647   \_regex_build_transition_right:nNn \_regex_action_free:n
1648   \l__regex_right_state_int \l__regex_max_state_int
1649   \int_set_eq:NN \l__regex_internal_a_int \l__regex_left_state_int
1650   \int_set_eq:NN \l__regex_internal_b_int \l__regex_max_state_int
1651   \if_int_compare:w \__int_eval:w #1 > \c_one
1652     \int_set:Nn \l__regex_internal_c_int
1653     {
1654       ( #1 - \c_one )
1655       * ( \l__regex_internal_b_int - \l__regex_internal_a_int )
1656     }
1657   \tex_advance:D \l__regex_right_state_int \l__regex_internal_c_int
1658   \tex_advance:D \l__regex_max_state_int \l__regex_internal_c_int
1659   \prg_replicate:nn \l__regex_internal_c_int
1660   {
1661     \tex_toks:D \l__regex_internal_b_int
1662     = \tex_toks:D \l__regex_internal_a_int
1663     \tex_advance:D \l__regex_internal_a_int \c_one
1664     \tex_advance:D \l__regex_internal_b_int \c_one
1665   }
1666   \fi:
1667 }

```

(End definition for `_regex_group_repeat_aux:n`.)

`_regex_group_repeat:nnN` This function is called to repeat a group at least `n` times; the case `n = 0` is very different from `n > 0`. Assume first that `n = 0`. Insert submatch tracking information at the start and end of the group, add a free transition from the right end to the “true” left state `a` (remember: in this case we had added an extra state before the left state). This forms the loop, which we break away from by adding a free transition from `a` to a new state.

Now consider the case `n > 0`. Repeat the group `n` times, chaining various copies with a free transition. Add submatch tracking only to the last copy, then add a free transition from the right end back to the left end of the last copy, either before or after the transition to move on towards the rest of the NFA. This transition can end up before submatch tracking, but that is irrelevant since it only does so when going again through the group, recording new matches. Finally, add a state; we already have a transition pointing to it from `_regex_group_repeat_aux:n`.

```

1668 \cs_new_protected:Npn \_regex_group_repeat:nnN #1#2#3
1669 {
1670   \if_int_compare:w #2 = \c_zero
1671     \_regex_group_submatches:nNN {#1}

```



```

1672     \l__regex_left_state_int \l__regex_right_state_int
1673     \int_set:Nn \l__regex_internal_a_int
1674     { \l__regex_left_state_int - \c_one }
1675     \__regex_build_transition_right:nNn \__regex_action_free:n
1676     \l__regex_right_state_int \l__regex_internal_a_int
1677     \__regex_build_new_state:
1678     \if_meaning:w \c_true_bool #3
1679     \__regex_build_transition_left:NNN \__regex_action_free:n
1680     \l__regex_internal_a_int \l__regex_right_state_int
1681     \else:
1682     \__regex_build_transition_right:nNn \__regex_action_free:n
1683     \l__regex_internal_a_int \l__regex_right_state_int
1684     \fi:
1685     \else:
1686     \__regex_group_repeat_aux:n {#2}
1687     \__regex_group_submatches:nNN {#1}
1688     \l__regex_internal_a_int \l__regex_right_state_int
1689     \if_meaning:w \c_true_bool #3
1690     \__regex_build_transition_right:nNn \__regex_action_free_group:n
1691     \l__regex_right_state_int \l__regex_internal_a_int
1692     \else:
1693     \__regex_build_transition_left:NNN \__regex_action_free_group:n
1694     \l__regex_right_state_int \l__regex_internal_a_int
1695     \fi:
1696     \__regex_build_new_state:
1697     \fi:
1698 }

```

(End definition for __regex_group_repeat:nnN.)

__regex_group_repeat:nnnN

We wish to repeat the group between #2 and #2 + #3 times, with a laziness controlled by #4. We insert submatch tracking up front: in principle, we could avoid recording submatches for the first #2 copies of the group, but that forces us to treat specially the case #2 = 0. Repeat that group with submatch tracking #2 + #3 times (the maximum number of repetitions). Then our goal is to add #3 transitions from the end of the #2-th group, and each subsequent groups, to the end. For a lazy quantifier, we add those transitions to the left states, before submatch tracking. For the greedy case, we add the transitions to the right states, after submatch tracking and the transitions which go on with more repetitions. In the greedy case with #2 = 0, the transition which skips over all copies of the group must be added separately, because its starting state does not follow the normal pattern: we had to add it “by hand” earlier.

```

1699 \cs_new_protected:Npn \__regex_group_repeat:nnnN #1#2#3#4
1700 {
1701     \__regex_group_submatches:nNN {#1}
1702     \l__regex_left_state_int \l__regex_right_state_int
1703     \__regex_group_repeat_aux:n { #2 + #3 }
1704     \if_meaning:w \c_true_bool #4
1705     \int_set_eq:NN \l__regex_left_state_int \l__regex_max_state_int
1706     \prg_replicate:nn { #3 }
1707     {

```

```

1708         \int_sub:Nn \l__regex_left_state_int
1709         { \l__regex_internal_b_int - \l__regex_internal_a_int }
1710         \__regex_build_transition_left:NNN \__regex_action_free:n
1711         \l__regex_left_state_int \l__regex_max_state_int
1712     }
1713 \else:
1714     \prg_replicate:nn { #3 - \c_one }
1715     {
1716         \int_sub:Nn \l__regex_right_state_int
1717         { \l__regex_internal_b_int - \l__regex_internal_a_int }
1718         \__regex_build_transition_right:nNn \__regex_action_free:n
1719         \l__regex_right_state_int \l__regex_max_state_int
1720     }
1721     \if_int_compare:w #2 = \c_zero
1722     \int_set:Nn \l__regex_right_state_int
1723     { \l__regex_left_state_int - \c_one }
1724 \else:
1725     \int_sub:Nn \l__regex_right_state_int
1726     { \l__regex_internal_b_int - \l__regex_internal_a_int }
1727 \fi:
1728     \__regex_build_transition_right:nNn \__regex_action_free:n
1729     \l__regex_right_state_int \l__regex_max_state_int
1730 \fi:
1731 \__regex_build_new_state:
1732 }

```

(End definition for __regex_group_repeat:nnnN.)

2.4.6 Others

__regex_assertion:Nn Usage: __regex_assertion:Nn *<boolean>* {*<test>*}, where the *<test>* is either of the two other functions. Add a free transition to a new state, conditionally to the assertion test.

__regex_b_test: The __regex_b_test: test is used by the \b and \B escape: check if the last character was a word character or not, and do the same to the current character. The boundary-markers of the string are non-word characters for this purpose. Anchors at the start or end of match use __regex_anchor:N, with a position controlled by the integer #1.

```

1733 \cs_new_protected:Npn \__regex_assertion:Nn #1#2
1734 {
1735     \__regex_build_new_state:
1736     \__regex_toks_put_right:Nx \l__regex_left_state_int
1737     {
1738         \exp_not:n {#2}
1739         \__regex_break_point:TF
1740         \bool_if:NF #1 { { } }
1741         {
1742             \__regex_action_free:n
1743             {
1744                 \int_eval:n
1745                 { \l__regex_right_state_int - \l__regex_left_state_int }
1746             }
1747         }
1748     }
1749 }

```

```

1747     }
1748     \bool_if:NT #1 { { } }
1749   }
1750 }
1751 \cs_new_protected:Npn \__regex_anchor:N #1
1752 {
1753   \if_int_compare:w #1 = \l__regex_current_pos_int
1754     \exp_after:wN \__regex_break_true:w
1755   \fi:
1756 }
1757 \cs_new_protected_nopar:Npn \__regex_b_test:
1758 {
1759   \group_begin:
1760   \int_set_eq:NN \l__regex_current_char_int \l__regex_last_char_int
1761   \__regex_prop_w:
1762   \__regex_break_point:TF
1763     { \group_end: \__regex_item_reverse:n \__regex_prop_w: }
1764     { \group_end: \__regex_prop_w: }
1765 }

```

(End definition for __regex_assertion:Nn, __regex_b_test:, and __regex_anchor:N.)

__regex_command_K: Change the starting point of the 0-th submatch (full match), and transition to a new state, pretending that this is a fresh thread.

```

1766 \cs_new_protected_nopar:Npn \__regex_command_K:
1767 {
1768   \__regex_build_new_state:
1769   \__regex_toks_put_right:Nx \l__regex_left_state_int
1770   {
1771     \__regex_action_submatch:n { 0< }
1772     \bool_set_true:N \l__regex_fresh_thread_bool
1773     \__regex_action_free:n
1774       { \int_eval:n { \l__regex_right_state_int - \l__regex_left_state_int } }
1775     \bool_set_false:N \l__regex_fresh_thread_bool
1776   }
1777 }

```

(End definition for __regex_command_K:.)

2.5 Matching

We search for matches by running all the execution threads through the NFA in parallel, reading one token of the query at each step. The NFA contains “free” transitions to other states, and transitions which “consume” the current token. For free transitions, the instruction at the new state of the NFA is performed immediately. When a transition consumes a character, the new state is appended to a list of “active states”, stored in `\skip` registers: this thread will be active again when the next token is read from the query. At every step (for each token in the query), we unpack that list of active states and the corresponding submatch props, and empty those.

If two paths through the NFA “collide” in the sense that they reach the same state after reading a given token, then they only differ in how they previously matched, and the future execution will be identical for both. (Note that this would be wrong in the presence of back-references.) Hence, we only need to keep one of the two threads: the thread with the highest priority. Our NFA is built in such a way that higher priority actions always come before lower priority actions, which makes things work.

The explanation in the previous paragraph may make us think that we simply need to keep track of which states were visited at a given step: after all, the loop generated when matching `(a?)*` against `a` is broken, isn’t it? No. The group first matches `a`, as it should, then repeats; it attempts to match `a` again but fails; it skips `a`, and finds out that this state has already been seen at this position in the query: the match stops. The capturing group is (wrongly) `a`. What went wrong is that a thread collided with itself, and the later version, which has gone through the group one more times with an empty match, should have a higher priority than not going through the group.

We solve this by distinguishing “normal” free transitions `_regex_action_free:n` from transitions `_regex_action_free_group:n` which go back to the start of the group. The former will keep threads unless they have been visited by a “completed” thread, while the latter kind of transition also prevents going back to a state visited by the current thread.

2.5.1 Variables used when matching

<code>\l__regex_min_pos_int</code> <code>\l__regex_max_pos_int</code> <code>\l__regex_current_pos_int</code> <code>\l__regex_start_pos_int</code> <code>\l__regex_success_pos_int</code>	<p>The tokens in the query are indexed from <code>min_pos</code> for the first to <code>max_pos - 1</code> for the last, and their information is stored in <code>\muskip</code> and <code>\toks</code> registers with those numbers. We don’t start from 0 because the <code>\toks</code> registers with low numbers are used to hold the states of the NFA. We match without backtracking, keeping all threads in lockstep at the <code>current_pos</code> in the query. The starting point of the current match attempt is <code>start_pos</code>, and <code>success_pos</code>, updated whenever a thread succeeds, is used as the next starting position.</p>
--	--

```

1778 \int_new:N \l__regex_min_pos_int
1779 \int_new:N \l__regex_max_pos_int
1780 \int_new:N \l__regex_current_pos_int
1781 \int_new:N \l__regex_start_pos_int
1782 \int_new:N \l__regex_success_pos_int

```

(End definition for \l__regex_min_pos_int and others. These variables are documented on page ??.)

<code>\l__regex_current_char_int</code> <code>\l__regex_current_catcode_int</code> <code>\l__regex_last_char_int</code> <code>\l__regex_case_changed_char_int</code>	<p>The character and category codes of the token at the current position; the character code of the token at the previous position; and the character code of the result of changing the case of the current token (<code>A-Z↔a-z</code>). This last integer is only computed when necessary, and is otherwise <code>\c_max_int</code>. The <code>current_char</code> variable is also used in various other phases to hold a character code.</p>
---	---

```

1783 \int_new:N \l__regex_current_char_int
1784 \int_new:N \l__regex_current_catcode_int
1785 \int_new:N \l__regex_last_char_int
1786 \int_new:N \l__regex_case_changed_char_int

```

(End definition for \l__regex_current_char_int and others. These variables are documented on page ??.)

`\l__regex_current_state_int` For every character in the token list, each of the active states is considered in turn. The variable `\l__regex_current_state_int` holds the state of the NFA which is currently considered: transitions are then given as shifts relative to the current state.

```
1787 \int_new:N \l__regex_current_state_int
```

(End definition for `\l__regex_current_state_int`. This variable is documented on page ??.)

`\l__regex_current_submatches_prop`
`\l__regex_success_submatches_prop` The submatches for the thread which is currently active are stored in the `current_submatches` property list variable. This property list is stored by `__regex_action_cost:n` into the `\toks` register for the target state of the transition, to be retrieved when matching at the next position. When a thread succeeds, this property list is copied to `\l__regex_success_submatches_prop`: only the last successful thread will remain there.

```
1788 \prop_new:N \l__regex_current_submatches_prop
1789 \prop_new:N \l__regex_success_submatches_prop
```

(End definition for `\l__regex_current_submatches_prop` and `\l__regex_success_submatches_prop`. These variables are documented on page ??.)

`\l__regex_step_int` This integer, always even, is increased every time a character in the query is read, and not reset when doing multiple matches. For each `\dimen<state>` in the NFA we store in `\dimen<state>` the last step in which this state was encountered. This lets us break infinite loops by not visiting the same state twice in the same step. In fact, `\dimen<state>` is equal `step` when we have started performing the operations of `\toks<state>`, but not finished yet. However, once we finish, we set `\dimen<state>` to `step + 1`. This is needed to track submatches properly (see building phase). The `step` is also used to attach each set of submatch information to a given iteration (and automatically discard it when it corresponds to a past step).

```
1790 \int_new:N \l__regex_step_int
```

(End definition for `\l__regex_step_int`. This variable is documented on page ??.)

`\l__regex_min_active_int`
`\l__regex_max_active_int` All the currently active states are kept in order of precedence in the `\skip` registers, and the corresponding submatches in the `\toks`. For our purposes, those serve as an array, indexed from `min_active` (inclusive) to `max_active` (excluded). At the start of every step, the whole array is unpacked, so that the space can immediately be reused, and `max_active` is reset to `min_active`, effectively clearing the array.

```
1791 \int_new:N \l__regex_min_active_int
1792 \int_new:N \l__regex_max_active_int
```

(End definition for `\l__regex_min_active_int` and `\l__regex_max_active_int`. These variables are documented on page ??.)

`\l__regex_every_match_tl` Every time a match is found, this token list is used. For single matching, the token list is empty. For multiple matching, the token list is set to repeat the matching, after performing some operation which depends on the user function. See `__regex_single_match:` and `__regex_multi_match:n`.

```
1793 \tl_new:N \l__regex_every_match_tl
```

(End definition for `\l__regex_every_match_tl`. This variable is documented on page ??.)

`\l__regex_fresh_thread_bool` When doing multiple matches, we need to avoid infinite loops where each iteration matches the same empty token list. When an empty token list is matched, the next successful match of the same empty token list is suppressed. We detect empty matches by setting `\l__regex_fresh_thread_bool` to `true` for threads which directly come from the start of the regex or from the `\K` command, and testing that boolean whenever a thread succeeds. The function `__regex_if_two_empty_matches:F` is redefined at every match attempt, depending on whether the previous match was empty or not: if it was, then the function must cancel a purported success if it is empty and at the same spot as the previous match; otherwise, we definitely don't have two identical empty matches, so the function is `\use:n`.

```

1794 \bool_new:N \l__regex_fresh_thread_bool
1795 \bool_new:N \l__regex_empty_success_bool
1796 \cs_new_eq:NN \__regex_if_two_empty_matches:F \use:n

```

(End definition for `\l__regex_fresh_thread_bool` and `\l__regex_empty_success_bool`. These variables are documented on page ??.)

`\g__regex_success_bool` The boolean `\l__regex_match_success_bool` is true if the current match attempt was successful, and `\g__regex_success_bool` is true if there was at least one successful match. This is the only global variable in this whole module, but we would need it to be local when matching a control sequence with `\c{...}`. This is done by saving the global variable into `\l__regex_saved_success_bool`, which is local, hence not affected by the changes due to inner regex functions.

```

1797 \bool_new:N \g__regex_success_bool
1798 \bool_new:N \l__regex_saved_success_bool
1799 \bool_new:N \l__regex_match_success_bool

```

(End definition for `\g__regex_success_bool`, `\l__regex_saved_success_bool`, and `\l__regex_match_success_bool`. These variables are documented on page ??.)

2.5.2 Matching: framework

`__regex_match:n` First store the query into `\toks` and `\muskip` registers (see `__regex_query_set:nnn`). Then initialize the variables that should be set once for each user function (even for multiple matches). Namely, the overall matching is not yet successful; none of the states should be marked as visited (`\dimen` registers), and we start at step 0; we pretend that there was a previous match ending at the start of the query, which was not empty (to avoid smothering an empty match at the start). Once all this is set up, we are ready for the ride. Find the first match.

```

1800 \cs_new_protected:Npn \__regex_match:n #1
1801 {
1802   \trace \trace_push:nnx { regex } { 1 } { __regex_match }
1803   \trace \trace:nnx { regex } { 1 } { analyzing~query~token~list }
1804   \int_zero:N \l__regex_balance_int
1805   \int_set:Nn \l__regex_current_pos_int { \c_two * \l__regex_max_state_int }
1806   \__regex_query_set:nnn { } { -1 } { -2 }
1807   \int_set_eq:NN \l__regex_min_pos_int \l__regex_current_pos_int
1808   \__tl_analysis_map_inline:nn {#1}
1809   { \__regex_query_set:nnn {##1} {"##2"} {##3} }

```

```

1810 \int_set_eq:NN \l__regex_max_pos_int \l__regex_current_pos_int
1811 \__regex_query_set:nnn { } { -1 } { -2 }
1812 <trace> \trace:nxx { regex } { 1 } { initializing }
1813 \bool_gset_false:N \g__regex_success_bool
1814 \int_step_inline:nnnn
1815 \l__regex_min_state_int \c_one { \l__regex_max_state_int - \c_one }
1816 { \tex_dimen:D ##1 \c_one sp \scan_stop: }
1817 \int_set_eq:NN \l__regex_min_active_int \l__regex_max_state_int
1818 \int_set_eq:NN \l__regex_step_int \c_zero
1819 \int_set_eq:NN \l__regex_success_pos_int \l__regex_min_pos_int
1820 \int_set:Nn \l__regex_submatch_int
1821 { \c_two * \l__regex_max_state_int }
1822 \bool_set_false:N \l__regex_empty_success_bool
1823 \__regex_match_once:
1824 <trace> \trace_pop:nxx { regex } { 1 } { __regex_match }
1825 }
(End definition for \__regex_match:n.)

```

__regex_match_once: This function finds one match, then does some action defined by the `every_match` token list, which may recursively call `__regex_match_once:`. First initialize some variables: set the conditional which detects identical empty matches; this match attempt starts at the previous `success_pos`, is not yet successful, and has no submatches yet; clear the array of active threads, and put the starting state 0 in it. We are then almost ready to read our first token in the query, but we actually start one position earlier than the start, and `get` that token, so that the `last_char` will be set properly for word boundaries. Then call `__regex_match_loop:`, which runs through the query until the end or until a successful match breaks early.

```

1826 \cs_new_protected_nopar:Npn \__regex_match_once:
1827 {
1828   \if_meaning:w \c_true_bool \l__regex_empty_success_bool
1829   \cs_set_nopar:Npn \__regex_if_two_empty_matches:F
1830   { \int_compare:nNnF \l__regex_start_pos_int = \l__regex_current_pos_int }
1831   \else:
1832     \cs_set_eq:NN \__regex_if_two_empty_matches:F \use:n
1833   \fi:
1834   \int_set_eq:NN \l__regex_start_pos_int \l__regex_success_pos_int
1835   \bool_set_false:N \l__regex_match_success_bool
1836   \prop_clear:N \l__regex_current_submatches_prop
1837   \int_set_eq:NN \l__regex_max_active_int \l__regex_min_active_int
1838   \__regex_store_state:n { \l__regex_min_state_int }
1839   \int_set:Nn \l__regex_current_pos_int
1840   { \l__regex_start_pos_int - \c_one }
1841   \__regex_query_get:
1842   \__regex_match_loop:
1843   \l__regex_every_match_tl
1844 }
(End definition for \__regex_match_once:.)

```

`__regex_single_match:` For a single match, the overall success is determined by whether the only match attempt
`__regex_multi_match:n` is a success. When doing multiple matches, the overall matching is successful as soon as
any match succeeds. Perform the action #1, then find the next match.

```

1845 \cs_new_protected_nopar:Npn \__regex_single_match:
1846 {
1847   \tl_set:Nn \l__regex_every_match_tl
1848     { \bool_gset_eq:NN \g__regex_success_bool \l__regex_match_success_bool }
1849 }
1850 \cs_new_protected:Npn \__regex_multi_match:n #1
1851 {
1852   \tl_set:Nn \l__regex_every_match_tl
1853   {
1854     \if_meaning:w \c_true_bool \l__regex_match_success_bool
1855       \bool_gset_true:N \g__regex_success_bool
1856       #1
1857     \exp_after:wN \__regex_match_once:
1858   } \fi:
1859 }
1860 }

```

(End definition for `__regex_single_match:` and `__regex_multi_match:n`)

`__regex_match_loop:` At each new position, set some variables and get the new character and category from
`__regex_match_one_active:w` the query. Then unpack the array of active threads, and clear it by resetting its length
(max_active). This results in a sequence of `__regex_use_state_and_submatches:nn`
`{\state}{\prop}`, and we consider those states one by one in order. As soon as a thread
succeeds, exit the step, and, if there are threads to consider at the next position, and
we have not reached the end of the string, repeat the loop. Otherwise, the last thread
that succeeded is what `__regex_match_once:` matches. We explain the `fresh_thread`
business when describing `__regex_action_wildcard:`.

```

1861 \cs_new_protected_nopar:Npn \__regex_match_loop:
1862 {
1863   \tex_advance:D \l__regex_step_int \c_two
1864   \int_incr:N \l__regex_current_pos_int
1865   \int_set_eq:NN \l__regex_last_char_int \l__regex_current_char_int
1866   \int_set_eq:NN \l__regex_case_changed_char_int \c_max_int
1867   \__regex_query_get:
1868   \use:x
1869   {
1870     \int_set_eq:NN \l__regex_max_active_int \l__regex_min_active_int
1871     \exp_after:wN \__regex_match_one_active:w
1872     \int_use:N \l__regex_min_active_int ;
1873   }
1874   \__prg_break_point:
1875   \bool_set_false:N \l__regex_fresh_thread_bool %^^A was arg of break_point:n
1876   \if_int_compare:w \l__regex_max_active_int > \l__regex_min_active_int
1877     \if_int_compare:w \l__regex_current_pos_int < \l__regex_max_pos_int
1878       \exp_after:wN \exp_after:wN \exp_after:wN \__regex_match_loop:
1879     \fi:

```



```

1880 \fi:
1881 }
1882 \cs_new:Npn \__regex_match_one_active:w #1;
1883 {
1884   \if_int_compare:w #1 < \l__regex_max_active_int
1885     \__regex_use_state_and_submatches:nn
1886     { \__int_value:w \tex_skip:D #1 }
1887     { \tex_the:D \tex_toks:D #1 }
1888   \exp_after:wN \__regex_match_one_active:w
1889   \int_use:N \__int_eval:w #1 + \c_one \exp_after:wN ;
1890 \fi:
1891 }

```

(End definition for __regex_match_loop:.)

`__regex_query_set:nnn` The arguments are: tokens that `o` and `x` expand to one token of the query, the catcode, and the character code. Store those, and the current brace balance (used later to check for overall brace balance) in a `\muskip` register and a `\toks`, then update the `balance`.

```

1892 \cs_new_protected:Npn \__regex_query_set:nnn #1#2#3
1893 {
1894   \tex_muskip:D \l__regex_current_pos_int
1895   = \etex_glue_tomu:D
1896   #3 sp
1897   plus #2 sp
1898   minus \l__regex_balance_int sp
1899   \scan_stop:
1900   \tex_toks:D \l__regex_current_pos_int {#1}
1901   \int_incr:N \l__regex_current_pos_int
1902   \if_case:w #2 \exp_stop_f:
1903   \or: \int_incr:N \l__regex_balance_int
1904   \or: \int_decr:N \l__regex_balance_int
1905   \fi:
1906 }

```

(End definition for __regex_query_set:nnn.)

`__regex_query_get:` Extract the current character and category codes from the `\muskip` register of the current position: those are the main and the stretch components, and we need a conversion to avoid T_EX's “incompatible glue units” error.

```

1907 \cs_new_protected_nopar:Npn \__regex_query_get:
1908 {
1909   \l__regex_current_char_int
1910   = \etex_mu_toglue:D \tex_muskip:D \l__regex_current_pos_int
1911   \l__regex_current_catcode_int = \etex_glue_stretch:D
1912   \etex_mu_toglue:D \tex_muskip:D \l__regex_current_pos_int
1913 }

```

(End definition for __regex_query_get:.)

2.5.3 Using states of the nfa

`_regex_use_state:` Use the current NFA instruction. The state is initially marked as belonging to the current **step**: this allows normal free transition to repeat, but group-repeating transitions won't. Once we are done exploring all the branches it spawned, the state is marked as **step + 1**: any thread hitting it at that point will be terminated.

```

1914 \cs_new_protected_nopar:Npn \_regex\_use\_state:
1915   {
1916   \*trace>
1917     \trace:nnx { regex } { 2 } { state-\int\_use:N \l\_regex\_current\_state\_int }
1918   </trace>
1919     \tex\_dimen:D \l\_regex\_current\_state\_int
1920     = \l\_regex\_step\_int sp \scan\_stop:
1921     \tex\_the:D \tex\_toks:D \l\_regex\_current\_state\_int
1922     \tex\_dimen:D \l\_regex\_current\_state\_int
1923     = \\_int\_eval:w \l\_regex\_step\_int + \c\_one \\_int\_eval\_end: sp \scan\_stop:
1924   }
(End definition for \_regex\_use\_state:.)

```

`_regex_use_state_and_submatches:nn` This function is called as one item in the array of active threads after that array has been unpacked for a new step. Update the `current_state` and `current_submatches` and use the state if it has not yet been encountered at this step.

```

1925 \cs_new_protected:Npn \_regex\_use\_state\_and\_submatches:nn #1 #2
1926   {
1927     \int\_set:Nn \l\_regex\_current\_state\_int {#1}
1928     \if\_int\_compare:w \tex\_dimen:D \l\_regex\_current\_state\_int
1929       < \l\_regex\_step\_int
1930       \tl\_set:Nn \l\_regex\_current\_submatches\_prop {#2}
1931       \exp\_after:wN \_regex\_use\_state:
1932     \fi:
1933     \scan\_stop:
1934   }
(End definition for \_regex\_use\_state\_and\_submatches:nn.)

```

2.5.4 Actions when matching

`_regex_action_start_wildcard:` For an unanchored match, state 0 has a free transition to the next and a costly one to itself, to repeat at the next position. To catch repeated identical empty matches, we need to know if a successful thread corresponds to an empty match. The instruction resetting `\l_regex_fresh_thread_bool` may be skipped by a successful thread, hence we had to add it to `_regex_match_loop:` too.

```

1935 \cs_new_protected_nopar:Npn \_regex\_action\_start\_wildcard:
1936   {
1937     \bool\_set\_true:N \l\_regex\_fresh\_thread\_bool
1938     \\_regex\_action\_free:n {1}
1939     \bool\_set\_false:N \l\_regex\_fresh\_thread\_bool
1940     \\_regex\_action\_cost:n {0}
1941   }

```

(End definition for _regex_action_start_wildcard:.)

_regex_action_free:n These functions copy a thread after checking that the NFA state has not already been used
_regex_action_free_group:n at this position. If not, store submatches in the new state, and insert the instructions
_regex_action_free_aux:nn for that state in the input stream. Then restore the old value of \l_regex_current_
state_int and of the current submatches. The two types of free transitions differ by how
they test that the state has not been encountered yet: the **group** version is stricter, and
will not use a state if it was used earlier in the current thread, hence forcefully breaking
the loop, while the “normal” version will revisit a state when within the thread itself.

```

1942 \cs_new_protected_nopar:Npn \_regex_action_free:n
1943   { \_regex_action_free_aux:nn { > \l\_regex_step_int \else: } }
1944 \cs_new_protected_nopar:Npn \_regex_action_free_group:n
1945   { \_regex_action_free_aux:nn { < \l\_regex_step_int } }
1946 \cs_new_protected:Npn \_regex_action_free_aux:nn #1#2
1947   {
1948     \use:x
1949     {
1950       \int_add:Nn \l\_regex_current_state_int {#2}
1951       \exp_not:n
1952       {
1953         \if_int_compare:w \tex_dimen:D \l\_regex_current_state_int #1
1954           \exp_after:wN \_regex_use_state:
1955           \fi:
1956       }
1957       \int_set:Nn \l\_regex_current_state_int
1958       { \int_use:N \l\_regex_current_state_int }
1959       \tl_set:Nn \exp_not:N \l\_regex_current_submatches_prop
1960       { \exp_not:o \l\_regex_current_submatches_prop }
1961     }
1962   }

```

(End definition for _regex_action_free:n and _regex_action_free_group:n.)

_regex_action_cost:n A transition which consumes the current character and shifts the state by #1. The
resulting state is stored in the \skip array for use at the next position, and we also store
the current submatches.

```

1963 \cs_new_protected:Npn \_regex_action_cost:n #1
1964   {
1965     \exp_args:No \_regex_store_state:n
1966     { \int_use:N \_int_eval:w \l\_regex_current_state_int + #1 }
1967   }

```

(End definition for _regex_action_cost:n.)

_regex_store_state:n Put the given state in the array of \skip registers (converted to a dimension in scaled
points), and increment the length of the array. Then store the current submatch in the
_regex_store_submatches: This is done by increasing the pointer \l_regex_max_active_int, and converting the
integer to a dimension (suitable for a \skip assignment) in scaled points.

```

1968 \cs_new_protected:Npn \_regex_store_state:n #1
1969   {

```

```

1970 \__regex_store_submatches:
1971 \tex_skip:D \l__regex_max_active_int = #1 sp \scan_stop:
1972 \int_incr:N \l__regex_max_active_int
1973 }
1974 \cs_new_protected_nopar:Npn \__regex_store_submatches:
1975 {
1976 \tex_toks:D \l__regex_max_active_int \exp_after:wN
1977 { \l__regex_current_submatches_prop }
1978 }

```

(End definition for __regex_store_state:n.)

__regex_disable_submatches: Some user functions don't require tracking submatches. We get a performance improvement by simply defining the relevant functions to remove their argument and do nothing with it.

```

1979 \cs_new_protected_nopar:Npn \__regex_disable_submatches:
1980 {
1981 \cs_set_protected_nopar:Npn \__regex_store_submatches: { }
1982 \cs_set_protected:Npn \__regex_action_submatch:n ##1 { }
1983 }

```

(End definition for __regex_disable_submatches:.)

__regex_action_submatch:n Update the current submatches with the information from the current position. Maybe a bottleneck.

```

1984 \cs_new_protected:Npn \__regex_action_submatch:n #1
1985 {
1986 \prop_put:Nno \l__regex_current_submatches_prop {#1}
1987 { \int_use:N \l__regex_current_pos_int }
1988 }

```

(End definition for __regex_action_submatch:n.)

__regex_action_success: There is a successful match when an execution path reaches the last state in the NFA, unless this marks a second identical empty match. Then mark that there was a successful match; it is empty if it is "fresh"; and we store the current position and submatches. The current step is then interrupted with __prg_break:, and only paths with higher precedence are pursued further. The values stored here may be overwritten by a later success of a path with higher precedence.

```

1989 \cs_new_protected_nopar:Npn \__regex_action_success:
1990 {
1991 \__regex_if_two_empty_matches:F
1992 {
1993 \bool_set_true:N \l__regex_match_success_bool
1994 \bool_set_eq:NN \l__regex_empty_success_bool
1995 \l__regex_fresh_thread_bool
1996 \int_set_eq:NN \l__regex_success_pos_int \l__regex_current_pos_int
1997 \prop_set_eq:NN \l__regex_success_submatches_prop
1998 \l__regex_current_submatches_prop
1999 \__prg_break:
2000 }
2001 }

```

(End definition for `__regex_action_success:.`)

2.6 Replacement

2.6.1 Variables and helpers used in replacement

`\l__regex_replacement_csnames_int` The behaviour of closing braces inside a replacement text depends on whether a sequences `\c{` or `\u{` has been encountered. The number of “open” such sequences that should be closed by `}` is stored in `\l__regex_replacement_csnames_int`, and decreased by 1 by each `}`.

2002 `\int_new:N \l__regex_replacement_csnames_int`

(End definition for `\l__regex_replacement_csnames_int`. This variable is documented on page ??.)

`\l__regex_balance_tl` This token list holds the replacement text for `__regex_replacement_balance_one_match:n` while it is being built incrementally.

2003 `\tl_new:N \l__regex_balance_tl`

(End definition for `\l__regex_balance_tl`. This variable is documented on page ??.)

`__regex_replacement_balance_one_match:n` This expects as an argument the first index of a range of `\skip` registers which hold the submatch information for a given match. It can be used within an integer expression to obtain the brace balance incurred by performing the replacement on that match. This combines the braces lost by removing the match, braces added by all the submatches appearing in the replacement, and braces appearing explicitly in the replacement. Even though it is always redefined before use, we initialize it as for an empty replacement. An important property is that concatenating several calls to that function must result in a valid integer expression (hence a leading `+` in the actual definition).

2004 `\cs_new:Npn __regex_replacement_balance_one_match:n #1`

2005 `{ - __regex_submatch_balance:n {#1} }`

(End definition for `__regex_replacement_balance_one_match:n`.)

`__regex_replacement_do_one_match:n` The input is the same as `__regex_replacement_balance_one_match:n`. This function is redefined to expand to the part of the token list from the end of the previous match to a given match, followed by the replacement text. Hence concatenating the result of this function with all possible arguments (one call for each match), as well as the range from the end of the last match to the end of the string, will produce the fully replaced token list. The initialization does not matter, but we set it as for an empty replacement.

2006 `\cs_new:Npn __regex_replacement_do_one_match:n #1`

2007 `{`

2008 `__regex_query_range:nn`

2009 `{ \etex_glueshrink:D \tex_skip:D #1 }`

2010 `{ \tex_skip:D #1 }`

2011 `}`

(End definition for `__regex_replacement_do_one_match:n`.)

`_regex_replacement_exp_not:N` This function lets us navigate around the fact that the primitive `\exp_not:n` requires a braced argument. As far as I can tell, it is only needed if the user tries to include in the replacement text a control sequence set equal to a macro parameter character, such as `\c_parameter_token`. Indeed, within an x-expanding assignment, `\exp_not:N #` behaves as a single `#`, whereas `\exp_not:n {#}` behaves as a doubled `##`.

```
2012 \cs_new:Npn \_regex_replacement_exp_not:N #1 { \exp_not:n {#1} }
(End definition for \_regex_replacement_exp_not:N.)
```

2.6.2 Query and brace balance

`_regex_query_range:nn` When it is time to extract submatches from the token list, the various tokens are stored in `\toks` registers numbered from `\l__regex_min_pos_int` inclusive to `\l__regex_max_pos_int` exclusive. The function `_regex_query_range:nn {<min>} {<max>}` unpacks registers from the position `<min>` to the position `<max> - 1` included. Once this is expanded, a second x-expansion will result in the actual tokens from the query. That second expansion is only done by user functions at the very end of their operation, after checking (and correcting) the brace balance first.

```
2013 \cs_new:Npn \_regex_query_range:nn #1#2
2014 {
2015   \exp_after:wN \_regex_query_range_loop:ww
2016   \int_use:N \_int_eval:w #1 \exp_after:wN ;
2017   \int_use:N \_int_eval:w #2 ;
2018   \__prg_break_point:
2019 }
2020 \cs_new:Npn \_regex_query_range_loop:ww #1 ; #2 ;
2021 {
2022   \if_int_compare:w #1 < #2 \exp_stop_f:
2023   \else:
2024     \exp_after:wN \__prg_break:
2025     \fi:
2026     \tex_the:D \tex_toks:D #1 \exp_stop_f:
2027     \exp_after:wN \_regex_query_range_loop:ww
2028     \int_use:N \_int_eval:w #1 + \c_one ; #2 ;
2029 }
```

(End definition for `_regex_query_range:nn`.)

`_regex_query_submatch:n` When this function is called, `\skipi` holds the start and end positions for the *i*-th overall submatch as its main and stretch components. In the case of repeated matches, submatches from all the matches are put one after the other in blocks of `\l__regex_capturing_group_int \skip` registers.

```
2030 \cs_new:Npn \_regex_query_submatch:n #1
2031 {
2032   \_regex_query_range:nn
2033   { \tex_skip:D \_int_eval:w #1 }
2034   { \etex_gluestretch:D \tex_skip:D \_int_eval:w #1 }
2035 }
```

(End definition for `_regex_query_submatch:n`.)

`_regex_submatch_balance:n` Every user function must result in a balanced token list (unbalanced token lists cannot be stored by TeX). When we unpacked the query, we kept track of the brace balance as the shrink component of `\muskip` registers, hence the contribution from a given range is the difference between the shrink components of `\muskip⟨max pos⟩` and `\muskip⟨min pos⟩`. For the i -th submatch, the end-points of the range are the main and stretch components of `\skipi`. The trailing `\scan_stop:` is gobbled by `\etex_muexpr:D`, and the whole expression can be cast safely to an integer (no trailing expansion).

```

2036 \cs_new_protected:Npn \_regex_submatch_balance:n #1
2037 {
2038   \etex_glueshrink:D \etex_mutogluue:D \etex_muexpr:D
2039   \tex_muskip:D \etex_gluestretch:D \tex_skip:D #1
2040   - \tex_muskip:D \tex_skip:D #1
2041   \scan_stop:
2042 }

```

(End definition for `_regex_submatch_balance:n`. This function is documented on page ??.)

2.6.3 Framework

`_regex_replacement:n` The replacement text is built incrementally by abusing `\toks` within a group (see `l3tl-build`). We keep track in `\l__regex_balance_int` of the balance of explicit begin- and end-group tokens and `\l__regex_balance_tl` will consist of some code to compute the brace balance from submatches (see its description). Detect unescaped right braces, and escaped characters, with trailing `\prg_do_nothing:` because some of the later function look-ahead. Once the whole replacement text has been parsed, make sure that there is no open csnames. Finally, define the `balance_one_match` and `do_one_match` functions.

```

2043 \cs_new_protected:Npn \_regex_replacement:n #1
2044 {
2045   <trace> \trace_push:nnn { regex } { 1 } { \_regex_replacement:n }
2046   \_tl_build:Nw \l__regex_internal_a_tl
2047   \int_zero:N \l__regex_balance_int
2048   \tl_clear:N \l__regex_balance_tl
2049   \_regex_escape_use:nnnn
2050   {
2051     \if_charcode:w \c_right_brace_str ##1
2052       \_regex_replacement_rbrace:N \else: \_tl_build_one:n \fi: ##1
2053   }
2054   { \_regex_replacement_escaped:N ##1 }
2055   { \_tl_build_one:n ##1 }
2056   {#1}
2057   \prg_do_nothing: \prg_do_nothing:
2058   \if_int_compare:w \l__regex_replacement_csnames_int > \c_zero
2059     \_msg_kernel_error:nnx { regex } { replacement-missing-rbrace }
2060     { \int_use:N \l__regex_replacement_csnames_int }
2061     \_tl_build_one:x
2062     { \prg_replicate:nn \l__regex_replacement_csnames_int \cs_end: }
2063   \fi:
2064   \cs_gset:Npx \_regex_replacement_balance_one_match:n ##1
2065   {

```

```

2066         + \int_use:N \l__regex_balance_int
2067         \l__regex_balance_tl
2068         - \__regex_submatch_balance:n {##1}
2069     }
2070     \__tl_build_end:
2071     \exp_args:No \__regex_replacement_aux:n \l__regex_internal_a_tl
2072     <trace> \trace_pop:nnn { regex } { 1 } { __regex_replacement:n }
2073 }
2074 \cs_new_protected:Npn \__regex_replacement_aux:n #1
2075 {
2076     \cs_set:Npn \__regex_replacement_do_one_match:n ##1
2077     {
2078         \__regex_query_range:nn
2079         { \etex_glueshrink:D \tex_skip:D ##1 }
2080         { \tex_skip:D ##1 }
2081         #1
2082     }
2083 }
(End definition for \__regex_replacement:n.)

```

`__regex_replacement_escaped:N` As in parsing a regular expression, we use an auxiliary built from #1 if defined. Otherwise, check for escaped digits (standing from submatches from 0 to 9): anything else is a raw character.

```

2084 \cs_new_protected:Npn \__regex_replacement_escaped:N #1
2085 {
2086     \cs_if_exist_use:cF { __regex_replacement_#1:w }
2087     {
2088         \if_int_compare:w \c_one < 1#1 \exp_stop_f:
2089         \__regex_replacement_put_submatch:n {#1}
2090         \else:
2091         \__tl_build_one:n #1
2092         \fi:
2093     }
2094 }
(End definition for \__regex_replacement_escaped:N.)

```

2.6.4 Submatches

`__regex_replacement_put_submatch:n` Insert a submatch in the replacement text. This is dropped if the submatch number is larger than the number of capturing groups. Unless the submatch appears inside a `\c{...}` or `\u{...}` construction, it must be taken into account in the brace balance. Here, ##1 will receive a pointer to the 0-th submatch for a given match. We cannot use `\int_eval:n` because it is expandable, and would be expanded too early (short of adding `\exp_not:N`, making the code messy again).

```

2095 \cs_new_protected:Npn \__regex_replacement_put_submatch:n #1
2096 {
2097     \if_int_compare:w #1 < \l__regex_capturing_group_int
2098     \__tl_build_one:n { \__regex_query_submatch:n { #1 + ##1 } }

```



```

2099     \if_int_compare:w \l__regex_replacement_csnames_int = \c_zero
2100     \tl_put_right:Nn \l__regex_balance_tl
2101     { + \__regex_submatch_balance:n { \__int_eval:w #1+##1 \__int_eval_end: } }
2102     \fi:
2103     \fi:
2104 }

```

(End definition for __regex_replacement_put_submatch:n.)

__regex_replacement_g:w An ugly method to grab digits for the \g escape sequence. At the end of the run of digits, check that it ends with a right brace.

__regex_replacement_g_digits:NN

```

2105 \cs_new_protected:Npn \__regex_replacement_g:w #1#2
2106 {
2107     \str_if_eq_x:nnTF { #1#2 } { \__tl_build_one:n \c_left_brace_str }
2108     {
2109         \int_zero:N \l__regex_internal_a_int
2110         \__regex_replacement_g_digits:NN
2111     }
2112     { \__regex_replacement_error:NNN g #1 #2 }
2113 }
2114 \cs_new_protected:Npn \__regex_replacement_g_digits:NN #1#2
2115 {
2116     \token_if_eq_meaning:NNTF #1 \__tl_build_one:n
2117     {
2118         \if_int_compare:w \c_one < 1#2 \exp_stop_f:
2119         \int_set:Nn \l__regex_internal_a_int
2120         { \c_ten * \l__regex_internal_a_int + #2 }
2121         \exp_after:wN \use_i:nnn
2122         \exp_after:wN \__regex_replacement_g_digits:NN
2123     \else:
2124         \exp_after:wN \__regex_replacement_error:NNN
2125         \exp_after:wN g
2126     \fi:
2127 }
2128 {
2129     \if_meaning:w \__regex_replacement_rbrace:N #1
2130     \exp_args:No \__regex_replacement_put_submatch:n
2131     { \int_use:N \l__regex_internal_a_int }
2132     \exp_after:wN \use_none:nn
2133 \else:
2134     \exp_after:wN \__regex_replacement_error:NNN
2135     \exp_after:wN g
2136 \fi:
2137 }
2138 #1 #2
2139 }

```

(End definition for __regex_replacement_g:w and __regex_replacement_g_digits:NN.)

2.6.5 Csnames in replacement

`__regex_replacement_c:w` `\c` can be followed by a left brace, or by a letter for which we have defined a way to produce that category of characters. The appropriate definitions for catcodes are introduced later. For control sequences, if we are within a control sequence, convert the token list to a string, otherwise simply prevent expansion, with a weird cross-over between `\exp_not:n` and `\exp_not:N` (see this helper's description for an explanation).

```

2140 \cs_new_protected:Npn __regex_replacement_c:w #1#2
2141 {
2142   \token_if_eq_meaning:NNTF #1 \__tl_build_one:n
2143   {
2144     \cs_if_exist_use:cF { __regex_replacement_c_#2:w }
2145     { __regex_replacement_error:NNN c #1#2 }
2146   }
2147   { __regex_replacement_error:NNN c #1#2 }
2148 }
2149 \cs_new_protected_nopar:cpn { __regex_replacement_c_ \c_left_brace_str :w }
2150 {
2151   \if_case:w \l__regex_replacement_csnames_int
2152   \__tl_build_one:n
2153   { \exp_not:n { \exp_after:wN __regex_replacement_exp_not:N \cs:w } }
2154   \else:
2155     \__tl_build_one:n { \exp_not:n { \exp_after:wN \tl_to_str:N \cs:w } }
2156   \fi:
2157   \int_incr:N \l__regex_replacement_csnames_int
2158 }

```

(End definition for __regex_replacement_c:w.)

`__regex_replacement_u:w` Check that `\u` is followed by a left brace. If so, start a control sequence with `\cs:w`, which is then unpacked either with `\exp_not:V` or `\tl_to_str:V` depending on the current context.

```

2159 \cs_new_protected:Npn __regex_replacement_u:w #1#2
2160 {
2161   \str_if_eq_x:nnTF { #1#2 } { \__tl_build_one:n \c_left_brace_str }
2162   {
2163     \if_case:w \l__regex_replacement_csnames_int
2164     \__tl_build_one:n { \exp_not:n { \exp_after:wN \exp_not:V \cs:w } }
2165     \else:
2166       \__tl_build_one:n { \exp_not:n { \exp_after:wN \tl_to_str:V \cs:w } }
2167     \fi:
2168     \int_incr:N \l__regex_replacement_csnames_int
2169   }
2170   { __regex_replacement_error:NNN u #1#2 }
2171 }

```

(End definition for __regex_replacement_u:w.)

`__regex_replacement_rbrace:N` Within a `\c{...}` or `\u{...}` construction, end the control sequence, and decrease the brace count. Otherwise, this is a raw right brace.

```

2172 \cs_new_protected:Npn __regex_replacement_rbrace:N #1

```

```

2173 {
2174   \if_int_compare:w \l__regex_replacement_csnames_int > \c_zero
2175     \__tl_build_one:n \cs_end:
2176     \int_decr:N \l__regex_replacement_csnames_int
2177   \else:
2178     \__tl_build_one:n #1
2179   \fi:
2180 }

```

(End definition for `__regex_replacement_rbrace:N`.)

2.6.6 Characters in replacement

We will need to change the category code of the null character many times, hence work in a group. The catcode-specific macros below are defined in alphabetical order; if you are trying to understand the code, start from the end of the alphabet as those categories are simpler than active or begin-group.

```

2181 \group_begin:

```

`__regex_replacement_char:nNN` The only way to produce an arbitrary character-catcode pair is to use the `\lowercase` or `\uppercase` primitives. This is a wrapper for our purposes. The first argument is the null character with various catcodes. The second and third arguments are grabbed from the input stream: `#3` is the character whose character code to reproduce.

```

2182 \cs_new_protected:Npn \__regex_replacement_char:nNN #1#2#3
2183 {
2184   \if_meaning:w \prg_do_nothing: #3
2185     \__msg_kernel_error:nn { regex } { replacement-catcode-end }
2186   \else:
2187     \tex_lccode:D \c_zero = '#3 \scan_stop:
2188     \tl_to_lowercase:n { \__tl_build_one:n {#1} }
2189   \fi:
2190 }

```

(End definition for `__regex_replacement_char:nNN`.)

`__regex_replacement_c_A:w` For an active character, expansion must be avoided, twice because we later do two x-expansions, to unpack `\toks` for the query, and to expand their contents to tokens of the query.

```

2191 \char_set_catcode_active:N \^^@
2192 \cs_new_protected_nopar:Npn \__regex_replacement_c_A:w
2193 { \__regex_replacement_char:nNN { \exp_not:n { \exp_not:N \^^@ } } }

```

(End definition for `__regex_replacement_c_A:w`.)

`__regex_replacement_c_B:w` An explicit begin-group token increases the balance, unless within a `\c{...}` or `\u{...}` construction. Add the desired begin-group character, using the standard `\if_false:` trick. We eventually x-expand twice. The first time must yield a balanced token list, and the second one gives the bare begin-group token. The `\exp_after:wN` is not strictly needed, but is more consistent with l3tl-analysis.

```

2194 \char_set_catcode_group_begin:N \^^@

```

```

2195 \cs_new_protected_nopar:Npn \__regex_replacement_c_B:w
2196 {
2197   \if_int_compare:w \l__regex_replacement_csnames_int = \c_zero
2198     \int_incr:N \l__regex_balance_int
2199     \fi:
2200     \__regex_replacement_char:nNN
2201     { \exp_not:n { \exp_after:wN ^^@ \if_false: } \fi: } }
2202 }
(End definition for \__regex_replacement_c_B:w.)

```

`__regex_replacement_c_C:w` This is not quite catcode-related: when the user requests a character with category “control sequence”, the one-character control symbol is returned. As for the active character, we prepare for two x-expansions.

```

2203 \cs_new_protected:Npn \__regex_replacement_c_C:w #1#2
2204 { \__tl_build_one:n { \exp_not:N \exp_not:N \exp_not:c {#2} } }
(End definition for \__regex_replacement_c_C:w.)

```

`__regex_replacement_c_D:w` Subscripts fit the mould: `\lowercase` the null byte with the correct category.

```

2205 \char_set_catcode_math_subscript:N \^^@
2206 \cs_new_protected_nopar:Npn \__regex_replacement_c_D:w
2207 { \__regex_replacement_char:nNN { ^^@ } }
(End definition for \__regex_replacement_c_D:w.)

```

`__regex_replacement_c_E:w` Similar to the begin-group case, the second x-expansion produces the bare end-group token.

```

2208 \char_set_catcode_group_end:N \^^@
2209 \cs_new_protected_nopar:Npn \__regex_replacement_c_E:w
2210 {
2211   \if_int_compare:w \l__regex_replacement_csnames_int = \c_zero
2212     \int_decr:N \l__regex_balance_int
2213     \fi:
2214     \__regex_replacement_char:nNN
2215     { \exp_not:n { \if_false: { \fi: ^^@ } } }
2216 }
(End definition for \__regex_replacement_c_E:w.)

```

`__regex_replacement_c_L:w` Simply `\lowercase` a letter null byte to produce an arbitrary letter.

```

2217 \char_set_catcode_letter:N \^^@
2218 \cs_new_protected_nopar:Npn \__regex_replacement_c_L:w
2219 { \__regex_replacement_char:nNN { ^^@ } }
(End definition for \__regex_replacement_c_L:w.)

```

`__regex_replacement_c_M:w` No surprise here, we lowercase the null math toggle.

```

2220 \char_set_catcode_math_toggle:N \^^@
2221 \cs_new_protected_nopar:Npn \__regex_replacement_c_M:w
2222 { \__regex_replacement_char:nNN { ^^@ } }
(End definition for \__regex_replacement_c_M:w.)

```

`_regex_replacement_c_0:w` Lowercase an other null byte.

```

2223 \char_set_catcode_other:N \^^@
2224 \cs_new_protected_nopar:Npn \_regex_replacement_c_0:w
2225 { \_regex_replacement_char:nNN { ^^@ } }
(End definition for \_regex_replacement_c_0:w.)

```

`_regex_replacement_c_P:w` For macro parameters, expansion is a tricky issue. We need to prepare for two x-expansions and passing through various macro definitions. Note that we cannot replace one `\exp_not:n` by doubling the macro parameter characters because this would misbehave if a mischievous user asks for `\c{\cP\#}`, since that macro parameter character would be doubled.

```

2226 \char_set_catcode_parameter:N \^^@
2227 \cs_new_protected_nopar:Npn \_regex_replacement_c_P:w
2228 {
2229   \_regex_replacement_char:nNN
2230   { \exp_not:n { \exp_not:n { ^^@^^@^^@^^@ } } }
2231 }
(End definition for \_regex_replacement_c_P:w.)

```

`_regex_replacement_c_S:w` Spaces are normalized on input by T_EX to have character code 32. It is in fact impossible to get a token with character code 0 and category code 10. Hence we use 32 instead of 0 as our base character.

```

2232 \cs_new_protected:Npn \_regex_replacement_c_S:w #1#2
2233 {
2234   \if_meaning:w \prg_do_nothing: #2
2235   \__msg_kernel_error:nn { regex } { replacement-catcode-end }
2236   \else:
2237     \if_int_compare:w '#2 = \c_zero
2238     \__msg_kernel_error:nn { regex } { replacement-null-space }
2239     \fi:
2240     \tex_lccode:D 32 = '#2 \scan_stop:
2241     \tl_to_lowercase:n { \_tl_build_one:n {~} }
2242     \fi:
2243   }
(End definition for \_regex_replacement_c_S:w.)

```

`_regex_replacement_c_T:w` No surprise for alignment tabs here. Those are surrounded by the appropriate braces whenever necessary, hence they don't cause trouble in alignment settings.

```

2244 \char_set_catcode_alignment:N \^^@
2245 \cs_new_protected_nopar:Npn \_regex_replacement_c_T:w
2246 { \_regex_replacement_char:nNN { ^^@ } }
(End definition for \_regex_replacement_c_T:w.)

```

`_regex_replacement_c_U:w` Simple call to `_regex_replacement_char:nNN` which lowercases the math superscript `^^@`.

```

2247 \char_set_catcode_math_superscript:N \^^@
2248 \cs_new_protected_nopar:Npn \_regex_replacement_c_U:w
2249 { \_regex_replacement_char:nNN { ^^@ } }

```

(End definition for `_regex_replacement_c_U:w`.)

Restore the catcode of the null byte.

2250 `\group_end:`

2.6.7 An error

`_regex_replacement_error:NNN` Simple error reporting by calling one of the messages `replacement-c`, `replacement-g`, or `replacement-u`.

```
2251 \cs_new_protected:Npn \_regex_replacement_error:NNN #1#2#3
2252 {
2253   \_msg_kernel_error:nxx { regex } { replacement-#1 } {#3}
2254   #2 #3
2255 }
```

(End definition for `_regex_replacement_error:NNN`.)

2.7 User functions

`\regex_new:N` Before being assigned a sensible value, a regex variable matches nothing.

```
2256 \cs_new_protected:Npn \regex_new:N #1
2257 { \cs_new_eq:NN #1 \c__regex_no_match_regex }
```

(End definition for `\regex_new:N`. This function is documented on page 7.)

`\regex_set:Nn` Compile, then store the result in the user variable with the appropriate assignment function.
`\regex_gset:Nn`
`\regex_const:Nn`

```
2258 \cs_new_protected_nopar:Npn \regex_set:Nn #1#2
2259 {
2260   \_regex_compile:n {#2}
2261   \tl_set_eq:NN #1 \l__regex_internal_regex
2262 }
2263 \cs_new_protected_nopar:Npn \regex_gset:Nn #1#2
2264 {
2265   \_regex_compile:n {#2}
2266   \tl_gset_eq:NN #1 \l__regex_internal_regex
2267 }
2268 \cs_new_protected_nopar:Npn \regex_const:Nn #1#2
2269 {
2270   \_regex_compile:n {#2}
2271   \tl_const:Nx #1 { \exp_not:o \l__regex_internal_regex }
2272 }
```

(End definition for `\regex_set:Nn`, `\regex_gset:Nn`, and `\regex_const:Nn`. These functions are documented on page 7.)

`\regex_show:N` User functions: the `n` variant requires compilation first. Then show the variable with some appropriate text. The auxiliary `_regex_show:Nx` is defined in a different section.
`\regex_show:n`

```
2273 \cs_new_protected:Npn \regex_show:n #1
2274 {
2275   \_regex_compile:n {#1}
2276   \_regex_show:Nx \l__regex_internal_regex
```

```

2277     { { \tl_to_str:n {#1} } }
2278   }

```

```

2279 \cs_new_protected:Npn \regex_show:N #1
2280 { \__regex_show:Nx #1 { variable~\token_to_str:N #1 } }

```

(End definition for \regex_show:N and \regex_show:n. These functions are documented on page 7.)

\regex_match:nnTF
\regex_match:NnTF

Those conditionals are based on a common auxiliary defined later. Its first argument builds the NFA corresponding to the regex, and the second argument is the query token list. Once we have performed the match, convert the resulting boolean to \prg_return_true: or false.

```

2281 \prg_new_protected_conditional:Npnn \regex_match:nn #1#2 { T , F , TF }
2282 {
2283   \__regex_if_match:nn { \__regex_build:n {#1} } {#2}
2284   \__regex_return:
2285 }
2286 \prg_new_protected_conditional:Npnn \regex_match:Nn #1#2 { T , F , TF }
2287 {
2288   \__regex_if_match:nn { \__regex_build:N #1 } {#2}
2289   \__regex_return:
2290 }

```

(End definition for \regex_match:nnTF and \regex_match:NnTF. These functions are documented on page ??.)

\regex_count:nnN
\regex_count:NnN

Again, use an auxiliary whose first argument builds the NFA.

```

2291 \cs_new_protected:Npn \regex_count:nnN #1
2292 { \__regex_count:nnN { \__regex_build:n {#1} } }
2293 \cs_new_protected:Npn \regex_count:NnN #1
2294 { \__regex_count:nnN { \__regex_build:N #1 } }

```

(End definition for \regex_count:nnN and \regex_count:NnN. These functions are documented on page ??.)

\regex_extract_once:nnN
\regex_extract_once:NnN
\regex_extract_all:nnN
\regex_extract_all:NnN
\regex_replace_once:nnN
\regex_replace_once:NnN
\regex_replace_all:nnN
\regex_replace_all:NnN
\regex_split:nnN
\regex_split:NnN

We define here 40 user functions, following a common pattern in terms of :nnN auxiliaries, defined in the coming subsections. The auxiliary is handed __regex_build:n or __regex_build:N with the appropriate regex argument, then all other necessary arguments (replacement text, token list, etc. The conditionals call __regex_return: to return either true or false once matching has been performed.

```

2295 \cs_set_protected:Npn \__regex_tmp:w #1#2#3
2296 {
2297   \cs_new_protected:Npn #2 ##1 { #1 { \__regex_build:n {##1} } }
2298   \cs_new_protected:Npn #3 ##1 { #1 { \__regex_build:N ##1 } }
2299   \prg_new_protected_conditional:Npnn #2 ##1##2##3 { T , F , TF }
2300   { #1 { \__regex_build:n {##1} } {##2} ##3 \__regex_return: }
2301   \prg_new_protected_conditional:Npnn #3 ##1##2##3 { T , F , TF }
2302   { #1 { \__regex_build:N ##1 } {##2} ##3 \__regex_return: }
2303 }
2304 \__regex_tmp:w \__regex_extract_once:nnN
2305 \regex_extract_once:nnN \regex_extract_once:NnN
2306 \__regex_tmp:w \__regex_extract_all:nnN
2307 \regex_extract_all:nnN \regex_extract_all:NnN

```

\regex_extract_once:nnNTF
\regex_extract_once:NnNTF
\regex_extract_all:nnNTF
\regex_extract_all:NnNTF
\regex_replace_once:nnNTF
\regex_replace_once:NnNTF
\regex_replace_all:nnNTF
\regex_replace_all:NnNTF
\regex_split:nnNTF
\regex_split:NnNTF

```

2308 \__regex_tmp:w \__regex_replace_once:nnN
2309 \regex_replace_once:nnN \regex_replace_once:NnN
2310 \__regex_tmp:w \__regex_replace_all:nnN
2311 \regex_replace_all:nnN \regex_replace_all:NnN
2312 \__regex_tmp:w \__regex_split:nnN \regex_split:nnN \regex_split:NnN
(End definition for \regex_extract_once:nnN and others. These functions are documented on page ??.)

```

2.7.1 Variables and helpers for user functions

`\l__regex_match_count_int` The number of matches found so far is stored in `\l__regex_match_count_int`. This is only used in the `\regex_count:nnN` functions.

```

2313 \int_new:N \l__regex_match_count_int
(End definition for \l__regex_match_count_int. This variable is documented on page ??.)

```

`__regex_begin` `__regex_end` Those flags are raised to indicate extra begin-group or end-group tokens when extracting submatches.

```

2314 \flag_new:n { __regex_begin }
2315 \flag_new:n { __regex_end }
(End definition for __regex_begin and __regex_end. These variables are documented on page ??.)

```

`\l__regex_submatch_int` `\l__regex_zeroth_submatch_int` The end-points of each submatch are stored as main and stretch components of `\skip⟨submatch⟩`, where `⟨submatch⟩` ranges from `\l__regex_max_state_int` (inclusive) to `\l__regex_submatch_int` (exclusive). Each successful match comes with a 0-th submatch (the full match), and one match for each capturing group: submatches corresponding to the last successful match are labelled starting at `zeroth_submatch`. Additionally, the shrink component of this 0-th submatch is the position at which that match attempt started: this is used for splitting and replacements.

```

2316 \int_new:N \l__regex_submatch_int
2317 \int_new:N \l__regex_zeroth_submatch_int
(End definition for \l__regex_submatch_int and \l__regex_zeroth_submatch_int. These variables are documented on page ??.)

```

`__regex_return:` This function triggers either `\prg_return_false:` or `\prg_return_true:` as appropriate to whether a match was found or not. It is used by all user conditionals.

```

2318 \cs_new_protected_nopar:Npn \__regex_return:
2319 {
2320   \if_meaning:w \c_true_bool \g__regex_success_bool
2321     \prg_return_true:
2322   \else:
2323     \prg_return_false:
2324   \fi:
2325 }
(End definition for \__regex_return:.)

```


2.7.2 Matching

`__regex_if_match:nn` We don't track submatches, and stop after a single match. Build the NFA with #1, and perform the match on the query #2.

```
2326 \cs_new_protected:Npn \__regex_if_match:nn #1#2
2327 {
2328   \group_begin:
2329     \__regex_disable_submatches:
2330     \__regex_single_match:
2331     #1
2332     \__regex_match:n {#2}
2333   \group_end:
2334 }
```

(End definition for __regex_if_match:nn.)

`__regex_count:nnN` Again, we don't care about submatches. Instead of aborting after the first “longest match” is found, we search for multiple matches, incrementing `\l__regex_match_count_int` every time to record the number of matches. Build the NFA and match. At the end, store the result in the user's variable.

```
2335 \cs_new_protected:Npn \__regex_count:nnN #1#2#3
2336 {
2337   \group_begin:
2338     \__regex_disable_submatches:
2339     \int_zero:N \l__regex_match_count_int
2340     \__regex_multi_match:n { \int_incr:N \l__regex_match_count_int }
2341     #1
2342     \__regex_match:n {#2}
2343     \exp_args:NNNo
2344   \group_end:
2345   \int_set:Nn #3 { \int_use:N \l__regex_match_count_int }
2346 }
```

(End definition for __regex_count:nnN.)

2.7.3 Extracting submatches

`__regex_extract_once:nnN` Match once or multiple times. After each match (or after the only match), extract the submatches using `__regex_extract:.` At the end, store the sequence containing all the submatches into the user variable #3 after closing the group.

```
2347 \cs_new_protected:Npn \__regex_extract_once:nnN #1#2#3
2348 {
2349   \group_begin:
2350     \__regex_single_match:
2351     #1
2352     \__regex_match:n {#2}
2353     \__regex_extract:
2354     \__regex_group_end_extract_seq:N #3
2355   }
2356 \cs_new_protected:Npn \__regex_extract_all:nnN #1#2#3
2357 {
```

```

2358 \group_begin:
2359   \_regex_multi_match:n { \_regex_extract: }
2360   #1
2361   \_regex_match:n {#2}
2362   \_regex_group_end_extract_seq:N #3
2363 }
(End definition for \_regex_extract_once:nnN and \_regex_extract_all:nnN.)

```

`_regex_split:nnN` Splitting at submatches is a bit more tricky. For each match, extract all submatches, and replace the zeroth submatch by the part of the query between the start of the match attempt and the start of the zeroth submatch. This is inhibited if the delimiter matched an empty token list at the start of this match attempt. After the last match, store the last part of the token list, which ranges from the start of the match attempt to the end of the query. This step is inhibited if the last match was empty and at the very end: decrement `\l__regex_submatch_int`, which controls which `\skip` registers will be used.

```

2364 \cs_new_protected:Npn \_regex_split:nnN #1#2#3
2365 {
2366   \group_begin:
2367   \_regex_multi_match:n
2368   {
2369     \if_int_compare:w \l__regex_start_pos_int < \l__regex_success_pos_int
2370     \_regex_extract:
2371     \tex_skip:D \l__regex_zeroth_submatch_int
2372     = \l__regex_start_pos_int sp
2373     plus \tex_skip:D \l__regex_zeroth_submatch_int \scan_stop:
2374     \fi:
2375   }
2376   #1
2377   \_regex_match:n {#2}
2378   <assert>\assert_int:n { \l__regex_current_pos_int = \l__regex_max_pos_int }
2379   \tex_skip:D \l__regex_submatch_int
2380   = \l__regex_start_pos_int sp plus \l__regex_max_pos_int sp \scan_stop:
2381   \int_incr:N \l__regex_submatch_int
2382   \if_meaning:w \c_true_bool \l__regex_empty_success_bool
2383   \if_int_compare:w \l__regex_start_pos_int = \l__regex_max_pos_int
2384   \int_decr:N \l__regex_submatch_int
2385   \fi:
2386   \fi:
2387   \_regex_group_end_extract_seq:N #3
2388 }
(End definition for \_regex_split:nnN.)

```

`_regex_group_end_extract_seq:N` The end-points of submatches are stored as the main and stretch components of `\skip` registers from `\l__regex_max_state_int` to `\l__regex_submatch_int` (exclusive). Extract the relevant ranges into `\l__regex_internal_a_tl`. We detect unbalanced results using the two flags `@@_begin` and `@@_end`, raised whenever we see too many begin-group or end-group tokens in a submatch. We disable `_seq_item:n` to prevent two x-expansions.

```

2389 \cs_new_protected:Npn \__regex_group_end_extract_seq:N #1
2390 {
2391     \cs_set_eq:NN \__seq_item:n \scan_stop:
2392     \flag_clear:n { __regex_begin }
2393     \flag_clear:n { __regex_end }
2394     \tl_set:Nx \l__regex_internal_a_tl
2395     {
2396         \s__seq
2397         \int_step_function:nnnN
2398             { \c_two * \l__regex_max_state_int }
2399             \c_one
2400             { \l__regex_submatch_int - \c_one }
2401             \__regex_extract_seq_aux:n
2402     }
2403     \int_compare:nNnF
2404         { \flag_height:n { __regex_begin } + \flag_height:n { __regex_end } }
2405         = \c_zero
2406     {
2407         \__msg_kernel_error:nnxxx { regex } { result-unbalanced }
2408         { splitting-or-extracting-submatches }
2409         { \flag_height:n { __regex_end } }
2410         { \flag_height:n { __regex_begin } }
2411     }
2412     \use:x
2413     {
2414         \group_end:
2415         \tl_set:Nn \exp_not:N #1 { \l__regex_internal_a_tl }
2416     }
2417 }

```

(End definition for __regex_group_end_extract_seq:N.)

__regex_extract_seq_aux:n
__regex_extract_seq_aux:ww

The :n auxiliary builds one item of the sequence of submatches. First compute the brace balance of the submatch, then extract the submatch from the query, adding the appropriate braces and raising a flag if the submatch is not balanced.

```

2418 \cs_new:Npn \__regex_extract_seq_aux:n #1
2419 {
2420     \__seq_item:n
2421     {
2422         \exp_after:wN \__regex_extract_seq_aux:ww
2423         \__int_value:w \__regex_submatch_balance:n {#1} ; #1;
2424     }
2425 }
2426 \cs_new:Npn \__regex_extract_seq_aux:ww #1; #2;
2427 {
2428     \if_int_compare:w #1 < \c_zero
2429         \flag_raise:n { __regex_end }
2430         \prg_replicate:nn {-#1} { \exp_not:n { { \if_false: } \fi: } }
2431     \fi:
2432     \__regex_query_submatch:n {#2}

```

```

2433 \if_int_compare:w #1 > \c_zero
2434   \flag_raise:n { __regex_begin }
2435   \prg_replicate:nn {#1} { \exp_not:n { \if_false: { \fi: } } }
2436   \fi:
2437 }

```

(End definition for `__regex_extract_seq_aux:n` and `__regex_extract_seq_aux:wn`.)

`__regex_extract:` Our task here is to extract from the property list `\l__regex_success_submatches_prop` the list of end-points of submatches, and store them in `\skip` registers, from `\l__regex_zeroth_submatch_int` upwards. We begin by emptying those `\skip` registers. Then for each $\langle key \rangle$ – $\langle value \rangle$ pair in the property list update the appropriate `\skip` component. This is somewhat a hack: the $\langle key \rangle$ is a non-negative integer followed by `<` or `>`, which we use in a comparison to `-1`. At the end, store the information about the position at which the match attempt started, as a shrink component.

```

2438 \cs_new_protected_nopar:Npn \__regex_extract:
2439 {
2440   \if_meaning:w \c_true_bool \g__regex_success_bool
2441     \int_set_eq:NN \l__regex_zeroth_submatch_int \l__regex_submatch_int
2442     \prg_replicate:nn \l__regex_capturing_group_int
2443     {
2444       \tex_skip:D \l__regex_submatch_int \c_zero sp \scan_stop:
2445       \int_incr:N \l__regex_submatch_int
2446     }
2447   \prop_map_inline:Nn \l__regex_success_submatches_prop
2448   {
2449     \if_int_compare:w ##1 \c_minus_one
2450       \exp_after:wN \__regex_extract_e:wn \__int_value:w
2451     \else:
2452       \exp_after:wN \__regex_extract_b:wn \__int_value:w
2453     \fi:
2454     \__int_eval:w \l__regex_zeroth_submatch_int + ##1 {##2}
2455   }
2456   \tex_skip:D \l__regex_zeroth_submatch_int
2457   = \tex_the:D \tex_skip:D \l__regex_zeroth_submatch_int
2458   minus \l__regex_start_pos_int sp \scan_stop:
2459   \fi:
2460 }
2461 \cs_new_protected:Npn \__regex_extract_b:wn #1 < #2
2462 {
2463   \tex_skip:D #1 = #2 sp
2464   plus \etex_gluestretch:D \tex_skip:D #1 \scan_stop:
2465 }
2466 \cs_new_protected:Npn \__regex_extract_e:wn #1 > #2
2467 {
2468   \tex_skip:D #1
2469   = 1 \tex_skip:D #1 plus #2 sp \scan_stop:
2470 }

```

(End definition for `__regex_extract:`, `__regex_extract_b:wn`, and `__regex_extract_e:wn`.)

2.7.4 Replacement

`__regex_replace_once:nnN` Build the NFA and the replacement functions, then find a single match. If the match failed, simply exit the group. Otherwise, we do the replacement. Extract submatches. Compute the brace balance corresponding to replacing this match by the replacement (this depends on submatches). Prepare the replaced token list: the replacement function produces the tokens from the start of the query to the start of the match and the replacement text for this match; we need to add the tokens from the end of the match to the end of the query. Finally, store the result in the user's variable after closing the group: this step involves an additional `x`-expansion, and checks that braces are balanced in the final result.

```

2471 \cs_new_protected:Npn \__regex_replace_once:nnN #1#2#3
2472 {
2473   \group_begin:
2474     \__regex_single_match:
2475     #1
2476     \__regex_replacement:n {#2}
2477     \exp_args:No \__regex_match:n { #3 }
2478     \if_meaning:w \c_false_bool \g__regex_success_bool
2479     \group_end:
2480   \else:
2481     \__regex_extract:
2482     \int_set:Nn \l__regex_balance_int
2483     {
2484       \__regex_replacement_balance_one_match:n
2485       { \l__regex_zeroth_submatch_int }
2486     }
2487     \tl_set:Nx \l__regex_internal_a_tl
2488     {
2489       \__regex_replacement_do_one_match:n { \l__regex_zeroth_submatch_int }
2490       \__regex_query_range:nn
2491       { \etex_gluestretch:D \tex_skip:D \l__regex_zeroth_submatch_int }
2492       { \l__regex_max_pos_int }
2493     }
2494     \__regex_group_end_replace:N #3
2495   \fi:
2496 }

```

(End definition for __regex_replace_once:nnN.)

`__regex_replace_all:nnN` Match multiple times, and for every match, extract submatches and additionally store the position at which the match attempt started (as the shrink component of a `\skip` register). The `\skip` registers from `\l__regex_max_state_int` to `\l__regex_submatch_int` hold information about submatches of every match in order; each match corresponds to `\l__regex_capturing_group_int` consecutive `\skip` registers. Compute the brace balance corresponding to doing all the replacements: this is the sum of brace balances for replacing each match. Join together the replacement texts for each match (including the part of the query before the match), and the end of the query.

```

2497 \cs_new_protected:Npn \__regex_replace_all:nnN #1#2#3
2498 {

```

```

2499 \group_begin:
2500 \__regex_multi_match:n { \__regex_extract: }
2501 #1
2502 \__regex_replacement:n {#2}
2503 \exp_args:No \__regex_match:n {#3}
2504 \int_set:Nn \l__regex_balance_int
2505 {
2506 0
2507 \int_step_function:nnnN
2508 { \c_two * \l__regex_max_state_int }
2509 \l__regex_capturing_group_int
2510 { \l__regex_submatch_int - \c_one }
2511 \__regex_replacement_balance_one_match:n
2512 }
2513 \tl_set:Nx \l__regex_internal_a_tl
2514 {
2515 \int_step_function:nnnN
2516 { \c_two * \l__regex_max_state_int }
2517 \l__regex_capturing_group_int
2518 { \l__regex_submatch_int - \c_one }
2519 \__regex_replacement_do_one_match:n
2520 \__regex_query_range:nn
2521 \l__regex_start_pos_int \l__regex_max_pos_int
2522 }
2523 \__regex_group_end_replace:N #3
2524 }

```

(End definition for __regex_replace_all:nnN.)

__regex_group_end_replace:N

If the brace balance is not 0, raise an error. Then set the user's variable #1 to the x-expansion of \l__regex_internal_a_tl, adding the appropriate braces to produce a balanced result. And end the group.

```

2525 \cs_new_protected_nopar:Npn \__regex_group_end_replace:N #1
2526 {
2527 \if_int_compare:w \l__regex_balance_int = \c_zero
2528 \else:
2529 \__msg_kernel_error:nnxxx { regex } { result-unbalanced }
2530 { replacing }
2531 { \int_max:nn { - \l__regex_balance_int } { \c_zero } }
2532 { \int_max:nn { \l__regex_balance_int } { \c_zero } }
2533 \fi:
2534 \use:x
2535 {
2536 \group_end:
2537 \tl_set:Nn \exp_not:N #1
2538 {
2539 \if_int_compare:w \l__regex_balance_int < \c_zero
2540 \prg_replicate:nn { - \l__regex_balance_int }
2541 { { \if_false: } \fi: }
2542 \fi:

```

```

2543         \l__regex_internal_a_tl
2544         \if_int_compare:w \l__regex_balance_int > \c_zero
2545             \prg_replicate:nn { \l__regex_balance_int }
2546             { \if_false: { \fi: } }
2547         \fi:
2548     }
2549 }
2550 }

```

(End definition for __regex_group_end_replace:N.)

2.7.5 Storing and showing compiled patterns

2.8 Messages

Messages for the preparsing phase.

```

2551 \__msg_kernel_new:nnnn { regex } { trailing-backslash }
2552 { Trailing~escape~character~\iow_char:N\\. }
2553 {
2554     A~regular~expression~or~its~replacement~text~ends~with~
2555     the~escape~character~\iow_char:N\\.~It~will~be~ignored.
2556 }
2557 \__msg_kernel_new:nnnn { regex } { x-missing-rbrace }
2558 { Missing~closing~brace~in~\iow_char:N\\x~hexadecimal~sequence. }
2559 {
2560     You~wrote~something~like~
2561     '\iow_char:N\\x\{\int_to_hexadecimal:n{#1}\}'~
2562     The~closing~brace~is~missing.
2563 }
2564 \__msg_kernel_new:nnnn { regex } { x-overflow }
2565 { Character~code~'#1'~too~large~in~\iow_char:N\\x~hexadecimal~sequence. }
2566 {
2567     You~wrote~something~like~
2568     '\iow_char:N\\x\{\int_to_hexadecimal:n{#1}\}'~
2569     The~character~code~'#1'~is~larger~than~\int_use:N \c_max_char_int.
2570 }

```

Invalid quantifier.

```

2571 \__msg_kernel_new:nnnn { regex } { invalid-quantifier }
2572 { Braced~quantifier~'#1'~may~not~be~followed~by~'#2'. }
2573 {
2574     The~character~'#2'~is~invalid~in~the~braced~quantifier~'#1'.~
2575     The~only~valid~quantifiers~are~'*',~'?',~'+',~'{<int>}',~
2576     '{<min>}',~and~'{<min>,<max>}',~followed~or~not~by~'?''.
2577 }

```

Messages for missing or extra closing brackets and parentheses, with some fancy singular/plural handling for the case of parentheses.

```

2578 \__msg_kernel_new:nnnn { regex } { missing-rbrack }
2579 { Missing~right~bracket~inserted~in~regular~expression. }
2580 {

```

```

2581 LaTeX~was~given~a~regular~expression~where~a~character~class~
2582 was~started~with~'[,~but~the~matching~'~is~missing.
2583 }
2584 \_msg_kernel_new:nnnn { regex } { missing-rparen }
2585 {
2586   Missing~right~parenthes\int_compare:nTF{#1=1}{i}{e}s~
2587   inserted~in~regular~expression.
2588 }
2589 {
2590   LaTeX~was~given~a~regular~expression~with~\int_eval:n{#1}~
2591   more~left~parenthes\int_compare:nTF{#1=1}{i}{e}s~than~right~
2592   parenthes\int_compare:nTF{#1=1}{i}{e}s.
2593 }
2594 \_msg_kernel_new:nnnn { regex } { extra-rparen }
2595 { Extra~right~parenthesis~ignored~in~regular~expression. }
2596 {
2597   LaTeX~came~across~a~closing~parenthesis~when~no~submatch~group~
2598   was~open.~The~parenthesis~will~be~ignored.
2599 }

```

Some escaped alphanumerics are not allowed everywhere.

```

2600 \_msg_kernel_new:nnnn { regex } { bad-escape }
2601 {
2602   Invalid~escape~\c_backslash_str #1~
2603   \_regex_if_in_cs:TF { within~a~control~sequence. }
2604   {
2605     \_regex_if_in_class:TF
2606     { in~a~character~class. }
2607     { following~a~category~test. }
2608   }
2609 }
2610 {
2611   The~escape~sequence~\iow_char:N\#1~may~not~appear~
2612   \_regex_if_in_cs:TF
2613   {
2614     within~a~control~sequence~test~introduced~by~
2615     \iow_char:N\c\iow_char:N{.
2616   }
2617   {
2618     \_regex_if_in_class:TF
2619     { within~a~character~class~ }
2620     { following~a~category~test~such~as~\iow_char:N\cL~ }
2621     because~it~does~not~match~exactly~one~character.
2622   }
2623 }

```

Range errors.

```

2624 \_msg_kernel_new:nnnn { regex } { range-missing-end }
2625 { Invalid~end~point~for~range~'#1-#2'~in~character~class. }
2626 {
2627   The~end~point~'#2'~of~the~range~'#1-#2'~may~not~serve~as~an~

```



```

2628     end-point~for~a~range:~alphanumeric~characters~should~not~be~
2629     escaped,~and~non~alphanumeric~characters~should~be~escaped.
2630 }
2631 \_msg_kernel_new:nnnn { regex } { range-backwards }
2632 { Range~[#1-#2]~out~of~order~in~character~class. }
2633 {
2634     In~ranges~of~characters~[x-y]~appearing~in~character~classes,~
2635     the~first~character~code~must~not~be~larger~than~the~second.~
2636     Here,~#1~has~character~code~\int_eval:n {'#1},~while~#2~has~
2637     character~code~\int_eval:n {'#2}.
2638 }

```

Errors related to \c and \u.

```

2639 \_msg_kernel_new:nnnn { regex } { c-bad-mode }
2640 { Invalid~nested~\iow_char:N\c~escape~in~regular~expression. }
2641 {
2642     The~\iow_char:N\c~escape~cannot~be~used~within~
2643     a~control~sequence~test~'\iow_char:N\c{...}'.~
2644     To~combine~several~category~tests,~use~'\iow_char:N\c[...}'.
2645 }
2646 \_msg_kernel_new:nnnn { regex } { c-missing-rbrace }
2647 { Missing~right~brace~inserted~for~\iow_char:N\c~escape. }
2648 {
2649     LaTeX~was~given~a~regular~expression~where~a~
2650     '\iow_char:N\c\iow_char:N{...}'~construction~was~not~ended~
2651     with~a~closing~brace~'\iow_char:N}'.
2652 }
2653 \_msg_kernel_new:nnnn { regex } { c-missing-rbrack }
2654 { Missing~right~bracket~inserted~for~\iow_char:N\c~escape. }
2655 {
2656     A~construction~'\iow_char:N\c[...]'~appears~in~a~
2657     regular~expression,~but~the~closing~']'~is~not~present.
2658 }
2659 \_msg_kernel_new:nnnn { regex } { c-missing-category }
2660 { Invalid~character~'#1'~following~\iow_char:N\c~escape. }
2661 {
2662     In~regular~expressions,~the~\iow_char:N\c~escape~sequence~
2663     may~only~be~followed~by~a~left~brace,~a~left~bracket,~or~a~
2664     capital~letter~representing~a~character~category,~namely~
2665     one~of~ABCDELMOPSTU.
2666 }
2667 \_msg_kernel_new:nnnn { regex } { u-missing-lbrace }
2668 { Missing~left~brace~following~\iow_char:N\u~escape. }
2669 {
2670     The~\iow_char:N\u~escape~sequence~must~be~followed~by~
2671     a~brace~group~with~the~name~of~the~variable~to~use.
2672 }
2673 \_msg_kernel_new:nnnn { regex } { u-missing-rbrace }
2674 { Missing~right~brace~inserted~for~\iow_char:N\u~escape. }
2675 {

```

```

2676 LaTeX~
2677 \str_if_eq_x:nnTF { } {#2}
2678 { reached-the-end-of-the-string~ }
2679 { encountered-an-escaped-alphanumeric-character '\iow_char:N\{#2'~ }
2680 when-parsing-the-argument-of-an'\iow_char:N\{...\}'~escape.
2681 }

```

Errors when encountering the POSIX syntax [:...:].

```

2682 \__msg_kernel_new:nnnn { regex } { posix-unsupported }
2683 { POSIX-collating-element~'[#1 ~ #1]~not-supported. }
2684 {
2685   The~[.foo.]~and~[=bar=]~syntaxes~have~a~special~meaning~in~POSIX~
2686   regular~expressions.~This~is~not~supported~by~LaTeX.~Maybe~you~
2687   forgot~to~escape~a~left~bracket~in~a~character~class?
2688 }
2689 \__msg_kernel_new:nnnn { regex } { posix-unknown }
2690 { POSIX-class~[:#1:]~unknown. }
2691 {
2692   [:#1:]~is~not~among~the~known~POSIX~classes~
2693   [:alnum:],~[:alpha:],~[:ascii:],~[:blank:],~
2694   [:cntrl:],~[:digit:],~[:graph:],~[:lower:],~
2695   [:print:],~[:punct:],~[:space:],~[:upper:],~
2696   [:word:],~and~[:xdigit:].
2697 }
2698 \__msg_kernel_new:nnnn { regex } { posix-missing-close }
2699 { Missing-closing~':'~for~POSIX-class. }
2700 { The~POSIX-syntax~'#1'~must~be~followed~by~':'',~not~'#2'. }

```

In various cases, the result of a `l3regex` operation can leave us with an unbalanced token list, which we must re-balance by adding begin-group or end-group character tokens.

```

2701 \__msg_kernel_new:nnnn { regex } { result-unbalanced }
2702 { Missing-brace-inserted-when~#1. }
2703 {
2704   LaTeX~was~asked~to~do~some~regular~expression~operation,~
2705   and~the~resulting~token~list~would~not~have~the~same~number~
2706   of~begin-group~and~end-group~tokens.~Braces~were~inserted:~
2707   #2~left,~#3~right.
2708 }

```

Error message for unknown options.

```

2709 \__msg_kernel_new:nnnn { regex } { unknown-option }
2710 { Unknown-option~'#1'~for~regular~expressions. }
2711 {
2712   The~only~available~option~is~'case-insensitive',~toggled~by~
2713   '(?i)'~and~'(?-i)'.
2714 }

```

Errors in the replacement text.

```

2715 \__msg_kernel_new:nnnn { regex } { replacement-c }
2716 { Misused~\iow_char:N\c~command~in~a~replacement~text. }

```

```

2717 {
2718   In~a~replacement~text,~the~\iow_char:N\\c~escape~sequence~
2719   can~be~followed~by~one~of~the~letters~ABCDELMOPTU~
2720   or~a~brace~group,~not~by~'#1'.
2721 }
2722 \__msg_kernel_new:nnnn { regex } { replacement-u }
2723 { Misused~\iow_char:N\\u~command~in~a~replacement~text. }
2724 {
2725   In~a~replacement~text,~the~\iow_char:N\\u~escape~sequence~
2726   must~be~followed~by~a~brace~group~holding~the~name~of~the~
2727   variable~to~use.
2728 }
2729 \__msg_kernel_new:nnnn { regex } { replacement-g }
2730 { Missing~brace~for~the~\iow_char:N\\g~construction~in~a~replacement~text. }
2731 {
2732   In~the~replacement~text~for~a~regular~expression~search,~
2733   submatches~are~represented~either~as~\iow_char:N \\g{dd..d},~
2734   or~\\d,~where~'d'~are~single~digits.~Here,~a~brace~is~missing.
2735 }
2736 \__msg_kernel_new:nnnn { regex } { replacement-catcode-end }
2737 {
2738   Missing~character~for~the~\iow_char:N\\c<category><character>~
2739   construction~in~a~replacement~text.
2740 }
2741 {
2742   In~a~replacement~text,~the~\iow_char:N\\c~escape~sequence~
2743   can~be~followed~by~one~of~the~letters~ABCDELMOPTU~representing~
2744   the~character~category.~Then,~a~character~must~follow.~LaTeX~
2745   reached~the~end~of~the~replacement~when~looking~for~that.
2746 }
2747 \__msg_kernel_new:nnnn { regex } { replacement-null-space }
2748 { TeX~cannot~build~a~space~token~with~character~code~0. }
2749 {
2750   You~asked~for~a~character~token~with~category~'space',~
2751   and~character~code~0,~for~instance~through~
2752   '\iow_char:N\\cS\iow_char:N\\x00'.~
2753   This~specific~case~is~impossible~and~will~be~replaced~
2754   by~a~normal~space.
2755 }
2756 \__msg_kernel_new:nnnn { regex } { replacement-missing-rbrace }
2757 { Missing~right~brace~inserted~in~replacement~text. }
2758 {
2759   There~were~\int_use:N \l__regex_replacement_csnames_int \
2760   missing~right~braces.
2761 }

```

__regex_msg_repeated:nnN

This is not technically a message, but seems related enough to go there. The arguments are: #1 is the minimum number of repetitions; #2 is the number of allowed extra repetitions (−1 for infinite number), and #3 tells us about laziness.

```

2762 \cs_new:Npn \__regex_msg_repeated:nnN #1#2#3
2763 {
2764   \str_if_eq_x:nnF { #1 #2 } { 1 0 }
2765   {
2766     , ~ repeated ~
2767     \int_case:nnF {#2}
2768     {
2769       { -1 } { #1~or~more~times,~\bool_if:NTF #3 { lazy } { greedy } }
2770       { 0 } { #1~times }
2771     }
2772     {
2773       between~#1~and~\int_eval:n {#1+#2}~times,~
2774       \bool_if:NTF #3 { lazy } { greedy }
2775     }
2776   }
2777 }
(End definition for \__regex_msg_repeated:nnN.)

```

2.9 Code for tracing

The tracing code is still very experimental, and is meant to be used with the `l3trace` package, currently in `l3trial`.

`__regex_trace_states:n` This function lists the contents of all states of the NFA, stored in `\toks` from 0 to `\l__regex_max_state_int` (excluded).

```

2778 \trace
2779 \cs_new_protected:Npn \__regex_trace_states:n #1
2780 {
2781   \int_step_inline:nnnn
2782     \l__regex_min_state_int
2783     \c_one
2784     { \l__regex_max_state_int - 1 }
2785   {
2786     \trace:nx { regex } { #1 }
2787     { \iow_char:N \toks ##1 = { \tex_the:D \tex_toks:D ##1 } }
2788   }
2789 }
2790 \trace
(End definition for \__regex_trace_states:n.)
2791 \initex |package>

```

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.

Symbols

`\$` 843
100

\\	1271, 1296, 1311, 1380, 1381, 1382, 2552, 2555, 2558, 2561, 2565, 2568, 2611, 2615, 2620, 2640, 2642, 2643, 2644, 2647, 2650, 2654, 2656, 2660, 2662, 2668, 2670, 2674, 2679, 2680, 2716, 2718, 2723, 2725, 2730, 2733, 2734, 2738, 2742, 2752, 2787	_regex_action_start_wildcard:	1420, 1935, 1935
{	2561, 2568, 2615, 2650, 2680	_regex_action_submatch:n	1641, 1642, 1771, 1982, 1984, 1984
}	2568, 2651, 2680	_regex_action_success:	1423, 1439, 1989, 1989
^	267, 269, 271, 273, 275, 277, 842, 2191, 2194, 2205, 2208, 2217, 2220, 2223, 2226, 2244, 2247	_regex_anchor:N	825, 1275, 1733, 1751
_int_eval:w	105, 1140, 1651, 1889, 1923, 1966, 2016, 2017, 2028, 2033, 2034, 2101, 2454	_regex_assertion:Nn	825, 850, 859, 1272, 1733, 1733
_int_eval_end:	105, 1140, 1923, 2101	_regex_b_test:	850, 859, 1274, 1733, 1757
_int_value:w	281, 736, 742, 774, 776, 785, 786, 909, 1226, 1240, 1886, 2423, 2450, 2452	_regex_begin	2314
_msg_kernel_error:nn	262, 360, 501, 530, 573, 1180, 2185, 2235, 2238	_regex_branch:n	20, 524, 996, 1039, 1256, 1611, 1611
_msg_kernel_error:nnx	300, 536, 725, 1097, 1133, 1198, 2059, 2253	_regex_break_point:TF	24, 25, 26, 30, 1514, 1515, 1739, 1762
_msg_kernel_error:nnxx	702, 767, 983	_regex_break_true:w	24, 24, 30, 35, 42, 49, 55, 62, 70, 117, 131, 151, 800, 1754
_msg_kernel_error:nnxxx	2407, 2529	_regex_build:N	1406, 1409, 1411, 2288, 2294, 2298, 2302
_msg_kernel_new:nnnn	2551, 2557, 2564, 2571, 2578, 2584, 2594, 2600, 2624, 2631, 2639, 2646, 2653, 2659, 2667, 2673, 2682, 2689, 2698, 2701, 2709, 2715, 2722, 2729, 2736, 2747, 2756	_regex_build:n	1406, 1406, 2283, 2292, 2297, 2300
_msg_kernel_warning:nn	1007	_regex_build_for_cs:n	142, 1427, 1427
_msg_kernel_warning:nnx	933, 934, 973, 1023, 1059, 1075	_regex_build_new_state:	1417, 1418, 1431, 1432, 1477, 1477, 1496, 1528, 1563, 1568, 1614, 1630, 1635, 1677, 1696, 1731, 1735, 1768
_msg_kernel_warning:nnxx	635, 781	_regex_build_transition_left:NNN	1473, 1473, 1679, 1693, 1710
_msg_show_variable:n	1304	_regex_build_transition_right:nNn	1473, 1475, 1529, 1565, 1617, 1621, 1647, 1675, 1682, 1690, 1718, 1728
_prg_break:	258, 1999, 2024	_regex_build_transitions_lazyness:NNNNN	1494, 1494, 1536, 1542, 1554
_prg_break:n	1598	_regex_char_if_alphanumeric:N	389
_prg_break_point:	235, 1599, 1874, 2018	_regex_char_if_alphanumeric:NTF	367, 564
_regex_action_cost:n	1514, 1515, 1523, 1940, 1963, 1963	_regex_char_if_special:N	367
_regex_action_free:n	1537, 1543, 1544, 1555, 1617, 1621, 1647, 1675, 1679, 1682, 1710, 1718, 1728, 1742, 1773, 1938, 1942, 1942	_regex_char_if_special:NTF	367, 560
_regex_action_free_aux:nn	1942, 1943, 1945, 1946	_regex_chk_c_allowed:T	493, 493, 1083
_regex_action_free_group:n	1565, 1690, 1693, 1942, 1944	_regex_class:NnnnN	21, 598, 893, 894, 900, 1225, 1235, 1269, 1508, 1508
		_regex_class_repeat:n	1518, 1524, 1524, 1540, 1549
		_regex_class_repeat:nN	1519, 1533, 1533
		_regex_class_repeat:nnN	1520, 1547, 1547
		_regex_command_K:	1250, 1270, 1766, 1766
		_regex_compile:n	553, 553, 1408, 2260, 2265, 2270, 2275
		_regex_compile:w	516, 516, 555, 1154
		_regex_compile_\$:	820

_regex_compile(:	1011	_regex_compile_class_posix:NNNNw	.
_regex_compile_):	1042		926, 932, 939
_regex_compile_::	791	_regex_compile_class_posix_end:w	.
_regex_compile_/A:	820		926, 957, 959
_regex_compile_/B:	844	_regex_compile_class_posix_loop:w	
_regex_compile_/D:	803		926, 945, 950, 953, 956
_regex_compile_/G:	820	_regex_compile_class_posix_test:w	
_regex_compile_/H:	803		880, 926, 926
_regex_compile_/K:	1247	_regex_compile_end:	
_regex_compile_/N:	803		516, 526, 574, 580, 1163
_regex_compile_/S:	803	_regex_compile_escaped:N	565, 582, 587
_regex_compile_/V:	803	_regex_compile_group_begin:N
_regex_compile_/W:	803		989, 989, 1025, 1030, 1048, 1050
_regex_compile_/Z:	820	_regex_compile_group_end:
_regex_compile_/b:	844		989, 998, 1045
_regex_compile_/c:	1082	_regex_compile_lparen:w	... 1014, 1016
_regex_compile_/d:	803	_regex_compile_one:x	575, 592, 592,
_regex_compile_/h:	803		735, 741, 795, 806, 809, 819, 966, 1164
_regex_compile_/s:	803	_regex_compile_quantifier:w
_regex_compile_/u:	1169		610, 621, 621, 872, 1005
_regex_compile_/v:	803	_regex_compile_quantifier*:w	... 648
_regex_compile_/w:	803	_regex_compile_quantifier+:w	... 648
_regex_compile_/z:	820	_regex_compile_quantifier?:w	... 648
_regex_compile_[:	877	_regex_compile_quantifier_abort:xNN	
_regex_compile_]:	862		630, 632, 658, 677, 691, 714
_regex_compile_^:	820	_regex_compile_quantifier_braced_auxi:w	
_regex_compile_abort_tokens:n		654, 657, 660
	612, 612, 620	_regex_compile_quantifier_braced_auxii:w	
_regex_compile_abort_tokens:x		654, 673, 682
	612, 636, 975, 985	_regex_compile_quantifier_braced_auxiii:w	
_regex_compile_anchor:Nf		654, 672, 696
	820, 820, 831, 840	_regex_compile_quantifier_lazyness:nnNN	
_regex_compile_c_[w	1101		639, 639, 649, 651, 653, 666, 687, 709
_regex_compile_c_lbrack_add:N	_regex_compile_quantifier_none:	..
	1101, 1123, 1138		626, 628, 630, 630, 634
_regex_compile_c_lbrack_end:	_regex_compile_range:Nw	733, 746, 762
	1101, 1130, 1134, 1145	_regex_compile_raw:N 443, 561,
_regex_compile_c_lbrack_loop:NN	..		565, 567, 585, 590, 617, 726, 728,
	1101, 1109, 1113, 1117, 1125		728, 748, 794, 840, 875, 923, 937,
_regex_compile_c_test:NN		955, 1008, 1013, 1026, 1036, 1044,
	1082, 1083, 1084		1060, 1061, 1062, 1069, 1076, 1077,
_regex_compile_class:NN		1078, 1086, 1119, 1167, 1181, 1187
	907, 913, 917, 920	_regex_compile_raw_error:N
_regex_compile_class:TFNN		723, 723, 831, 847, 856, 1172, 1251
	892, 903, 907, 907	_regex_compile_special:N
_regex_compile_class_catcode:w		561, 582, 582, 623, 641, 664, 669,
	884, 896, 896		685, 699, 732, 751, 910, 928, 941,
_regex_compile_class_normal:w		962, 1018, 1053, 1069, 1106, 1174, 1190
	887, 890, 890	_regex_compile_special_group-:w	1051
		_regex_compile_special_group::w	1047

_regex_compile_special_group_i:w .	_regex_extract_b:wn ..	2438, 2452, 2461
..... 1051, 1051	_regex_extract_e:wn ..	2438, 2450, 2466
_regex_compile_u_end:	_regex_extract_once:nnN	2304, 2347, 2347
..... 1193, 1199, 1204, 1204	_regex_extract_seq_aux:n	
_regex_compile_u_in_cs: 1210, 1213, 1213 2401, 2418, 2418	
_regex_compile_u_in_cs_aux:n	_regex_extract_seq_aux:ww	
..... 1220, 1223 2418, 2422, 2426	
_regex_compile_u_loop:NN	_regex_get_digits:NTFW	
..... 1169, 1177, 1185, 1188, 1194 429, 429, 656, 671	
_regex_compile_u_not_cs:	_regex_get_digits_loop:nw	432, 435, 438
..... 1208, 1229, 1229	_regex_get_digits_loop:w	429
_regex_compute_case_changed_char:	_regex_group:nnnN	
..... 52, 66, 74, 74	.. 1025, 1030, 1263, 1421, 1579, 1579	
_regex_count:nnN 2292, 2294, 2335, 2335	_regex_group_aux:nnnnN	
_regex_disable_submatches: 1559, 1559, 1581, 1589, 1592	
... 141, 1155, 1979, 1979, 2329, 2338	_regex_group_end_extract_seq:N ...	
_regex_end 2354, 2362, 2387, 2389, 2389	
_regex_escape_\w:w	_regex_group_end_replace:N	
..... 258 2494, 2523, 2525, 2525	
_regex_escape_/a:w	_regex_group_no_capture:nnnN	
..... 258 1048, 1265, 1579, 1588	
_regex_escape_/break:w	_regex_group_repeat:nn	1573, 1625, 1625
..... 258	_regex_group_repeat:nnN	1574, 1668, 1668
_regex_escape_/e:w	_regex_group_repeat:nnnN	
..... 258 1575, 1699, 1699	
_regex_escape_/f:w	_regex_group_repeat_aux:n	
..... 258 1632, 1645, 1645, 1686, 1703	
_regex_escape_/n:w	_regex_group_resetting:nnnN	
..... 258 1050, 1267, 1590, 1590	
_regex_escape_/r:w	_regex_group_resetting_loop:nnNn .	
..... 258 1590, 1594, 1602, 1607	
_regex_escape_/t:w	_regex_group_submatches:nnN	
..... 258	.. 1633, 1638, 1638, 1671, 1687, 1701	
_regex_escape_/x:w	_regex_if_end_range:NN	746
..... 278	_regex_if_end_range:NNTF	746, 764
_regex_escape_\w:w	_regex_if_in_class:TF	
..... 242	.. 453, 453, 528, 595, 610, 730, 793,	
_regex_escape_break:w	864, 879, 1013, 1036, 1044, 2605, 2618	
..... 258, 258	_regex_if_in_class_or_catcode:TF .	
_regex_escape_escaped:N 473, 473, 822, 846, 855, 1171	
..... 228, 252, 255, 256	_regex_if_in_cs:TF	
_regex_escape_loop:N 461, 461, 1161, 2603, 2612	
..... 234, 242, 242, 246,	_regex_if_match:nn	2283, 2288, 2326, 2326
249, 253, 278, 334, 342, 343, 356, 362	_regex_if_raw_digit:NN	441
_regex_escape_raw:N . 229, 255, 257,	_regex_if_raw_digit:NNTF	431, 437, 441
267, 269, 271, 273, 275, 277, 287, 310	_regex_if_two_empty_matches:F	
_regex_escape_unescaped:N 1794, 1796, 1829, 1832, 1991	
..... 227, 245, 255, 255	_regex_if_within_catcode:TF	
_regex_escape_use:nnnn 485, 485, 882	
..... 223, 223, 558, 2049		
_regex_escape_x:N		
..... 333, 337, 337		
_regex_escape_x_end:w ..		
..... 278, 280, 283		
_regex_escape_x_large:n .		
..... 278, 290, 294		
_regex_escape_x_loop:N		
..... 330, 346, 346, 349, 352		
_regex_escape_x_test:N		
..... 281, 315, 315, 320		
_regex_escape_x_test_two:N ..		
..... 322, 327		
_regex_extract:		
..... 2353,		
2359, 2370, 2438, 2438, 2481, 2500		
_regex_extract_all:nnN		
2306, 2347, 2356		

_regex_item_caseful_equal:n	_regex_posix_space:	183, 209
.....	32, 32, 161,	_regex_posix_upper:	183, 186, 214
.....	162, 166, 167, 168, 169, 170, 179,	_regex_posix_word:	183, 216
.....	182, 193, 211, 522, 1071, 1226, 1276	_regex_posix_xdigit:	183, 217
_regex_item_caseful_range:nn	. 32,	_regex_prop_:	791
.....	38, 158, 173, 176, 177, 178, 188,	_regex_prop_N:	157, 181, 819
.....	192, 197, 199, 201, 204, 205, 206,	_regex_prop_d:	157, 157, 195
.....	207, 212, 215, 220, 221, 523, 1072, 1278	_regex_prop_h:	157, 159, 189
_regex_item_caseless_equal:n	_regex_prop_s:	157, 164
.....	46, 46, 1055, 1280	_regex_prop_v:	157, 172
_regex_item_caseless_range:nn	_regex_prop_w:	157, 174, 216, 1761, 1763, 1764
.....	46, 58, 1056, 1282	_regex_push_lr_states:	1433, 1445, 1445, 1569
_regex_item_catcode:	_regex_query_get:	1841, 1867, 1907, 1907
_regex_item_catcode:nT	_regex_query_range:nn	2008,
.....	93, 103, 112, 603, 904, 1287	2013, 2013, 2032, 2078, 2490, 2520	
_regex_item_catcode_reverse:nT	...	_regex_query_range_loop:ww	2013, 2015, 2020, 2027
.....	93, 111, 905, 1289	_regex_query_set:nnn	1806, 1809, 1811, 1892, 1892
_regex_item_cs:n	135, 135, 576, 1165, 1297	_regex_query_submatch:n	2030, 2030, 2098, 2432
_regex_item_equal:n	_regex_replace_all:nnN	2310, 2497, 2497	
.....	91, 91,	_regex_replace_once:nnN	2308, 2471, 2471	
.....	522, 736, 742, 772, 785, 786, 1055, 1071	_regex_replacement:n	2043, 2043, 2476, 2502
_regex_item_exact:nn	113, 113, 1240, 1293	_regex_replacement_aux:n	2043, 2071, 2074
_regex_item_exact_cs:c	_regex_replacement_balance_one_match:n	2004, 2004, 2064, 2484, 2511
.....	113, 121, 1238, 1295	_regex_replacement_c:w	2140, 2140
_regex_item_range:nn	_regex_replacement_c_A:w	..	2191, 2192
.....	91, 92, 523, 774, 1056, 1072	_regex_replacement_c_B:w	..	2194, 2195
_regex_item_reverse:n	_regex_replacement_c_C:w	..	2203, 2203
.....	27, 27, 112, 182, 810, 968, 1291, 1763	_regex_replacement_c_D:w	..	2205, 2206
_regex_match:n	144, 1800, 1800, 2332,	_regex_replacement_c_E:w	..	2208, 2209
.....	2342, 2352, 2361, 2377, 2477, 2503	_regex_replacement_c_L:w	..	2217, 2218
_regex_match_loop:	1842, 1861, 1861, 1878	_regex_replacement_c_M:w	..	2220, 2221
_regex_match_once:	1823, 1826, 1826, 1857	_regex_replacement_c_O:w	..	2223, 2224
_regex_match_one_active:w	_regex_replacement_c_P:w	..	2226, 2227
.....	1861, 1871, 1882, 1888	_regex_replacement_c_S:w	..	2232, 2232
_regex_mode_quit:c:	.. 506, 506, 594, 992	_regex_replacement_c_T:w	..	2244, 2245
_regex_msg_repeated:nnN	_regex_replacement_c_U:w	..	2247, 2248
..	1334, 1356, 1360, 1370, 2762, 2762	_regex_replacement_char:nnN	2182, 2182, 2193, 2200, 2207, 2214,
_regex_multi_match:n	2219, 2222, 2225, 2229, 2246, 2249	
..	1845, 1850, 2340, 2359, 2367, 2500	_regex_replacement_do_one_match:n	2006, 2006, 2076, 2489, 2519
_regex_pop_lr_states:			
.....	1435, 1445, 1452, 1571			
_regex_posix_alnum:			
.....	183, 183			
_regex_posix_alpha:			
.....	183, 184, 185			
_regex_posix_ascii:			
.....	183, 187			
_regex_posix_blank:			
.....	183, 189			
_regex_posix_cntrl:			
.....	183, 190			
_regex_posix_digit:	. 183, 184, 195, 219			
_regex_posix_graph:			
.....	183, 196			
_regex_posix_lower:			
.....	183, 186, 198			
_regex_posix_print:			
.....	183, 200			
_regex_posix_punct:			
.....	183, 202			

\bool_set_false:N	943, 1108, 1775, 1822, 1835, 1875, 1939
\bool_set_true:N	948, 1112, 1772, 1937, 1993	
C		
\c_regex_all_catcodes_int	413, 425, 520, 602, 1149
\c_regex_catcode_A_int	413, 424
\c_regex_catcode_B_int	413, 414
\c_regex_catcode_C_int	413, 413
\c_regex_catcode_D_int	413, 420
\c_regex_catcode_E_int	413, 415
\c_regex_catcode_L_int	413, 422
\c_regex_catcode_M_int	413, 416
\c_regex_catcode_O_int	413, 423
\c_regex_catcode_P_int	413, 418
\c_regex_catcode_S_int	413, 421
\c_regex_catcode_T_int	413, 417
\c_regex_catcode_U_int	413, 419
\c_regex_no_match_regex	18, 18, 426, 2257	
\c_backslash_str	248, 782, 2602
\c_false_bool	631, 644, 859, 894, 1147, 1227, 1243, 1290, 1421, 2478
\c_fifteen	869
\c_fifty_eight	370, 392
\c_forty_eight	158, 178, 371, 393
\c_left_brace_str	329, 654, 658, 678, 692, 716, 1152, 1174, 2107, 2149, 2161	
\c_max_char_int	298, 2569
\c_max_int	51, 65, 1866
\c_minus_one	1640, 2449
\c_nine	162, 167, 212
\c_ninety_one	77, 369, 391
\c_ninety_seven	84, 176, 199, 221, 381, 401	
\c_one	444, 1486, 1551, 1629, 1651, 1654, 1663, 1664, 1674, 1714, 1723, 1815, 1816, 1840, 1889, 1923, 2028, 2088, 2118, 2399, 2400, 2510, 2518, 2783	
\c_one_hundred_twenty_seven	188, 193, 384	
\c_one_hundred_twenty_three	83, 380, 400	
\c_right_brace_str	355, 664, 685, 699, 1159, 1167, 1192, 2051
\c_six	511, 1092, 1104, 1157
\c_sixteen	866
\c_sixty_five	78, 177, 215, 220, 376, 396	
\c_space_token	319, 351
\c_ten	168, 173, 182, 2120
\c_thirteen	170, 173, 212, 870
\c_thirty_two	80, 86, 161, 166, 201, 211, 372	
\c_three	498, 512
\c_true_bool	21, 150, 544, 598, 642, 825, 850, 893, 900, 1225, 1235, 1288, 1499, 1678, 1689, 1704, 1828, 1854, 2320, 2382, 2440
\c_twelve	169
\c_two	508, 799, 898, 1092, 1104, 1157, 1805, 1821, 1863, 2398, 2508, 2516	
\c_two_hundred_fifty_six	285
\c_zero	100, 123, 137, 188, 192, 305, 466, 478, 487, 495, 509, 535, 557, 571, 1000, 1207, 1237, 1249, 1535, 1562, 1627, 1670, 1721, 1818, 2058, 2099, 2174, 2187, 2197, 2211, 2237, 2405, 2428, 2433, 2444, 2527, 2531, 2532, 2539, 2544
\char_set_catcode_active:N	2191
\char_set_catcode_alignment:N	...	2244
\char_set_catcode_group_begin:N	..	2194
\char_set_catcode_group_end:N	...	2208
\char_set_catcode_letter:N	2217
\char_set_catcode_math_subscript:N	2205
\char_set_catcode_math_superscript:N	2247
\char_set_catcode_math_toggle:N	..	2220
\char_set_catcode_other:N	2223
\char_set_catcode_other:n	293
\char_set_catcode_parameter:N	...	2226
\char_set_lccode:nn	305
\cs:w	2153, 2155, 2164, 2166
\cs_end:	2062, 2175
\cs_generate_variant:Nn	8, 620
\cs_gset:Npx	2064
\cs_if_exist:cTF	964
\cs_if_exist_use:cF	244, 251, 584, 589, 625, 1020, 1095, 2086, 2144
\cs_new:cpn	278
\cs_new:Npn	9, 242, 283, 294, 315, 327, 337, 346, 435, 953, 1185, 1223, 1375, 1523, 1882, 2004, 2006, 2012, 2013, 2020, 2030, 2418, 2426, 2762
\cs_new_eq:NN	91, 92, 189, 195, 216, 255, 256, 257, 258, 426, 1796, 2257
\cs_new_nopar:cpn	248, 259, 265
\cs_new_nopar:cpx	266, 268, 270, 272, 274, 276
\cs_new_nopar:Npn	453, 461, 473, 485
\cs_new_protected:cpn	654, 862, 1082, 1101, 1159, 1169
\cs_new_protected:Npn	24, 26, 27, 32, 38, 46, 58, 93,

103, 111, 113, 121, 135, 223, 429, 493, 506, 553, 582, 587, 592, 612, 621, 630, 632, 639, 660, 682, 696, 723, 728, 762, 820, 896, 907, 920, 926, 939, 959, 989, 998, 1016, 1051, 1084, 1117, 1138, 1204, 1213, 1229, 1253, 1306, 1316, 1318, 1320, 1327, 1386, 1406, 1411, 1427, 1459, 1465, 1471, 1473, 1475, 1494, 1508, 1524, 1533, 1547, 1559, 1579, 1590, 1602, 1611, 1625, 1638, 1645, 1668, 1699, 1733, 1751, 1800, 1850, 1892, 1925, 1946, 1963, 1968, 1984, 2036, 2043, 2074, 2084, 2095, 2105, 2114, 2140, 2159, 2172, 2182, 2203, 2232, 2251, 2256, 2273, 2279, 2291, 2293, 2297, 2298, 2326, 2335, 2347, 2356, 2364, 2389, 2461, 2466, 2471, 2497, 2779	375, 377, 379, 382, 383, 385, 394, 395, 397, 399, 402, 403, 446, 449, 457, 465, 468, 477, 480, 489, 497, 500, 510, 605, 706, 750, 754, 757, 768, 773, 871, 1006, 1092, 1104, 1141, 1157, 1209, 1239, 1502, 1520, 1539, 1575, 1631, 1681, 1685, 1692, 1713, 1724, 1831, 1943, 2023, 2052, 2090, 2123, 2133, 2154, 2165, 2177, 2186, 2236, 2322, 2451, 2480, 2528
\cs_new_protected_nopar:cpn	\etex_glueshrink:D 2009, 2038, 2079
. 648, 650, 652, 797, 818, 830, 839, 844, 853, 877, 1011, 1034, 1042, 1047, 1049, 1066, 1152, 1247, 2149	\etex_gluestretch:D 1911, 2034, 2039, 2464, 2491
\cs_new_protected_nopar:cpx 791, 805, 807	\etex_gluetomu:D 1895
\cs_new_protected_nopar:Npn	\etex_muexpr:D 2038
. 74, 157, 159, 164, 172, 174, 181, 183, 185, 187, 190, 196, 198, 200, 202, 209, 214, 217, 516, 526, 890, 1145, 1445, 1452, 1477, 1588, 1757, 1766, 1826, 1845, 1861, 1907, 1914, 1935, 1942, 1944, 1974, 1979, 1989, 2192, 2195, 2206, 2209, 2218, 2221, 2224, 2227, 2245, 2248, 2258, 2263, 2268, 2318, 2438, 2525	\etex_mutogluue:D 1910, 1912, 2038
\cs_set:cpn 1299	\exp_after:wN 35, 42, 49, 55, 62, 70, 106, 108, 117, 127, 146, 147, 234, 263, 280, 320, 322, 323, 330, 333, 334, 355, 456, 458, 464, 467, 469, 476, 479, 481, 488, 490, 496, 499, 502, 800, 872, 884, 1005, 1008, 1167, 1192, 1238, 1439, 1462, 1463, 1468, 1469, 1472, 1754, 1857, 1871, 1878, 1888, 1889, 1931, 1954, 1976, 2015, 2016, 2024, 2027, 2121, 2122, 2124, 2125, 2132, 2134, 2135, 2153, 2155, 2164, 2166, 2201, 2422, 2450, 2452
\cs_set:Npn 1336, 2076	\exp_args:Nc 1123
\cs_set_eq:NN 1269, 1275, 1832, 2391	\exp_args:Nf 1607
\cs_set_nopar:Npn . 227, 228, 229, 522, 523, 1055, 1056, 1071, 1072, 1274, 1829	\exp_args:NNNo 2343
\cs_set_nopar:Npx 1461, 1467, 1510	\exp_args:NNo 1215
\cs_set_protected:Npn	\exp_args:No 616, 666, 687, 1231, 1581, 1965, 2071, 2130, 2477, 2503
. 803, 828, 837, 1272, 1276, 1278, 1280, 1282, 1287, 1289, 1291, 1293, 1295, 1297, 1982, 2295	\exp_args:Noo 709, 1594
\cs_set_protected_nopar:Npn	\exp_args:Nx 144, 842, 843
. . 1256, 1263, 1265, 1267, 1270, 1981	\exp_last_unbraced:Nf 287, 1342
\cs_to_str:N 127, 147, 1238	\exp_not:c 795, 806, 810, 969, 2204
E	\exp_not:N 267, 269, 271, 273, 275, 277, 604, 606, 793, 1311, 1959, 2193, 2204, 2415, 2537
\else: 79, 82, 85, 100, 107, 321, 331, 373, 374,	\exp_not:n 604, 606, 1512, 1738, 1951, 2012, 2153, 2155, 2164, 2166, 2193, 2201, 2215, 2230, 2430, 2435
	\exp_not:o 576, 1165, 1960, 2271
	\exp_not:V 2164
	\exp_stop_f: . 444, 766, 771, 909, 1517, 1572, 1902, 2022, 2026, 2088, 2118
	\ExplFileDate 5
	\ExplFileDescription 5

\ExplFileName	5	\if_false:	261, 263,
\ExplFileVersion	5		296, 302, 309, 358, 362, 524, 543,
			544, 550, 598, 631, 642, 644, 867,
			900, 912, 916, 944, 949, 957, 991,
			996, 1001, 1039, 1176, 1193, 1197,
			2201, 2215, 2430, 2435, 2541, 2546
F			
\fi:	36,	\if_int_compare:w	34,
	43, 44, 50, 53, 56, 63, 64, 67, 71, 72,		40, 41, 48, 51, 54, 60, 61, 65, 68,
	81, 87, 88, 89, 101, 109, 118, 119,		69, 77, 78, 83, 84, 115, 116, 369,
	152, 261, 263, 296, 302, 309, 324,		370, 371, 372, 376, 380, 381, 384,
	335, 358, 362, 373, 374, 377, 378,		391, 392, 393, 396, 400, 401, 444,
	382, 385, 386, 387, 394, 397, 398,		466, 478, 487, 495, 498, 508, 511,
	402, 405, 406, 448, 451, 459, 470,		535, 602, 701, 766, 771, 799, 866,
	471, 482, 483, 491, 503, 504, 513,		898, 1000, 1207, 1237, 1438, 1535,
	514, 524, 543, 544, 549, 550, 598,		1562, 1627, 1640, 1651, 1670, 1721,
	607, 631, 642, 644, 708, 756, 759,		1753, 1876, 1877, 1884, 1928, 1953,
	760, 775, 778, 801, 867, 868, 873,		2022, 2058, 2088, 2097, 2099, 2118,
	900, 901, 912, 916, 944, 949, 957,		2174, 2197, 2211, 2237, 2369, 2383,
	991, 996, 1001, 1009, 1039, 1092,		2428, 2433, 2449, 2527, 2539, 2544
	1104, 1143, 1150, 1157, 1176, 1193,	\if_int_odd:w	105, 455, 463, 475, 871, 1140
	1197, 1211, 1241, 1440, 1505, 1521,	\if_meaning:w	
	1545, 1567, 1576, 1636, 1643, 1666,		.. 150, 443, 748, 751, 1147, 1499,
	1684, 1695, 1697, 1727, 1730, 1755,		1678, 1689, 1704, 1828, 1854, 2129,
	1833, 1858, 1879, 1880, 1890, 1905,		2184, 2234, 2320, 2382, 2440, 2478
	1932, 1955, 2025, 2052, 2063, 2092,	\int_add:Nn	80, 1950
	2102, 2103, 2126, 2136, 2156, 2167,	\int_case:nnF	1343, 2767
	2179, 2189, 2199, 2201, 2213, 2215,	\int_compare:nNnF	1830, 2403
	2239, 2242, 2324, 2374, 2385, 2386,	\int_compare:nNnT	137, 571
	2430, 2431, 2435, 2436, 2453, 2459,	\int_compare:nNnTF ...	123, 285, 298, 1249
	2495, 2533, 2541, 2542, 2546, 2547	\int_compare:nTF	2586, 2591, 2592
\flag_clear:n	2392, 2393	\int_const:Nn .	413, 414, 415, 416, 417,
\flag_height:n	2404, 2409, 2410		418, 419, 420, 421, 422, 423, 424, 425
\flag_new:n	2314, 2315	\int_decr:N	1904, 2176, 2212, 2384
\flag_raise:n	2429, 2434	\int_eval:n	1277, 1279, 1281, 1285, 1474,
			1476, 1486, 1500, 1501, 1503, 1504,
			1744, 1774, 2590, 2636, 2637, 2773
G			
\g_regex_internal_tl		\int_if_exist:cTF	1088, 1121
 10, 17, 231, 234, 1215, 1219	\int_if_odd_p:n	1390
\g_regex_success_bool		\int_incr:N	995, 1308,
 143, 150, 153, 1797, 1797,		1492, 1584, 1864, 1901, 1903, 1972,
	1813, 1848, 1855, 2320, 2440, 2478		2157, 2168, 2198, 2340, 2381, 2445
\group_begin: 139, 292, 1759, 2181, 2328,		\int_max:nn	1608, 2531, 2532
2337, 2349, 2358, 2366, 2473, 2499		\int_new:N	12, 13, 14, 23, 408, 409, 410,
\group_end:	154, 314, 1763, 1764,		411, 428, 1399, 1400, 1401, 1402,
2250, 2333, 2344, 2414, 2479, 2536			1405, 1778, 1779, 1780, 1781, 1782,
\group_insert_after:N	151		1783, 1784, 1785, 1786, 1787, 1790,
			1791, 1792, 2002, 2313, 2316, 2317
I			
\if_case:w	96, 1092, 1104,	\int_set:Nn	230, 556, 1148, 1414, 1455,
	1157, 1517, 1572, 1902, 2151, 2163		1457, 1550, 1604, 1605, 1616, 1628,
\if_charcode:w	319, 329, 752, 2051		

1652, 1673, 1722, 1805, 1820, 1839,	\l__regex_current_catcode_int
1927, 1957, 2119, 2345, 2482, 2504	.. 96, 115, 123, 137, <u>1783</u> , 1784, 1911
\int_set_eq:Nc 1090	\l__regex_current_char_int 34,
\int_set_eq:NN . 76, 520, 521, 557, 609,	40, 41, 48, 60, 61, 76, 77, 78, 83, 84,
902, 994, 1003, 1416, 1430, 1490,	116, 799, 1760, <u>1783</u> , 1783, 1865, 1909
1491, 1541, 1649, 1650, 1705, 1760,	\l__regex_current_pos_int
1807, 1810, 1817, 1818, 1819, 1834, 128, 148, 1438,
1837, 1865, 1866, 1870, 1996, 2441	<u>1753</u> , <u>1778</u> , 1780, 1805, 1807, 1810,
\int_step_function:nnnN 2397, 2507, 2515	1830, 1839, 1864, 1877, 1894, 1900,
\int_step_inline:nnnn 1814, 2781	1901, 1910, 1912, 1987, 1996, 2378
\int_sub:Nn 86, 707, 1708, 1716, 1725	\l__regex_current_state_int .. <u>1787</u> ,
\int_to_hexadecimal:n 2561, 2568	<u>1787</u> , 1917, 1919, 1921, 1922, 1927,
\int_use:c 1390	1928, 1950, 1953, 1957, 1958, 1966
\int_use:N . . . 537, 603, 667, 678, 688,	\l__regex_current_submatches_prop ..
692, 703, 704, 710, 711, 717, 718, <u>1788</u> , 1788, 1836,
885, 1448, 1450, 1483, 1484, 1485,	1930, 1959, 1960, 1977, 1986, 1998
1582, 1595, 1596, 1872, 1889, 1917,	\l__regex_default_catcodes_int
1958, 1966, 1987, 2016, 2017, 2028,	<u>410</u> , 411, 520, 521, 609, 902, 994, 1003
2060, 2066, 2131, 2345, 2569, 2759	\l__regex_empty_success_bool
\int_zero:N 519, 705, 1105,	.. <u>1794</u> , 1795, 1822, 1828, 1994, 2382
1339, 1415, 1804, 2047, 2109, 2339	\l__regex_every_match_tl
\iow_char:N 267, 269, 271, 273, 275, 277, <u>1793</u> , 1793, 1843, 1847, 1852
842, 843, 1271, 1296, 1380, 1381,	\l__regex_fresh_thread_bool
1382, 2552, 2555, 2558, 2561, 2565, 1772, 1775,
2568, 2611, 2615, 2620, 2640, 2642,	<u>1794</u> , 1794, 1875, 1937, 1939, 1995
2643, 2644, 2647, 2650, 2651, 2654,	\l__regex_group_level_int
2656, 2660, 2662, 2668, 2670, 2674,	<u>408</u> , 408, 519, 535, 537, 539, 995, 1000
2679, 2680, 2716, 2718, 2723, 2725,	\l__regex_internal_a_int
2730, 2733, 2738, 2742, 2752, 2787 10, 12, 656, 667,
	678, 688, 692, 701, 703, 707, 710,
	717, 1541, 1544, 1550, 1555, 1634,
	1649, 1655, 1662, 1663, 1673, 1676,
	1680, 1683, 1688, 1691, 1694, 1709,
	1717, 1726, 2109, 2119, 2120, 2131
	\l__regex_internal_a_tl 10, 10,
	226, 240, 944, 949, 964, 969, 974,
	978, 984, 985, 1176, 1206, 1216,
	1231, 1255, 1258, 1260, 1304, 1319,
	1338, 1361, 1372, 1454, 1455, 1456,
	1457, 1615, 1616, 1620, 1622, 2046,
	2071, 2394, 2415, 2487, 2513, 2543
	\l__regex_internal_b_int . 10, 13, 671,
	701, 704, 705, 707, 711, 718, 1650,
	1655, 1661, 1664, 1709, 1717, 1726
	\l__regex_internal_b_tl 10,
	11, 232, 237, 297, 301, 308, 359, 361
	\l__regex_internal_bool
 10, 15, 943, 948, 968, 977
	\l__regex_internal_c_int
 10, 14, 1652, 1657, 1658, 1659
L	
\l__regex_balance_int <u>23</u> ,	
23, 1804, 1898, 1903, 1904, 2047,	
2066, 2198, 2212, 2482, 2504, 2527,	
2531, 2532, 2539, 2540, 2544, 2545	
\l__regex_balance_tl	
..... <u>2003</u> , 2003, 2048, 2067, 2100	
\l__regex_capturing_group_int	
..... <u>1405</u> , 1405,	
1415, 1582, 1584, 1595, 1596, 1604,	
1605, 1608, 2097, 2442, 2509, 2517	
\l__regex_case_changed_char_int	
..... 51, 54,	
65, 68, 69, 76, 80, 86, <u>1783</u> , 1786, 1866	
\l__regex_catcodes_bool	
..... <u>410</u> , 412, 1108, 1112, 1147	
\l__regex_catcodes_int <u>410</u> , 410, 521,	
602, 603, 609, 885, 902, 994, 1003,	
1090, 1105, 1140, 1142, 1148, 1149	

\l__regex_internal_regex	1634, 1648, 1657, 1672, 1676, 1680,
426, 426, 518, 547, 576, 993, 1004,	1683, 1688, 1691, 1694, 1702, 1716,
1165, 1409, 2261, 2266, 2271, 2276	1719, 1722, 1725, 1729, 1745, 1774
\l__regex_internal_seq	\l__regex_right_state_seq
..... 10, 16, 1388, 1389, 1394 1401, 1404, 1449, 1456, 1620
\l__regex_last_char_int	\l__regex_saved_success_bool
..... 1760, 1783, 1785, 1865 143, 153, 1797, 1798
\l__regex_left_state_int	\l__regex_show_lines_int
..... 1401, 428, 428, 1308, 1339, 1343
1401, 1419, 1448, 1455, 1483, 1490,	\l__regex_show_prefix_seq
1497, 1500, 1501, 1503, 1504, 1530, 427,
1538, 1541, 1566, 1616, 1618, 1629,	427, 1258, 1260, 1300, 1312, 1317, 1319
1649, 1672, 1674, 1702, 1705, 1708,	\l__regex_start_pos_int
1711, 1723, 1736, 1745, 1769, 1774 834,
\l__regex_left_state_seq	1381, 1778, 1781, 1830, 1834, 1840,
..... 1401, 1403, 1447, 1454, 1615	2369, 2372, 2380, 2383, 2458, 2521
\l__regex_match_count_int	\l__regex_step_int ..
..... 2313, 2313, 2339, 2340, 2345 1790, 1790, 1818,
\l__regex_match_success_bool	1863, 1920, 1923, 1929, 1943, 1945
.. 1797, 1799, 1835, 1848, 1854, 1993	\l__regex_submatch_int
\l__regex_max_active_int 1820, 2316, 2316, 2379, 2381, 2384,
..... 1791, 1792, 1837,	2400, 2441, 2444, 2445, 2510, 2518
1870, 1876, 1884, 1971, 1972, 1976	\l__regex_success_pos_int
\l__regex_max_pos_int 1778, 1782, 1819, 1834, 1996, 2369
..... 835, 836,	\l__regex_success_submatches_prop ..
843, 1382, 1438, 1778, 1779, 1810, 1788, 1789, 1997, 2447
1877, 2378, 2380, 2383, 2492, 2521	\l__regex_zeroth_submatch_int
\l__regex_max_state_int 2316, 2317, 2371, 2373, 2441,
..... 1399, 1400, 1416,	2454, 2456, 2457, 2485, 2489, 2491
1430, 1485, 1486, 1489, 1491, 1492,	
1551, 1564, 1628, 1648, 1650, 1658,	
1705, 1711, 1719, 1729, 1805, 1815,	
1817, 1821, 2398, 2508, 2516, 2784	
\l__regex_min_active_int	
..... 1791,	
1791, 1817, 1837, 1870, 1872, 1876	
\l__regex_min_pos_int	
..... 833, 842, 1380, 1778, 1778, 1807, 1819	
\l__regex_min_state_int	
..... 1399,	
1399, 1416, 1430, 1815, 1838, 2782	
\l__regex_mode_int	
..... 409, 409, 455,	
463, 466, 475, 478, 487, 495, 498,	
508, 509, 511, 512, 557, 571, 866,	
869, 870, 871, 898, 909, 1091, 1092,	
1103, 1104, 1156, 1157, 1207, 1249	
\l__regex_replacement_csnames_int ..	
..... 2002, 2002, 2058,	
2060, 2062, 2099, 2151, 2157, 2163,	
2168, 2174, 2176, 2197, 2211, 2759	
\l__regex_right_state_int	
... 1401, 1402, 1422, 1436, 1450,	
1457, 1484, 1490, 1491, 1530, 1537,	
1543, 1556, 1564, 1566, 1618, 1622,	

O

\or: . 97, 98, 99, 100, 1519, 1574, 1903, 1904

P

\prg_do_nothing:	532,
569, 570, 577, 578, 2057, 2184, 2234	
\prg_new_conditional:Npnn .	367, 389, 441
\prg_new_protected_conditional:Npnn	
.....	746, 2281, 2286, 2299, 2301
\prg_replicate:nn	
538, 1526, 1552, 1659, 1706, 1714,	
2062, 2430, 2435, 2442, 2540, 2545	
\prg_return_false:	
.....	373, 374, 377, 382, 385, 394,
397, 402, 404, 447, 450, 753, 758, 2323	
\prg_return_true:	373, 377, 382,
385, 394, 397, 402, 445, 749, 755, 2321	
\prop_clear:N	1836
\prop_map_inline:Nn	2447
\prop_new:N	1788, 1789
\prop_put:Nno	1986
\prop_set_eq:NN	1997
\ProvidesExplPackage	4

R	
\regex_const:Nn	7, 2258, 2268
\regex_count:NnN	2291, 2293
\regex_count:nnN	8, 2291, 2291
\regex_extract_all:NnN	2295, 2307
\regex_extract_all:nnN	2295, 2307
\regex_extract_all:NnNTF	2295
\regex_extract_all:nnNTF	9, 2295
\regex_extract_once:NnN	2295, 2305
\regex_extract_once:nnN	2295, 2305
\regex_extract_once:NnNTF	2295
\regex_extract_once:nnNTF	8, 2295
\regex_gset:Nn	7, 2258, 2263
\regex_match:Nn	2286
\regex_match:nn	2281
\regex_match:NnTF	2281
\regex_match:nnTF	7, 2281
\regex_new:N	7, 2256, 2256
\regex_replace_all:NnN	2295, 2311
\regex_replace_all:nnN	2295, 2311
\regex_replace_all:NnNTF	2295
\regex_replace_all:nnNTF	10, 2295
\regex_replace_once:NnN	2295, 2309
\regex_replace_once:nnN	2295, 2309
\regex_replace_once:NnNTF	2295
\regex_replace_once:nnNTF	9, 2295
\regex_set:Nn	7, 2258, 2258
\regex_show:N	2273, 2279
\regex_show:n	7, 2273, 2273
\regex_split:NnN	2295, 2312
\regex_split:nnN	2295, 2312
\regex_split:NnNTF	2295
\regex_split:nnNTF	9, 2295
\RequirePackage	6
\reverse_if:N	40, 41, 60, 61, 68, 69
S	
\s_seq	2396
\scan_stop:	439, 1816, 1899, 1920, 1923, 1933, 1971, 2041, 2187, 2240, 2373, 2380, 2391, 2444, 2458, 2464, 2469
\seq_clear:N	1300
\seq_get:NN	1615, 1620
\seq_map_function:NN	1312, 1394
\seq_new:N	16, 427, 1403, 1404
\seq_pop:NN	1454, 1456
\seq_pop_right:NN	1258, 1319
\seq_push:No	1447, 1449
\seq_put_right:No	1260
\seq_put_right:Nx	1317
\seq_set_filter:NNn	1389
\seq_set_split:Nnn	1388
\str_case:nn	930
\str_case:nnF	1378
\str_case_x:nnn	662
\str_if_eq:nnTF	641, 732, 910, 941, 961, 1018, 1053, 1068, 1106
\str_if_eq_x:nnF	2764
\str_if_eq_x:nnTF	125, 317, 339, 684, 698, 1174, 2107, 2161, 2677
T	
\tex_advance:D	869, 1142, 1657, 1658, 1663, 1664, 1863
\tex_dimen:D	1816, 1919, 1922, 1928, 1953
\tex_divide:D	870
\tex_escapechar:D	230, 556, 1414
\tex_lccode:D	2187, 2240
\tex_muskip:D	1894, 1910, 1912, 2039, 2040
\tex_skip:D	1886, 1971, 2009, 2010, 2033, 2034, 2039, 2040, 2079, 2080, 2371, 2373, 2379, 2444, 2456, 2457, 2463, 2464, 2468, 2469, 2491
\tex_the:D	128, 148, 1463, 1469, 1472, 1887, 1921, 2026, 2457, 2787
\tex_toks:D	128, 148, 1462, 1463, 1468, 1469, 1472, 1489, 1661, 1662, 1887, 1900, 1921, 1976, 2026, 2787
\tl_clear:N	2048
\tl_const:Nn	18
\tl_const:Nx	2271
\tl_gset_eq:NN	2266
\tl_map_function:NN	1219
\tl_map_function:nN	616
\tl_new:N	10, 11, 17, 1793, 2003
\tl_put_right:Nn	2100
\tl_set:Nn	1847, 1852, 1930, 1959, 2415, 2537
\tl_set:Nv	1206
\tl_set:Nx	232, 301, 308, 361, 944, 949, 1176, 2394, 2487, 2513
\tl_set_eq:NN	2261
\tl_to_lowercase:n	306, 2188, 2241
\tl_to_str:N	2155
\tl_to_str:n	8, 616, 2277
\tl_to_str:V	8, 2166
\token_if_eq_charcode:NNTF	351, 355, 922, 1129, 1192
\token_if_eq_meaning:NNT	928
\token_if_eq_meaning:NNTF	623, 955, 1086, 1119, 1187, 1190, 2116, 2142

\token_to_str:N	244, 251, 1021, 1024, 2280	U	
\trace:nmx	.. 1480, 1803, 1812, 1917, 2786	\use:c 531
\trace_pop:nnn	\use:n 106,
...	239, 1425, 1443, 1577, 1623, 2072		496, 499, 1261, 1312, 1394, 1796, 1832
\trace_pop:nmx 1824	\use:x 614, 1868, 1948, 2412, 2534
\trace_push:nnn	\use_i:nn 456, 467, 476, 479, 488
...	225, 1413, 1429, 1561, 1613, 2045	\use_i:nnn 2121
\trace_push:nmx 1802	\use_ii:nn	.. 458, 464, 469, 481, 490, 1331
		\use_none:n 108, 263, 502
		\use_none:nn 1604, 2132