1.4. Regular Expressions

Prev Chapter 1. TROVA: Contents search

1.4. Regular Expressions

Regular expressions allow users to create complicated queries. Below follows a list of most commonly used regular expressions together with explanations and some potential uses.

- [abc] means "a or b or c", e.g. query "[br]ang" will match both "adbarnirrang" and "bang"
- [^abc] means "begins with any character but a,b,c", e.g. query [^aeou]ang will match "rang" but not "baang"
- [a-zA-Z] means "a charater from a/A through z/Z", e.g. b[a-zA-Z] will match "bang", "bLang" or "baang" but not "b8ng"
- . (the dot) means "any character", e.g. "b.ng" will match "bang", "b8ng", but not "baang"
- X* means "X zero or more times", e.g. "ba*ng" will match "bng", "bang", "baang", "baaang" etc.
- X+ means "X one or more time", e.g. "ba+ng" will match "bang", "baang" but not "bng"
- ^ means "the beginning of the annotation", e.g. "^ng" will match "ngabi" but not "bukung"
- $ means "the end of the annotation", e.g. "ung$" will match "bukung" but not "ngabi"

  Examples

  - ^[pbtd][^aeiou]

    You can use this expression to search for complex onsets. It will find words that start with one of the plosives ("p","b","t","d") followed by a character that is not a vowel ("a","e","i","o","u"). An example of a matching word is "tsakeha"

  - [^n]g$

    You can use this expression in case you want to search for annotations ending with a "g", but not with "ng". In Dutch, you will find "snelweg" and "maandag" as the results but not words as "bang".

  - ^k.+k$

    You can use this expression if you want to search for annotation starting and ending with "k" and with one or more character between them, e.g. "kitik" or "kanak-kanak"

  - ^(.+)\1$

    You can use this expression to search for words that are reduplicated. When you put something in bracketes, you create a variable (.+), which you can refer to as "\1". This expression then searches for an annotation that starts with one or more random characters followed by that same sequence of characters. This expression will match for instance "kulukulu".

More about regular expressions...

The following tables have been created by a user of ELAN (an annotation tool which has the same search mechanism as TROVA). They may result quite useful also for other users since they offer a simple and clear overview of the main symbols (partly different from the ones just seen) used in regular expressions, with a short explanation and an example for each of them. Bear in mind that the examples are taken from the language that the user is being researching, so do not pay attention to the meaning of the words but to the working mechanism of the regular expressions.

Table 1.1. Symbols

Symbols	Place	Meaning
\b	at the beginning and/or end of a string	word boundary
\w+	at the end of a string	variable end of word
.	anywhere	any letter
.*	between spaces	any string of letters between spaces/any word
.*\	between spaces	any string of words
(x|y)	anywhere	either x or y
[^x]	place at the beginning	not x
(....)\l	anywhere	words with four reduplicated letters
?	after a letter	the preceding letter is optional

Table 1.2. Search for particular word forms (examples)

Symbols	Hits	Examples
sa	all words containing the string sa	sa, vasaku, sahata, tisa
\bsa	all words starting with sa	sa, sahata, sana; NOT vasaku, tisa
\bsa\b	all words sa	sa
\bsa..\b	all words consisting of sa + two letters that follow sa	saka, saku, sana
\bsa\w+	all words beginning with sa, but not the word sa by itself	sahata, sana
\b.*ana\b	al words ending in ana	sinana, tamuana, sana, bana, maana
(....)\l	all words with four reduplicated letters	pakupaku, vapakupaku, mahumahun, vamahumahun
\b(....)\l	all words beginning with four reduplicated letters	pakupaku; NOT vapakupaku
\b(....)\lana\b	all words beginning with four reduplicated letters and ending in ana	vasuvasuana, hunuhunuana
\bva(....)\l	all words consisting of the prefix va- + four reduplicated letters	vapakupaku, vagunagunaha
\bvahaa?\b	all tokens of vahaa and vaha	vahaa and vaha

Table 1.3. Search for particular sequences of words (examples)

Symbols	Hits	Examples
\bsaka\b .* \bhaa	string of 3 words: (1) saka; (2) any word; (3) the word haa by itself or with suffixes	saka antee haa; saka abana haari; saka kabuu haana

Symbols	Hits	Examples
saka .* \bhaa\w+	string of 3 words: (1) saka; (2) any word; (3) a word beginning with haa, but NOT the word haa by itself	saka abana haari; saka kabuu haana
(\bsaka\b|\bsa\b) \bpaku\b	2-word string consisting of saka or sa and paku	saka paku; sa paku
(\bsaka\b|\bsa\b) .* \bvaha\b	strings of 3 words: (1) saka or sa; (2) any word; (3) vaha	saka tii vaha; sa tapaku vaha
(\bsaka\b|\bsa\b) (....)\l \bhaa	strings of 3 words: (1) saka or sa; (2) any word with four reduplicated letters; (3) the word haa or a word beginning with haa	sa natanata haa; saka natanata haana

Prev Up
1.3. Multiple Layer Search Tab Home