

Searches with Regular Expressions in ELAN corpora

Ulrike Mosel, Kiel University (umosel@gmx.de)

(13 July 2015)

Contents

- 1 Regular Expression (Regex) characters and simple searches
- 2 Searches for complex wordforms
- 3 Searches for syntactic constructions
- 4 Multi-layer searches for transcriptions and their translation equivalents
- 5 Multi-layer and multi-column searches for syntactic and morphological constructions

These guidelines are based on my experiences with searches for Teop and English wordforms and constructions in the Teop Language Corpus that is compiled in ELAN (Brugman & Russel 2004). They do not cover all kinds of searches you can do with regular expressions in ELAN, but only those I most frequently use myself for finding

- wordforms with variable spellings, particular phonotactic patterns, e.g. consonant clusters in the end of the word
- wordforms with particular derivational and inflectional morphemes, reduplication, etc.;
- collocations, sequences of particular wordforms with or without gaps, e.g. *the _____ woman*;
- particular senses of homonymous wordforms, e.g. Teop *ta* 1. some, any, 2. part, 3. 'to'
- all English translational equivalents of a particular wordform of the researched language

Search modes (see ELAN Manual chapter 7)

Single Layer Search, Multiple Layer Search

View

Keyword in context (KWIC), Frequency view, Alignment view

References:

Brugman, H., Russel, A. (2004). Annotating Multimedia/ Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.

Friedl, Jeffrey E. F. 2006. *Mastering Regular expressions*. Beijing, Cambridge etc.: O'Reilly.

URL: <http://tla.mpi.nl/tools/tla-tools/elan/>

1. Regular Expression (Regex) characters and simple searches

Table 1: Regex characters

symbol	place	meaning and examples
white space	anywhere	white space, e.g. < d.p > matches <i>had put, and pulled</i>
.	anywhere	matches any one character including white space, e.g. < d.p > matches <i>had put, and pulled, adapt, departure</i>
\	before a character	the character following \ does not have the value it usually has in Regex, e.g. \. 'period', not 'any character' as in < d.p >, e.g. < \. I\b matches all clauses that start with I within an annotation.
\b	before/after a string	word boundary, not the letter <i>b</i>
\w	anywhere	matches any word character, i.e. any letter <i>a</i> to <i>z</i> , any digit 0 to 9, and <u></u>
\w+	anywhere	one or more word characters; useful for searches of wordforms with a variable beginning, middle or end; \w+men matches <i>women, fishermen moment</i> , etc., by itself between white spaces it means "any wordform", the white space functions as a character, but it is not a word character
.*		matches any character 0 or more times, including white space
(abc xyv)	anywhere	matches the sequence <abc> or <xyz>, e.g. < (mo fa)ther > matches <i>mother, father</i> , but also <i>grandmother, grandfather, motherhood, fatherhood</i>
[abc]	anywhere	matches < a > or < b > or < c >, e.g. < wom[ae]n > matches <i>woman</i> and <i>women</i> .
[a-z][0-9]	anywhere	any letter <i>a</i> to <i>z</i> , e.g. < \bb[a-z]a > matches <i>beach, board, black, branch</i> ; any digit 0 to 9
[^x]	before a character	any character except < x >, e.g. < \bb[^eo]a > matches <i>black, branch</i> , but not <i>beach, board</i> , because <i>e</i> and <i>o</i> are excepted
(..)\1	anywhere	matches any sequence of two characters that are repeated once (reduplicated), e.g. <i>coconut, banana, competition, weeded</i>
?	after a character	preceding character is optional, e.g. < \bboys?\b > matches <i>boy</i> and <i>boys</i>
(xyz)?	anywhere	the string xyz is optional, e.g. < \bchild(ren)?\b > matches <i>child</i> and <i>children</i>
x{n}	after a character	x exactly n times, e.g. < \b[aeiou]{2} > matches all words starting with two vowels <i>early, out, air</i> etc.
x{n, }	after a character	x at least n times
\A	before a Regex	matches the beginning of an annotation
\Z	after a Regex	matches the end of an annotation

2. Searches for complex word forms

Table 2: Combinations of symbols on word level (case insensitive)

symbols	matches	examples
<code>in</code>	<i>in</i> and all wordforms containing <i>in</i>	<i><u>in</u>, <u>into</u>, <u>drinking</u>, <u>again</u></i>
<code>\bin</code>	<i>in</i> and all wordforms starting with <i>in</i>	<i><u>in</u>, <u>into</u>, <u>inside</u>, <u>indeed</u></i>
<code>\bin\w+</code>	all wordforms starting with <i>in</i>	<i><u>into</u>, <u>inside</u>, , <u>indeed</u>, <u>invitation</u></i>
<code>\b(i u)n\w+ed\b</code>	all wordforms starting either with <i>in</i> or with <i>un</i> and ending with <i>ed</i>	<i><u>indeed</u>, <u>informed</u>, <u>unsalted</u></i>
<code>\w+in</code>	all wordforms starting with one or more wordcharacters followed by <i>in</i>	<i><u>again</u>, <u>rain</u>, <u>morning</u>, <u>find</u>, <u>finished</u></i>
<code>\w+in\b</code>	all wordforms ending in <i>in</i>	<i><u>again</u>, <u>rain</u>,</i>
<code>\b\w(in){2}</code>	all wordforms starting with one word character followed by two groups of (<i>in</i>)	<i><u>lining</u></i>
<code>\b\w(\w{2})1</code>	all words starting with any word character followed by any 2 word characters that are repeated once.	<i><u>banana</u>, <u>remember</u></i>
<code>\block.*?\bat\b</code>	the wordform <i>look</i> and all wordforms starting with <i>look</i> that are directly or later in the same annotation followed by <i>at</i>	<i><u>look at</u>, <u>looking at</u>, <u>looked with disgust at</u></i>

3 Searches for phonotactics/combinations of word characters

Table 3: Searches for Vowels and consonant clusters

<code>[b-df-hj-np-tvwxz]</code>	<i>b</i> or <i>c</i> or <i>d</i> or <i>f</i> or <i>g</i> ..., i.e. any consonant letter of the English alphabet	
<code>[aeiou]</code>	<i>a</i> or <i>e</i> or <i>i</i> or <i>o</i> or <i>u</i> , i.e. any vowel letter of the English alphabet	
<code>\b[aeiou]{2}\w+</code>	all wordforms starting with any combination of two vowels	<i><u>early</u>, <u>out</u>, <u>eat</u></i>
<code>\b[b-df-hj-np-tvwxz]{3}\w+</code>	all wordforms starting with CCC	<i><u>splashed</u>, <u>threw</u>, <u>scrape</u></i>
<code>\b[aeiou][b-df-hj-np-tvwxz]{3}\b</code>	all wordforms starting with a vowel and ending with three consonants	<i><u>arms</u>, <u>ends</u>, <u>ants</u></i>
<code>\b[aeiou][b-df-hj-np-tvwxz]{3,}</code>	all wordforms starting with a vowel and followed by at least three consonants	<i><u>instead</u>, <u>angry</u>, <u>instrument</u>, <u>asthma</u></i>

4 Searches for Numbers

Table 4: Digits and sequences of digits

19..	match 1 followed by 9 followed by any 2 characters	1952, 19:1, 19a)
\d	any digit	1, 2, 3, 4, 5, 6, 7, 8, 9, 0
19\d\d	match 1 followed by 9 followed by any two digits	1901, 1999
\d{4}	any sequences of 4 digits	1964, 2003, 9999

5. Searches for constructions

Table 5: Search for NPs, progressive forms and sentence initial items

	Regex	examples
1.	NPs formed by the article <i>a</i> , <i>an</i> or <i>the</i> , with the headnoun <i>woman</i> or <i>women</i> , and any word between the article and the headnoun	
	\b(an?!the)\b \w+ \bwom.n\b	<i>a pregnant woman</i> <i>an other woman</i> <i>the old woman, the two women,</i>
	1) boundary followed by either <i>a</i> or <i>an</i> or <i>the</i> , followed by a boundary, followed by white space 2) string of one or more word characters, followed by white space 3) boundary followed by the string <i>wom</i> , followed by any character, followed by <i>n</i> , followed by boundary	
2.	The wordform <i>is</i> or <i>was</i> followed by a word form ending in <i>-ly</i> followed by a wordform ending in <i>-ing</i>	
	\b(i wa)s\b \w+ly\b \w+ing\b	<i>was slowly crawling</i> <i>was strongly holding</i> <i>is only sitting</i>
	1) boundary, followed by either <i>i</i> or <i>wa</i> , followed by <i>s</i> , followed by a boundary, followed by a white space, 2) boundary, followed one or more word characters, followed by <i>ly</i> , followed by boundary, followed by white space 3) wordform starting with one or more word characters, followed by <i>ing</i> , followed by boundary	
3.	Sentences that start with the pronoun I	
	(\A \.)bI\b (case: sensitive)	<i>It's alright. I am not well.</i> <i>I am going to tell a story.</i>
	Either at the beginning of an annotation or after a period followed by white space, followed by a boundary followed by < I > followed by a boundary	

6 Multi-layer search for transcriptions and their translation equivalents

Multilayer search is useful if you want to find examples of a particular sense of a **homonymous** lexical item or a functional word as, for instance, the Teop non-specific article *ta* ‘any, some’ which is homonymous with the noun *ta* ‘part’ and the complementizer *ta*.

Figure 1: Multilayer search for *ta* with the translation ‘any’ or ‘some’, Frequency by frequency view

The screenshot shows the 'Search eaf files' window with the 'Multiple Layer Search' tab selected. The search criteria are as follows:

- Domain: 218 eaf files
- Query History: < > New Query
- Mode: case insensitive, regular expression
- Search criteria:
 - Minimal Duration: \bta\b
 - Maximal Duration: (empty)
 - Begin After: (empty)
 - End Before: (empty)
 - Overlap: Overlap
 - Search criteria: \b(some|any)\b
 - Tier Type: t (transcription)
 - Must be in same file: Must be in same file
 - Tier Type: f (free translation)

Results: Found 141 hits in 141 annotations (of 252344). Frequency 1 - 11 of 137.

Percentage	Count	Annotation
1.42%	2	#1 ["Uh, dee maa ta iana!"] #2 ["Uu, give me some fish!"]
1.42%	2	#1 ["Uul dee maa ta iana!"] #2 ["Uul Give me some fish!"]
1.42%	2	#1 ["Uuu, dee maa ta iana!"] #2 ["Uuu, give me some fish!"]
1.42%	2	#1 [Ahiki ta kiu nae.] #2 [It does not have any use.]
0.71%	1	#1 [Ahiki ta otei to rakerake unoman?] #2 ["Doesn't any boy want to marry you?"]
0.71%	1	#1 ["Bara! Gaga koa nasu ta ruene."] #2 ["Alright! Go ahead and drink some water."]
0.71%	1	...a maa kara ariono, eam repaa vaave koa nio obai." #2 ["If there is not any ariono vine, we do the thatching with obai vine."]

Multilayer searches are also practical, if you do not know the language well and want to search for a word and all its translations. Then you search on the free translation tier with <.*>. For example, the search for *mararae* gives you the translations ‘happy’, ‘joyful’ and ‘joy’.

Figure 2: Multilayer search for *mararae* with any translation, Concordance view

The screenshot shows the 'Search eaf files' window with the 'Multiple Layer Search' tab selected. The search criteria are as follows:

- Search criteria: \bmararae\b
- Overlap: Overlap
- Search criteria: .*
- Tier Type: t
- Must be in same file: Must be in same file
- Tier Type: f

Results: Found 70 hits in 70 annotations (of 260302). hit 1 - 11 of 70.

Annotation
#1 [Eori he mararae bata, eori he tea karavi bata "Eh,] #2 [They were happy, they were surprised, "Hey,]
#1 [175. eori repaa paku bona maa hagi teori, amaa mararae teori.] #2 [they do their dances, their joyful actions/ excited movements.]
#1 [176. Kahi vataaree ni rori bona mararae teori,] #2 [They will show their joy,]
#1 [327. A meha otei vai to mararae kurusu batana paa sue, "Ah!] #2 [One man who was very happy said, "Ah!]
#1 [E sinanae sa mararae kurus haa.] #2 [His mother was not very happy.]

7 Multi-layer and multi-column search for syntactic and morphological constructions

This kind of search is useful if you want to search for two items in juxtaposed annotations or annotations that are separated by one or more annotations, for example when searching on a tier with morphological glossings. The search below is on two adjacent annotations on the transcription and translation tiers. The task is: "Find all examples for the Teop translation equivalent of English *say, says, said* followed by direct speech marked by <"> in the next annotation.

Table 6: Find the Teop translation equivalents of 'say, says, said' followed by direct speech

first layer	.* (any annotation)	-0 ann. (no annotation between first and second column)	"	Tier type: t (transcription)
	overlap		overlap	must be in the same file
second layer	sa(ys? id)\b (say or says or said)	-0 ann. (no annotation between first and second column)	"	Tier type: f (free translation)

Figure 3: Find the Teop translation equivalents of 'say, says, said' followed by direct speech, Alignment view

The screenshot shows a search interface with the following components:

- Mode:** case insensitive, regular expression
- Search Criteria:** Minimal Duration, Maximal Duration, Begin After, End Before
- Search Results:** Found 671 hits in 671 annotations (of 280603)
- Time scale:** 1 sec.
- Search Results Table:**

Time	Teop	English
00:01:39.382	Erau, ore paa tei, voosu vai maa e sinariori ore paa sue,	"Gaga han tea ruene vai to paa gono mau e subuava ei"
	And so they stayed, and now their mother came home and he said,	"Drink from this water that this old woman got."
00:02:41.658	Erau kao tavusu mepaa ma sue ki bene - e mutanae,	"Eh, havee to gonogono vasaku koa maanae subuava bona ruene?"
	And so he came and said to his wife,	"Hey, where does the old woman get the water so quickly?"