Searches with Regular Expressions in ELAN corpora

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

(13 July 2015)

Contents

These guidelines are based on my experiences with searches for Teop and English wordforms and constructions in the Teop Language Corpus that is compiuled 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 *had put, and pulled*
.	anywhere	matches any one character including white space, e.g. < d.p > matches *had put, and pulled, adapt, departure*
\	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 *b*
\w	anywhere	matches any word character, i.e. any letter *a* to *z,* any digit 0 to 9, and _
\w+	anywhere	one or more word characters; useful for searches of wordforms with a variable beginning, middle or end; \w+men matches *wo<u>men</u>, fisher<u>men</u> mo<u>ment</u>,* 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|xvy)	anywhere	matches the sequence <abc> or <xyz>, e.g. < (mo|fa)ther > matches *mother, father,* but also *grandmother, grandfather, motherhood, fatherhood*
[abc]	anywhere	matches < a > or < b > or < c > , e.g. < wom[ae]n > matches *woman* and *women.*
[a-z] [0-9]	anywhere	any letter *a* to *z*, e.g. < \bb[a-z]a > matches *<u>bea</u>ch, <u>boa</u>rd, <u>bla</u>ck, <u>bra</u>nch* ; any digit 0 to 9
[^x]	before a character	any character except < x >, e.g. < \bb[^eo]a > matches *black, branch*, but not *beach, board*, because *e* and *o* are excepted
(..)\1	anywhere	matches any sequence of two characters that are repeated once (reduplicated), e.g. *<u>coco</u>nut, ba<u>nana</u>, competi<u>tion</u>, <u>weeded</u>*
?	after a character	preceding character is optional, e.g. < \bboys?\b > matches *boy* and *boys*
(xyz)?	anywhere	the string xyz is optional, e.g. < \bchild(ren)?\b > matches *child* and *children*
x{n}	after a character	x exactly n times, e.g. < \b[aeiou]{2} > matches all words starting with two vowels *early, out, air* 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
in	*in* and all wordforms containing *in*	*in, into, drinking, again*
\bin	*in* and all wordforms starting with *in*	*in, into, inside, indeed*
\bin\w+	all wordforms starting with *in*	*into, inside, , indeed, invitation*
\b(i|u)n\w+ed\b	all wordforms starting either with *in* or with *un* and ending with *ed*	*indeed, informed, unsalted*
\w+in	all wordforms starting with one or more wordcharacters followed by *in*	*again, rain,morning, find, finished*
\w+in\b	all wordforms ending in *in*	*again, rain,*
\b\w(in){2}	all wordforms starting with one word character followed by two groups of *(in)*	*lining*
\b\w(\w{2})\1	all words starting with any word character followed by any 2 word characters that are repeated once.	*banana, remember*
\blook.*?\bat\b	the wordform *look* and all wordforms starting with *look* that are directly or later in the same annotation followed by *at*	*look at, looking at, looked with disgust at*

3 Searches for phonotactics/combinations of word characters

Table 3: Searches for Vowels and consonant clusters

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

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 sequencs of 4 digits	1964, 2003, 9999

5. Searches for constructions

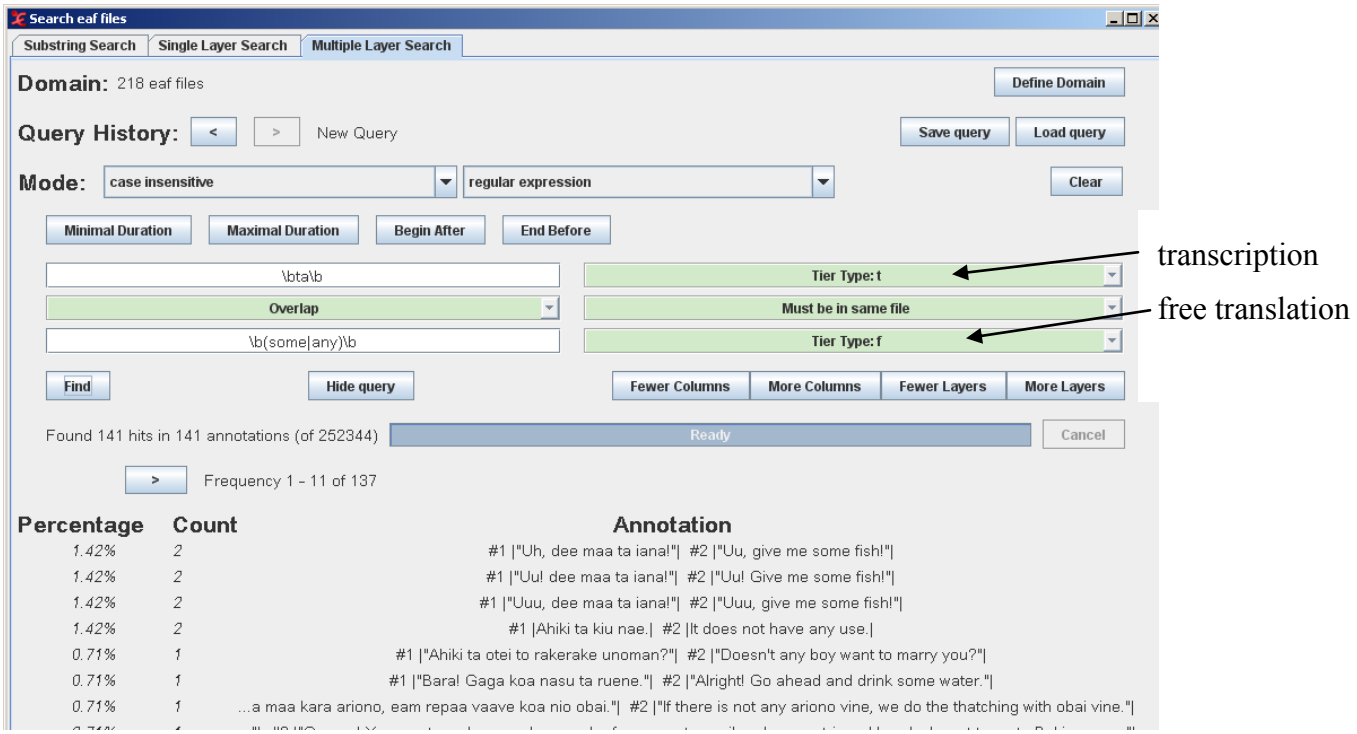
Table 5: Search for NPs, progressive forms and sentence initial items

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

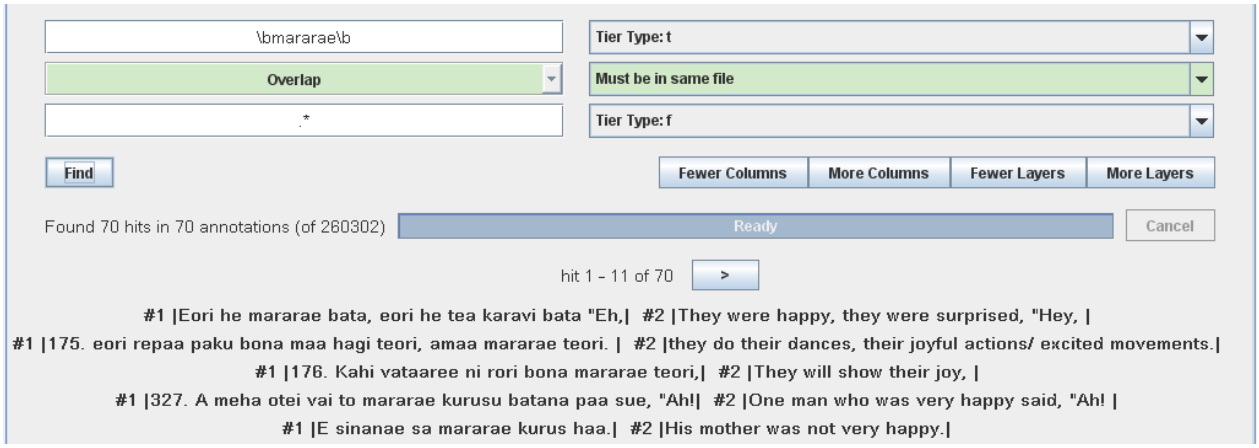
Multilayer search is useful if you want to find examples of a particular sense of a **homonymous** lexical item or a functional word as, for instance, the Teop non-specific article *ta* 'any, some' which is homonymous with the noun *ta* 'part' and the complementizer *ta.*

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



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



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 transation equivalents of 'say, says, said' followed by direct speech

	first column		second column	
first layer	.* (any annotation)	-0 ann. (no annotation between first and second column)	"	Tier type: t (transcription)
	overlap		overlap	must be n 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