Regensburg, Sep. 2011

Regular expressions for ELAN users

Ulrike Mosel
umosel@linguistik.uni-kiel.de

Symbols

Table 1

symbol	place	meaning
\b	at the beginning and/or the 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
(....)\1	anywhere	words with four reduplicated letters
?	after a letter	preceding letter is optional

Search for particular word forms

Table 2: Combine symbols to find words with particular beginnings, endings and reduplications

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* and two letters that follow *sa*	*saka, saku, sana,*
\bsa\w+	all words beginning with *sa,* but not *sa* by itself	*sahata, sana*
\b.*ana\b	all words ending in *ana*	*sinana, tamuana, sana, bana, maana*
\b[^(bana|maana)].*ana\b	all words ending in *ana*, but not *bana* or *maana*	*sinana, tamuana, sana*
(....)\1	all words with four reduplicated letters	*pakupaku, vapakupaku, mahumahun, vamahumahun*
\b(....)\1	all words beginning with four reduplicated letters	*pakupaku* NOT: *vapakupaku*
\b(....)\1ana\b	all words beginning with four reduplicated letters and ending in *ana*	*vasuvasuana, hunuhunuana*
\bva(....)\1	all words with the prefix *va-* and	*vapakupaku, vagunagunaha*

Regensburg, Sep. 2011

	four reduplicated letters	
\bvahaa?\b	all tokens of *vahaa* and *vaha*	*vahaa* and *vaha*

Searching for particular sequences of words

Table 3: Combine: \b, .*. \w+ and (x|y)

	symbols	hits	examples
1.	\bsaka\b .* \bhaa	string of 3 words: (1) *saka* (2) any word, and (3) the word *haa* by itself or with suffixes	*saka antee haa;* *saka abana haari;* *saka kabuu haana*
2.	saka .* \bhaa\w+	string of 3 words: (1) *saka* (2) any word, and (3) a words beginning with *haa*, but not *haa* by itself	*saka abana haari;* *saka kabuu haana*
3.	(\bsaka\b|\bsa\b) \bpaku\b	all 2 word strings that consist of *saka* or *sa* and *paku*	*saka paku, sa paku*
4.	(\bsaka\b|\bsa\b) .* \bvaha\b	all 3 word strings with (1) *saka* or *sa,* (2) any word (3) *vaha*	*saka tii vaha* *sa tapaku vaha*
5.	(\bsaka\b|\bsa\b) (....)\1 \bhaa	all 3 word strings with (1) *saka* or *sa,* (2) a word with four reduplicated letters (3) the word *haa* or a word beginning with *haa*	*sa natanata haa,* *saka natanata haana*

Comments on Table 4:

saka/sa ... haa is a discontinuous negation. The last component *haa* can have a suffix that indicates imperfective aspect and person, e.g. *haana, haari, haara.* The formulars above provide data for the following questions:

1. Which words are used inbetween *saka* and *haa/haana/haari/haara* ?

2. Which words are used inbetween *saka* and *haana/haari/haara* ?

3. Are there examples for *saka/sa* followed by *paku* ‘do’?

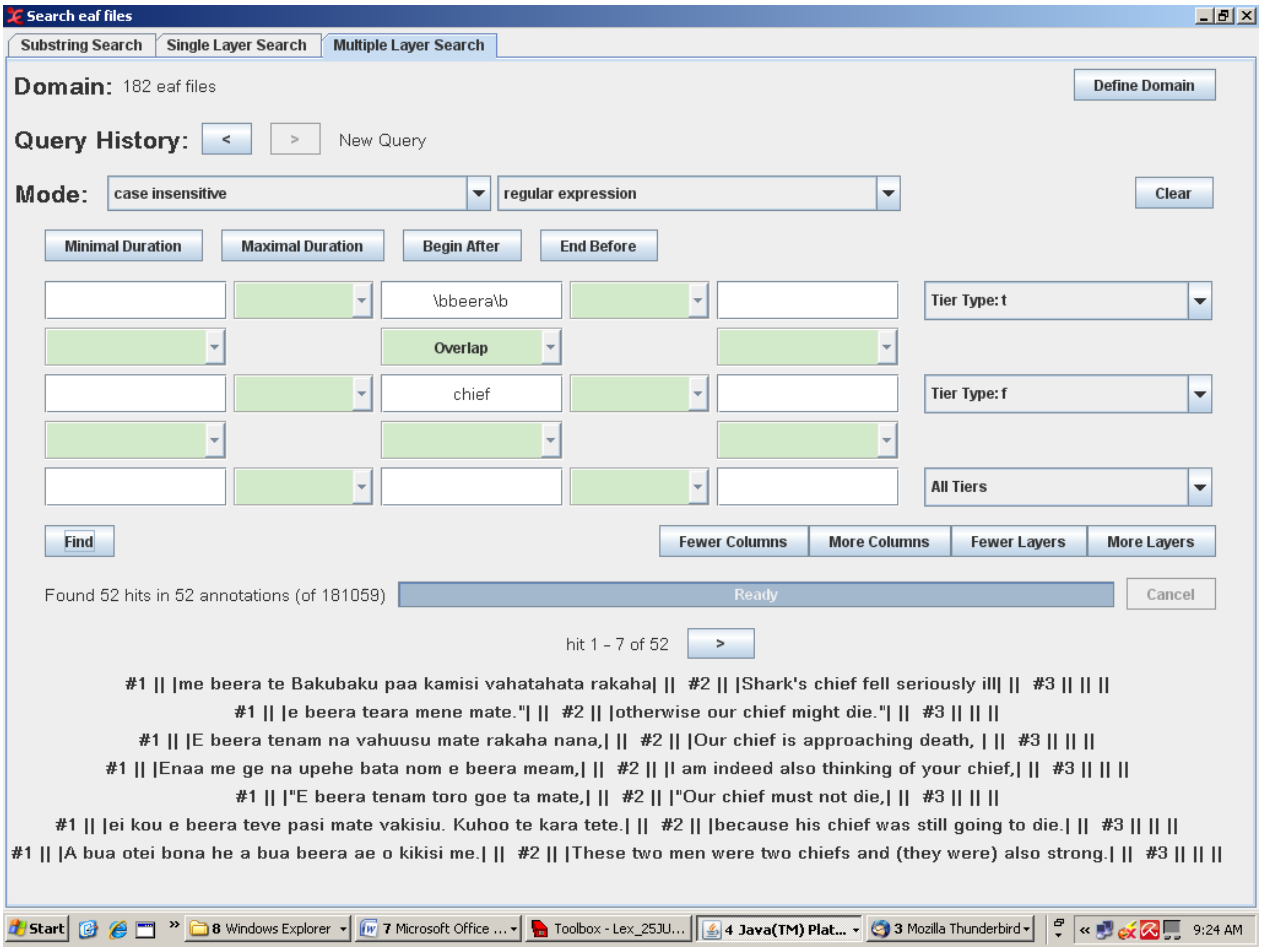
4. Which words are used between *saka/sa* and *vaha* ‘back, also, again, anymore’?

5. Does *saka/sa ... haa* combine with reduplicated words?

Regensburg, Sep. 2011

Multilayer search with regular expressions

Multilayer search is useful if you want to find examples for one meaning of polysemous or homonymous lexical items. For example, *beera* means 'big' and 'chief, chiefs and chiefly'. If I want to search only for the second meaning, I use multilayer search on the transcription (t) and the free translation tier (f):



Multilayer search is also pratical, if you do not know the language well and you want to search for the lexical item and any of its translations. Then you search on the free translation tier with a wild card: .*

Note that there are still some bugs in ELAN. In the following example you see in line 4 and 6 funny things on the right hand side. This is not the translation tier, but our 'notes' tier. But otherwise the hits are fine. You see that *beera* is translated by 'older', 'big' and 'important'.

Regensburg, Sep. 2011

Find all contexts with *beera* and its translation:

transcription tier: \bbeera\b

translation tier: .*