Regular Expressions

Primer 2

First Edition

Table of contents

Preface

This primer builds on the previous primer and is designed to help users learn regular expressions in the Language technology context. Primer 1 introduced the basic "alphabet" to help users read real-life examples of regular expressions to help with Paratext.

It noted that making **regular** changes to your data requires writing an **expression** to match the text.

In the Primer 2, users will revise the basic "alphabet" and build regular expressions to clean-up a dictionary conversion.

The RegEx Alphabet

In additon to English letters, numbers and punctuation, here are the basic symbols used in RegEx

Symbol	Meaning
.	any character
\	turn special character back to normal character
^	beginning of line, NOT
*	0 or more
+	one or more
$	end of line
[ ]	set
{ }	repetitions
( )	store in a group use \0 \1 \2 \3 ... \9 to recall
|	OR
?	optional, lazy
\d	numbers
\s	white space (spaces, tabs, enter)
\w	word building characters
\b	word break
\uxxxx	Unicode character xxxx

Summary of Primer 1

In Primer 1 we learnt the following:

Lesson	RegEx	Meaning	Example
1	.	anything	(.*)
	\.	.	
	\r\n	End of line/paragraph mark	(.*)\r\n
2	\d	Any digit 0123456789	\\v\d
	\D	Any non-digit	
3	\w	Word building	\\f \w \d
	\W	Non-word building	
4	\s	White space (space, tab, enter…)	\s*(\\f*)
	\S	Any non-space	
5	Review		
6	()	Save in group	\\r(.*)\r\n
	\0 \1 \2 \…\9	Recall group (\0 all matched text \1 first brackets \2 second brackets …)	

Lesson	RegEx	Meaning	Example
7	*	0 or more	
	+	One or more	\w+(-\w+)+
	?	Optional (lazy)	
8	{x,y}	Between x and y times	\\\w{1,3}
9	[]	Define a set of characters	[\w*.,!;:]
	[^]	Any characters not in set	
10	Review		
11	\b	Word break	t\b
	\B	Non-word break	
12	\u	Unicode character	\u2013
13	^	Beginning of line	^\\[^q]
	$	End of line	
14	|	OR	(\\f|\\x)

Common SFMs in dictionaries

RegEx for dictionaries can look confusing with a mix of RegEx symbols and MDF markers. Keep in mind some of the common MDF markers

MDF	RegEx	used for
\lx	\\lx	lexeme
\hm	\\hm	Homograph number
\ph	\\ph	phonetic
\gn	\\gn	Gloss (n)
\dn	\\dn	Definition (n)
\ps	\\ps	Part of speech
\se	\\se	Sub-entry
\sn	\\sn	Sense number
\xv	\\xv	Example in vernacular
\xn	\\xn	Example in national
|b	\\|b	Bold
|r	\\|r	Regular

Here are some very useful RegExs for Dictionaries.

(\\[^l].*\r\n)(\\hm .*\r\n) (task 1)

Find: (\\[^l].*\r\n)(\\hm .*\r\n) (task 2)
Replace: \2\1

\r\n[^\\] (task 3)

([\w*.,?!;:])"([\s.,?!;:\u2013\u2014\\]) (task 4)

Find: \\q (task 5)
Replace: \\q1

Find: (\\[^lh].*\r\n)(\\lc .*r\n) (task 6)
Replace: \2\1

Find: t\b (task 7)
Replace: ɗ

Find: (\\ph)\[(.*)\] (task 8)
Replace: \1\2

Find: \\([^x].*)[\.|,]\s*\r\n (task 9)
Replace: \\\1\r\n

Find: (\\sn \d)\) (task 10)
Replace: \1

We are going to find which of these tasks you can use them for.

1. Find all \hm markers which are not immediately after \lx.
2. Move the \hm marker so that it is directly after the \lx.
3. Find a new line which doesn't begin with an SFM marker.
4. Convert straight quotes used at the opening and closing of quotations to angled (curved) quotes.
5. Change all \q markers to \q1.
6. Make sure all \lc are directly after either \lx or \hm if it exists.
7. Change word final 't' to 'ɗ'.
8. Remove the phonetic brackets from the \ph field.
9. Remove punctuation marks from some markers (leave the example sentences and translations).
10. Remove brackets from the sense numbers.

Tasks

Task 1 – check ordering of markers

Find all **\hm** markers which are **not** immediately after **\lx**.

What is the \hm marker: \\hm .*\r\n

What should be above it? \\lx .* \r\n

What shouldn't be above it? \\[^l]

Answer: \\[^l].*\r\n\\hm .*\r\n

Task 2 – reorder markers

Move the **\hm** marker so that it is directly after the **\lx**.

We found the misplaced \hm markers in the previous task. Now we need to move the \hm to be above it, that is change the order of the markers. But we need to store the text first.

Answer: Find: (\\[^l].*\r\n)(\\hm .*\r\n)
Replace: \2\1

Task 3 – using NOT

Find *a* **new line** which **doesn't begin** with *an* **SFM** marker.

What is *a* new line? _____ What is NOT? ___

What does *a* SFM marker begin with? ___

Answer: \r\n[^\\]

Task 4 – convert straight quotes to curved quotes

Convert straight quotes used *at* the opening *and* closing of quotations to angled (curved) quotes.

This is identical to the *task* in Primer 1 (for Paratext).

Task 5 – changing markers

Change all \q markers to \q1.

This is identical to a task in Primer 1 (for Paratext) and can also be done with a normal search/replace.

Task 6 - check ordering of markers

Make sure all \lc are directly after either \lx or \hm if it exists.

This is very similar to task 1.

Task 7 – change text

Change word final 't' to 'd'.

This is identical to a task in Primer 1 for Paratext. Here we need a regular expression because we only want the word final 't' and not the others.

Task 8 – remove punctuation

Remove the phonetic brackets from the \ph field.

Hint: You need to break up the \ph field into the marker then brackets and whatever comes between the brackets and the new line. Remember that brackets have a special meaning in RegEx so you need to make sure you find literal brackets.

Answer: Find: (\\ph)\[(.*)\]
Replace: \1\2

Task 9 – remove punctuation several markers

Remove punctuation marks from some markers (leave the example sentences and translations).

Hint: You could either use OR to specify the ones you want, or use NOT to exclude the ones you don't want.

Answer: Find: \\([^x].*)[\.|,]\s*\r\n
Replace: \\\1\r\n

Task 10 – remove final punctuation

Remove brackets from the sense numbers.

Hint: You want to keep the marker and whatever the number is, but remember that brackets have a special meaning in RegEx so you need to make sure you find literal brackets.

Answer: Find: (\\sn \d)\)
Replace: \1

Assignment

Cleanup the Baatonum dictionary (which you converted to Unicode in the previous session). Several macros have been run to convert the dictionary to SFM. However there are three more changes that need to be done to prepare the file for import into FLEx. (these are details in Tasks 8-10 above).

Datafile: Baatonum-u.sfm

Tasks: Perform the final cleanup to get the file ready to import into FLEx.

Suggestions:

- Open the Baatonum-u.sfm file in an editor that will accept regular expression searches.
- Remove the phonetic brackets from the \ph field.
- Remove the final punctuation from all markers except the example sentence (\xv) and it's translation (\xn).
- Remove the bracket from the \sn line.
- Check the order of the \hm fields.