Thoughts on Regular Expressions for Ethiopic

በዳንኤል፡ያዕቆብ

Supposing that we could write regular expressions natively in Ethiopic; is the regular expressions language itself sufficient for identifying Ethiopic tokens? At the very least regex implementations would have to be updated to distinguish between Ethiopic syllables, numerals and punctuations for the detection of word boundaries.

Applying the `UNICODE` basic range for Ethiopic some simple definitions fall out:

`[ሀ-፼]`	`:==`	Any Single Ethiopic Character (`\ce` in Emacs)
`[^ሀ-፼]`	`:==`	Any Single None-Ethiopic Character (`\Ce` in Emacs)
`[ሀ-ፚ]+`	`:==`	Any Syllabic Element
`-?([0-9]+|[፩-፼]+)`	`:==`	Any Integer Value
`[። - ፨]`	`:==`	Any Single Ethiopic Punctuation *excluding wordspace*
`[ ፡\n\t]*`	`:==`	Zero or more space characters

For lack of the address range continuity, the implementation of the above gets messy if the ISO/`UNICODE` extensions are ever made to the Ethiopic character set...

Shortly however we would find that the traditional operators and syntax will not be sufficient, or at least not convenient, to meet all of the needs in Ethiopic text processing. Consider if you will a need to detect any 2^{nd} (**ካዕብ**) order of an Ethiopic syllable (as in the case of the male definite and 3^{rd} person male possessive articles). We might then be lead to construct a macro of the form:

ካዕባት	`:==`	`[ሁሉሑሙሡሩሱሹቁቑቡቩቱቹኁኑኙኡኩኹዉዑዙዡዩዱዹጁጉጙጡጩጱጹፁፉፑ]`

Which suffices but does not offer the same utility as would an operator available for this same purpose. Existing operators offering this service are not known to the author. The modulus operator, '`%`' in C, is suggested here as it is otherwise without special meaning in regex languages. At the implementation level for Ethiopic under `UNICODE` the operation is nearly just that but here we are really specifying the remainder in keeping with practices in speech:

`[ሀ-ፗ]%7`	Any syllable of the 7^{th} order.
`[ሀ-ፗ]%ኦ`	The same expressed with a vowel.
`[ሀ-ፗ]%[1-5,7-]`	Any syllable of orders 1^{st} through 5^{th}, and 7^{th} onward[1].
`[ሀ-ፗ]%[ኡ-ኤ,ኦ-]`	The same expressed with vowels.
`[ሀ-ፗ]%[^እ]`	The negative expressions is simplest of all.
`([ሀ-ፖ]%እ)+`	A consonant cluster such a "**ትምህርት**".

[1] `UNICODE` specifies 12 orders for the **ቀ**, **ቐ**, **ኀ**, **ከ**, **ኸ** and **ገ** series. But **ሟ**, **ፘ** and **ፚ** may be considered 13^{th} orders of their respective bases.

The use of vowels is convenient here, however a complication arises were we to try and specify an 11^{th} order syllable class. In this event it would be simpler to allow for mixed use of numbers and vowels within the braces as per `[ኡ,11]` rather than to try and contrive additional vowels just for this purpose or to make one more operator (such a '+') active as per `[ኡ,አ+10]` or `[ኡ,አ+4]`.

The above would then be sufficient to form regular expression to detect Amharic plural suffixes -ኣት, -ኦች, -ኣዎች as follows:

```
AmharicPlurals  :==  ([ሀ-ፖ]+)(%ኣ/ት)?(%ኦ/ች)?
```

Without an Ethiopic aware lexicographical analyzer the same expression might be formed to detect the tokens in a buffer of `SERA` transliterated text as follows:

```
 SERAConsonant  :==  [b-df-hj-np-tv-zB-DF-HJ-NP-TV-Z]
AmharicPlurals  :==  (<SERAConsonant>)+(at)?(o/c)
```

Which ultimately becomes as unsatisfactory as our **ካዕባት** macro and in addition we have lost the multilingual context of our text altogether and need to take careful steps to be sure of the language context of the stream.

The operator '%' may until this point have appeared to be an end fix operator but could also be applied in prefix notation:

%ኦ/ች	Equivalent to `([ሀ-ፖ]+)%ኦ/ች`
\ኣዊ	Alternately, Ethiopic vowels could be applied to define escapes to serve the same purpose as '%'.
(%[ኡ,ኣ-ኦ])ን?(ም|ስ)?+(%ኣ|ና)?	A common formation at the end of words.

The regular expression languages applied in `UNIX`, `Perl` and `Lex` are drawn upon in the above discussion. It is acknowledged that the creation of a new operator could have detrimental effects on backwards compatibility in regex languages. However, given the other requirements of Ethiopic and other non-Roman writing systems upon lexicographical analyzers, not the least of which will be basic token recognition of the text elements in multibyte text streams, the consequences to the languages applying the operator will be little or none.

A new operator to detect the syllabic order will be of great utility for the text processing of Ethiopic and other syllabaries, it should however be carefully and exhaustively considered first if existing operators, perhaps not meaningful to syllabaries, may be simply applied to serve in the desired context.