Improving Language Translation accuracy in Mosip

The documentation clearly explains how the improvement has been made between Sinhala to English and English to Sinhala translation.

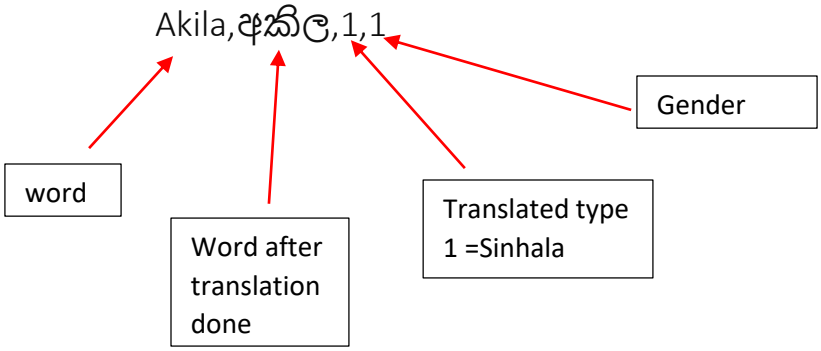
We should need to do the following changes before run the translit application.

1. Clone the translit repo in GitHub.
2. If it is not the latest update, adding below txt files to resources
    - en-to-si.txt
    - rules-en.txt
    - rules-si.txt
3. Run the java application or build the project using this command [mvn clean install].
4. Check the accuracy of the transliteration words.
5. Integrate jar file to Mosip kernel core and check the translation performance level of the Mosip.

1. *en-to-si.txt*

```
Akshima,අක්ෂිමා,1,1
Akila,අකිල,1,1
Akila,අකිලා,1,1
Akuressa,අකුරැස්ස,0,0
Agamanpodi,අගමංපොඩි,1,1
Agampodi,අගම්පොඩි,0,0
Agampodige,අගම්පොඩිගේ,1,1
Agnes,අග්නස්,1,1
Agida,අගිදා,1,1
Ages,අගේස්,0,0
Agosthina,අගොස්තිනා,1,1
```

The en-to-si.txt describe the English to Sinhala dictionary words in translit application. Further deeply refer the format of the dictionary words like below.



2. *rules-en.txt*

```
0,290,mahathmaya#%,11,m.ah.at.m.aj.a:#%
0,291,mahatmaya#%,10,m.ah.at.m.aj.a:#%
0,143,vitharana%,9,ʋ.it.a:r.aɳ.a%
0,144,witharana%,9,ʋ.it.a:r.aɳ.a%
0,160,thilaka#%,8,t.il.ak.a#%
0,260,sriyanth%,8,ʂ.r.ij.an.t.%
0,266,aratchie%,8,a:r.ac.c.i%
0,273,#ambalan%,8,#am.b.al.an.%
0,156,rathna#%,7,r.at.n.a#%
0,161,tilaka#%,7,t.il.ak.a#%
0,162,nayaka#%,7,n.a:j.ak.a#%
1,166,wimala#%,7,ʋ.im.al.a#%
2,167,wimala#%,7,ʋ.im.al.a:#%
0,170,anthony%,7,æn.t.an.i%
```

The rules-en.txt describe the English language to phonetic rules which have been used in the translit application. Further deeply refer the format of “en-rules” like below.

0,290,mahathmaya#%,11,m.ah.at.m.aj.a:#%

- The first comma separates describe the gender of each rule.
- The second index of the rule set.
- The Third comma separates describe the English rule.
- Fourth comma separates explain the length of the rule.

- fifth comma will separate the phonetic of each rules.

3. *rules-si.txt*

```
0,91,කොලඹ%,6,colombo %
0,90,දූව%,3,duwa%
0,89,ඩ%,2,D%
0,88,්රුවන්%,6,.r.uw.an.%
0,80,්ර%,3,.r%
0,81,්ය%,3,.j%
0,87,න්ර%,3,n.r%
0,89,ල්ර%,3,l.r%
0,83,ඔ%,2,j.%
```

The rules-en.txt describe the Sinhala language to phonetic rules which have been used in the translit module. Further describe the format of this rule like below.

0,18,ඔඟ%,1,au%

- The first comma separates describe the gender of each rule.
- The second index of the rule set.
- The Third comma separates describe the Sinhala rule.
- Fourth comma separates explain the length of the rule.
- fifth comma will separate the phonetic of each rules.

Conclusion:

We can add more rules to rules-en.txt and also rules-si.txt. through that we can get comprehensive improvement of the translation process. Adding rules might be support to increase the accuracy of the translation process.