Myanmar Spell Correction

MANGO AI

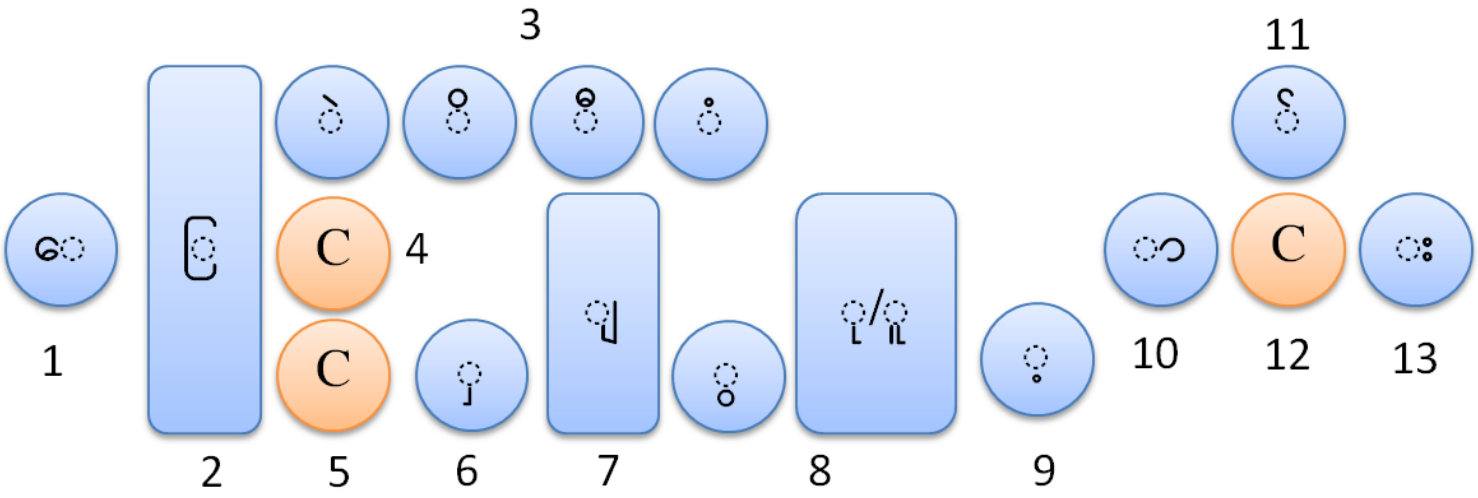
Myanmar Language

- Myanmar Languages are made up of different glyphs, derived from Brahmic Scripts. Each separable syllable constitutes different combination of glyphs, following the well-defined syllabic rules (Maung & Mikami, 2008).
- In addition, Myanmar Languages do not specifically define Word-Boundary, unlike Latin Languages.
- Therefore, spell-checking and correction pose different challenges for Myanmar Languages.
- Naturally, grammatical analysis starts with validation of individual syllable, which comprises of different glyphs. Invalid glyphs will always result in invalid syllables, which in turn results in Invalid Words.

Myanmar Unicode

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+100x	က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	စျ	ဉ	ည	ဋ	ဌ	ဍ	ဎ	ဏ
U+101x	တ	ထ	ဒ	ဓ	န	ပ	ဖ	ဗ	ဘ	မ	ယ	ရ	လ	ဝ	သ	ဟ
U+102x	ဠ	အ	ဣ	ဤ	ဥ	ဦ	ဧ	ဩ	ဪ	ါ	ာ	ိ	ီ	ု	ူ
U+103x	ေ	ဲ	ံ	့	း	္	်	ျ	ြ	ွ	ှ	ဿ
U+104x	၀	၁	၂	၃	၄	၅	၆	၇	၈	၉	၊	။	၌	၍	၎	၏
U+105x	ၐ	ၑ	ၒ	ၓ	ၔ	ၕ	ၖ	ၗ	ၘ	ၙ	ၚ	ၛ	ၜ	ၝ	ၞ	ၟ
U+106x	ၠ	ၡ	ၢ	ၣ	ၤ	ၥ	ၦ	ၧ	ၨ	ၩ	ၪ	ၫ	ၬ	ၭ	ၮ	ၯ
U+107x	ၰ	ၱ	ၲ	ၳ	ၴ	ၵ	ၶ	ၷ	ၸ	ၹ	ၺ	ၻ	ၼ	ၽ	ၾ	ၿ
U+108x	ႀ	ႁ	ႂ	ႃ	၍	၎	၏	ၐ	ၑ	ၒ	ၓ	ၔ	ၕ	ၖ	ၗ	ၘ
U+109x	ၙ	ၚ	ၛ	ၜ	ၝ	ၞ	ၟ	ၠ	ၡ	ၢ	ၣ	ၤ	ၥ	ၦ	ၧ	ၨ

Myanmar Syllable



Fundamental Problems

- The basic idea of Spell-Checking itself is pretty much straight-forward.
- There is a collection of words (dictionary), and we can check every word in the text to check against the dictionary.
- However, there are 3 major problems:
  - **Out of Dictionary Words**: some words in the text may not appear in the dictionary.
  - **Speed**: if we have **M** Words in the text and **N** words in the dictionary, then we will have to check **M x N** times, which is practically impossibly slow.
  - **Accuracy**: Some words may be similar but not particularly the same.

Specific Problems

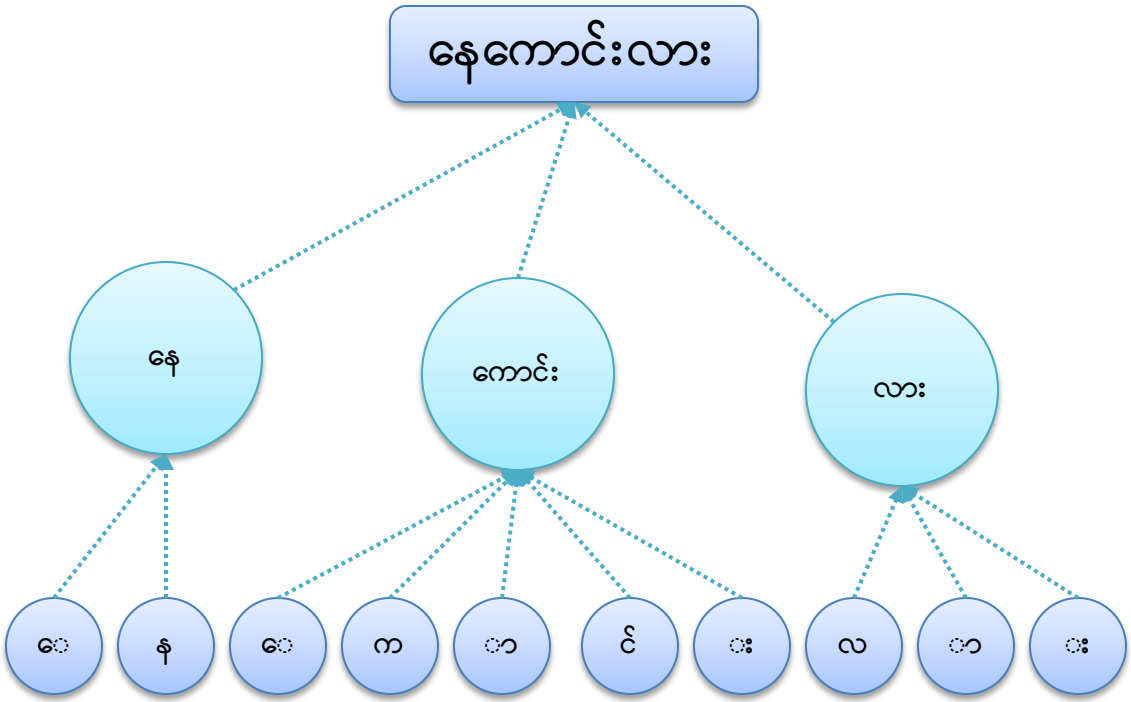
- In addition to Fundamental Problems of Spell Checking, Myanmar Languages pose additional specific problems.
- Firstly, Myanmar Languages do not define specific Word Boundary, which makes it quite challenging to extract individual words in Myanmar Sentences.

မနေ့ကထမင်းစားနေတုန်းမှာ

- Secondly, Myanmar Languages are agonistic to Word Orders.

ထမင်းမနေ့ကစားနေတုန်းမှာ

Glyphs to Words

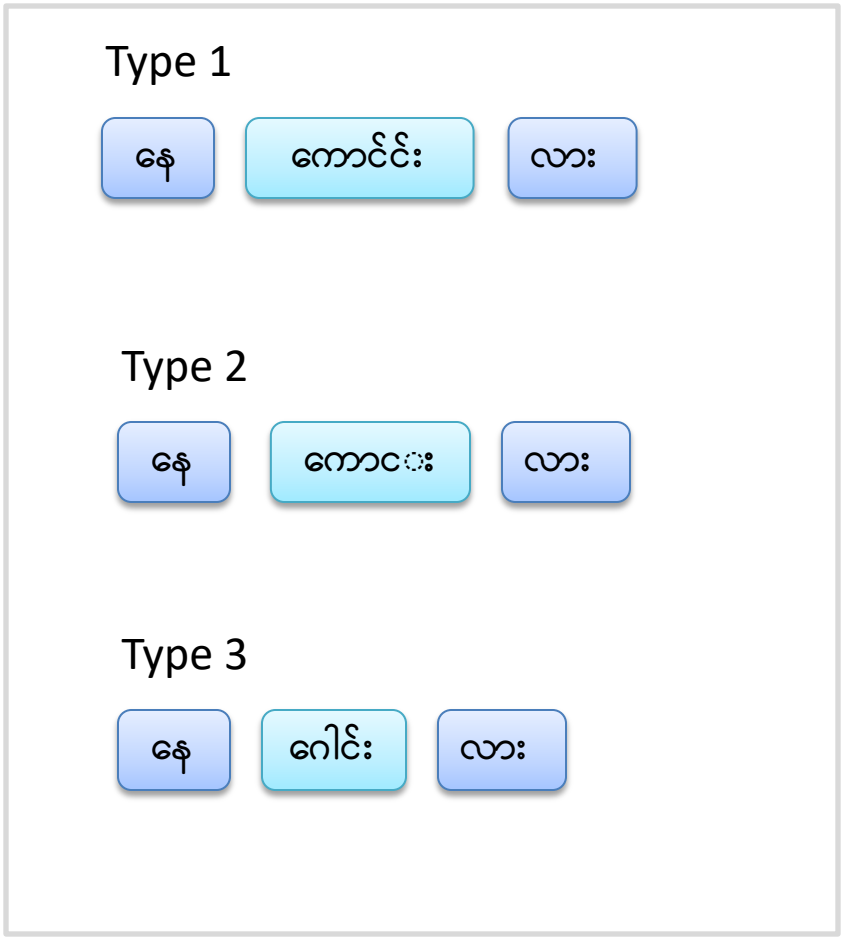


Glyphs

Simplification

- However, if we assume Myanmar Spelling Errors are the result of Glyph Errors to Syllables, namely:

- Glyph Addition
- Glyph Subtraction
- Glyph Change



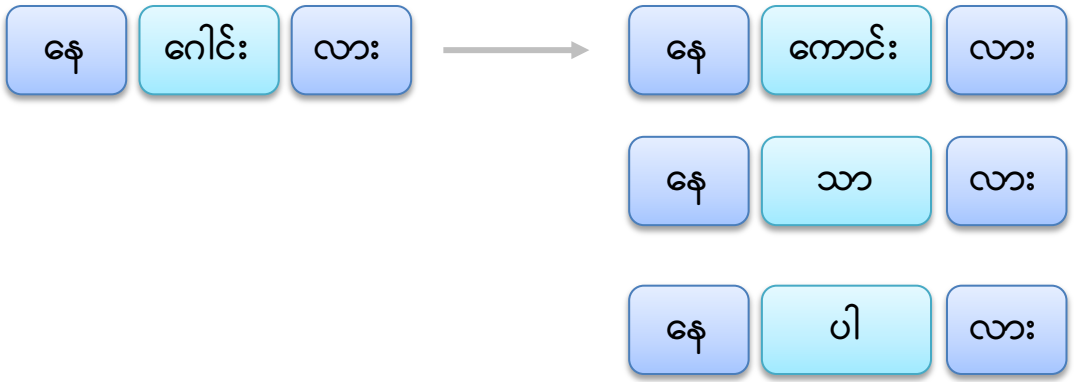
Spell Checking

- Sentences in the text are segmented into Syllables. If there are some errors in Glyphs, they will result in some improper Syllables, which do **not follow** Syllabic Rules. Then, we will know that there are spelling errors. This will solve **Type 1** and **Type 2** Errors.

- **Type 3** Errors are a bit more challenging to detect, as they all have Proper Syllables. However, if we have well-defined Collection of bi-grams, then it is relatively straight-forward to detect as well.

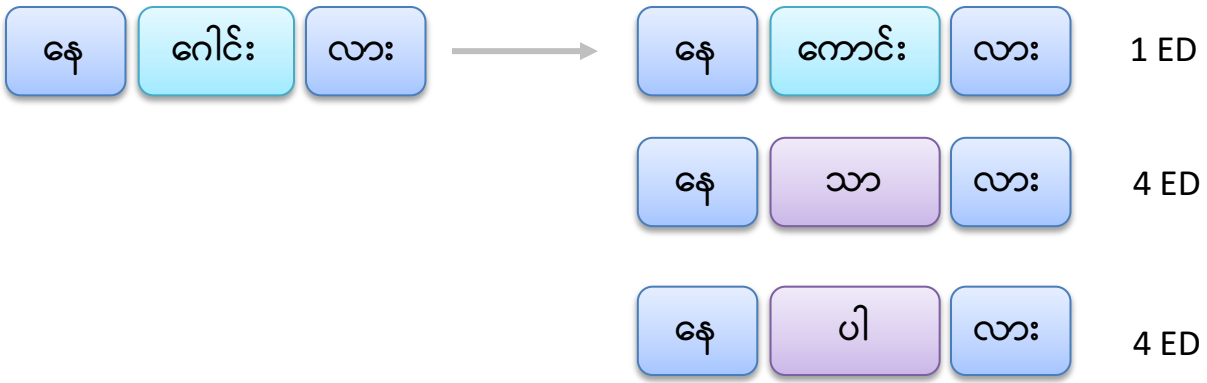
Spell Correction

- Once spelling errors are detected, it is possible to correct such errors.
- However, Spell Correction is altogether a challenging problem as well, as there are many possibilities.



Minimum Edit Distance

- Out of many possibilities, the most likely corrections are calculated from Minimum Edit Distance, which defines how many Inserts, Deletes or Edits (Edit Distance) are necessary for the words to be the same.

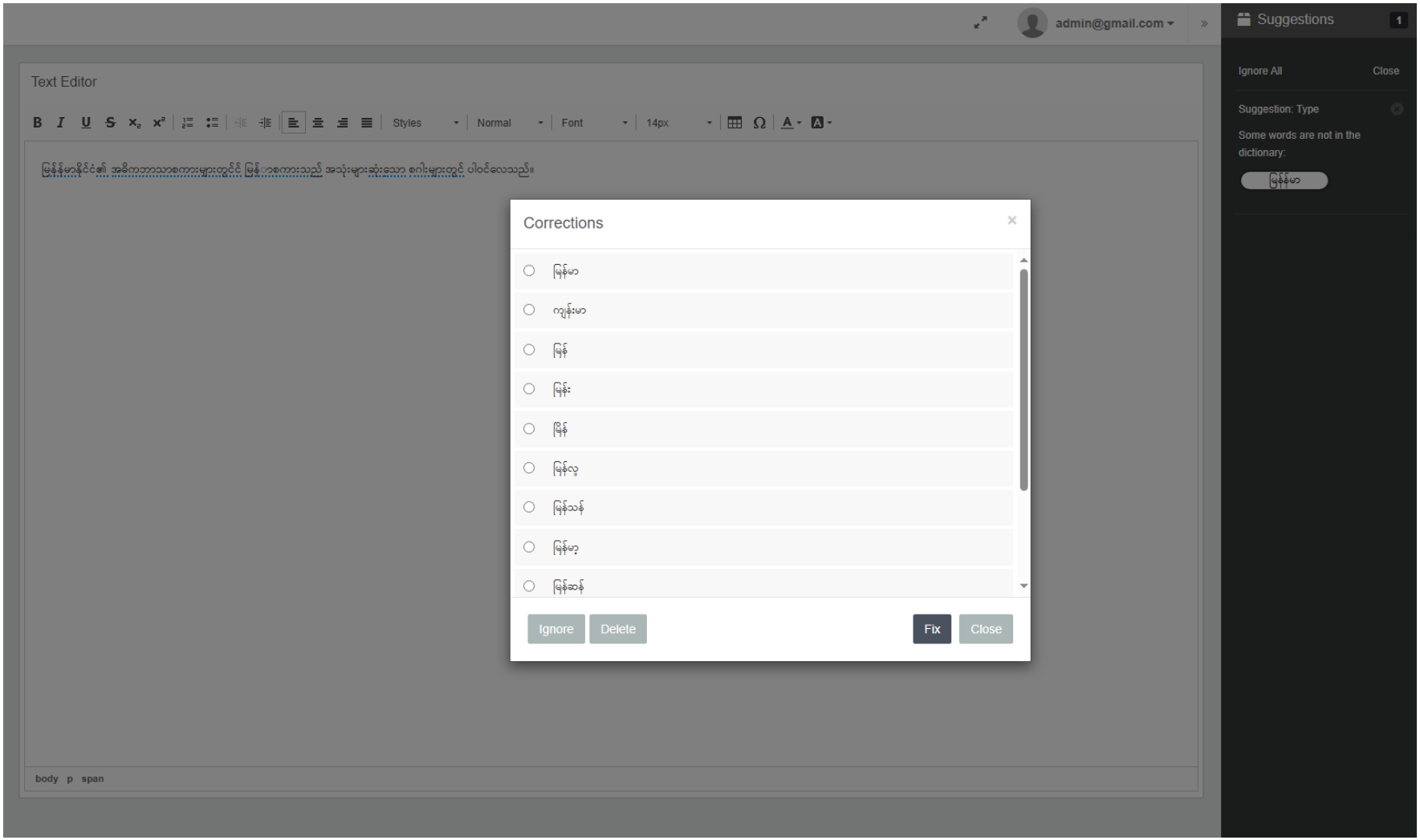


Optimization

- However, calculating Minimum Edit Distance is a really time-consuming process, and therefore, a simple MED Computation is not enough.
- Fortunately, we have a really fast MED Computation Algorithm, known as SymSpell, developed by Wolf Garbe[1].
- For SymSpell to work, we need a collection of words along with their relative frequency, and we have created a collection of 1 million words with relative frequency.

1. https://github.com/wolfgarbe/SymSpell

Results


admin@gmail.com
Suggestions
1
Text Editor
Styles
Normal
Font
14px
မြန်နိုင်ငံ၏ အဓိကဘာသာစကားများတွင်ပင် မြန်ာစကားသည် အသုံးများဆုံးသော စကါးများတွင် ပါဝင်လေသည်။
Ignore All
Close
Suggestion: Type
Some words are not in the dictionary:
မြန်နိုင်မာ
Corrections
မြန်မာ
ကျန်းမာ
မြန်
မြန်း
မြိန်
မြန်လူ
မြန်သန်
မြန်မာ့
မြန်ဆန်
Ignore
Delete
Fix
Close
body p span