



---

# Khmer Spell Checker

---

Presented By: Sochenda KHEM



---

# Outline



- Problem
- Discussion/Solution
- Results
- Conclusion



# Problem



→ The rich of Khmer character confuses user to write one word in many different ways.

→ Ex:

→ (correct) សម្រេច

→ (incorrect) សំរេច

→ after passing segmentation សំរេច = សំ + រេច

→ For human being:

→ សំរេច is a one word error.



# Discussion/Solution



## → Spelling errors:

→ non-word error: Typographic errors that result in a valid dictionary not intended by typist

→ Ex: Form <> From

<>

→ real word error: Words that are invalid in the dictionary

→ Ex: address <> adres

<>

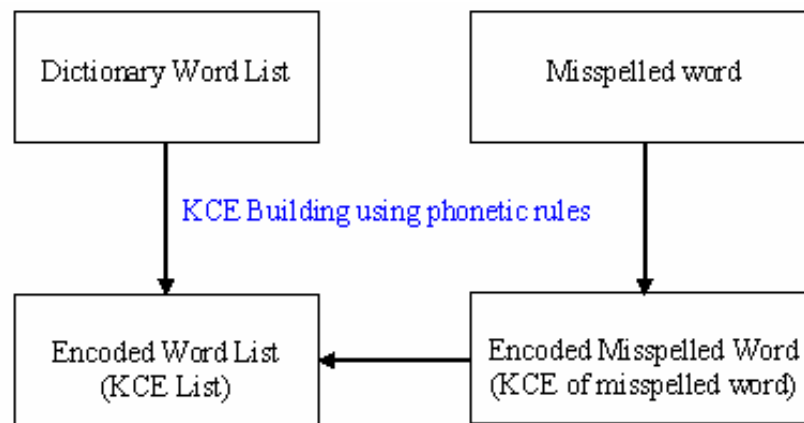


# Discussion/Solution



## → Khmer Common Expression (KCE)

- Detect sound similarity
- Encode the misspelled word and the dictionary word list into expression that is based on how it is pronounced



To search the KCE of the misspelled word in the encoded word list rather than the misspelled word in the dictionary word list

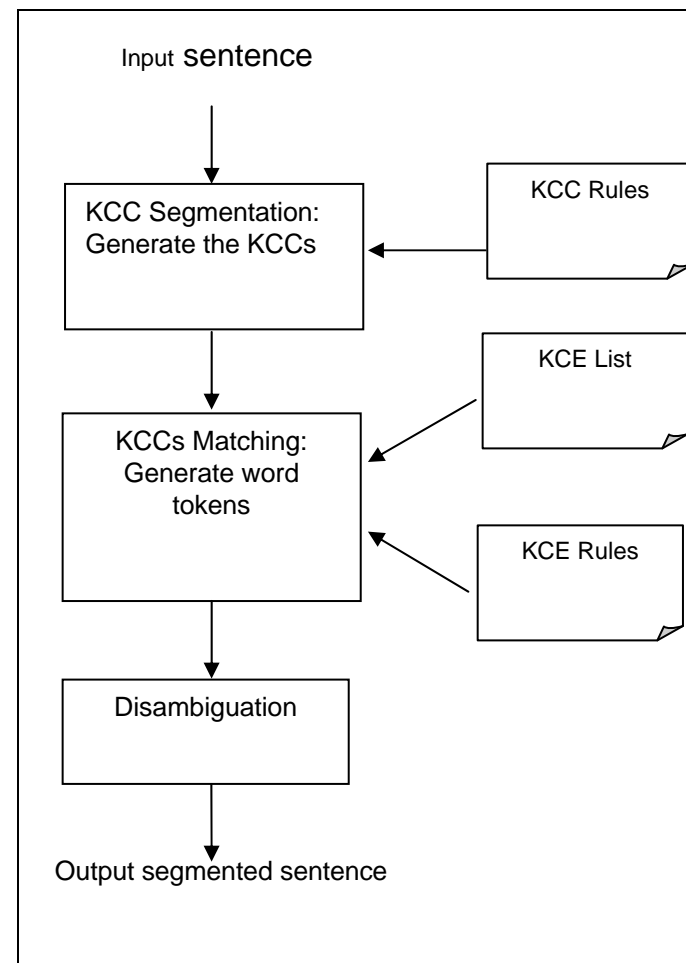


# Word segmentation



## → Word segmentation results:

- Khmer Consonants Cluster (KCC)
- Khmer Common Expression (KCE)
  - Matching rules
  - Mapping rules





# Khmer Consonant Cluster (KCC)



→ The word

→ Three KCCs:        +        +

→ The word

→ Four KCCs:        +        +        +

→ The word

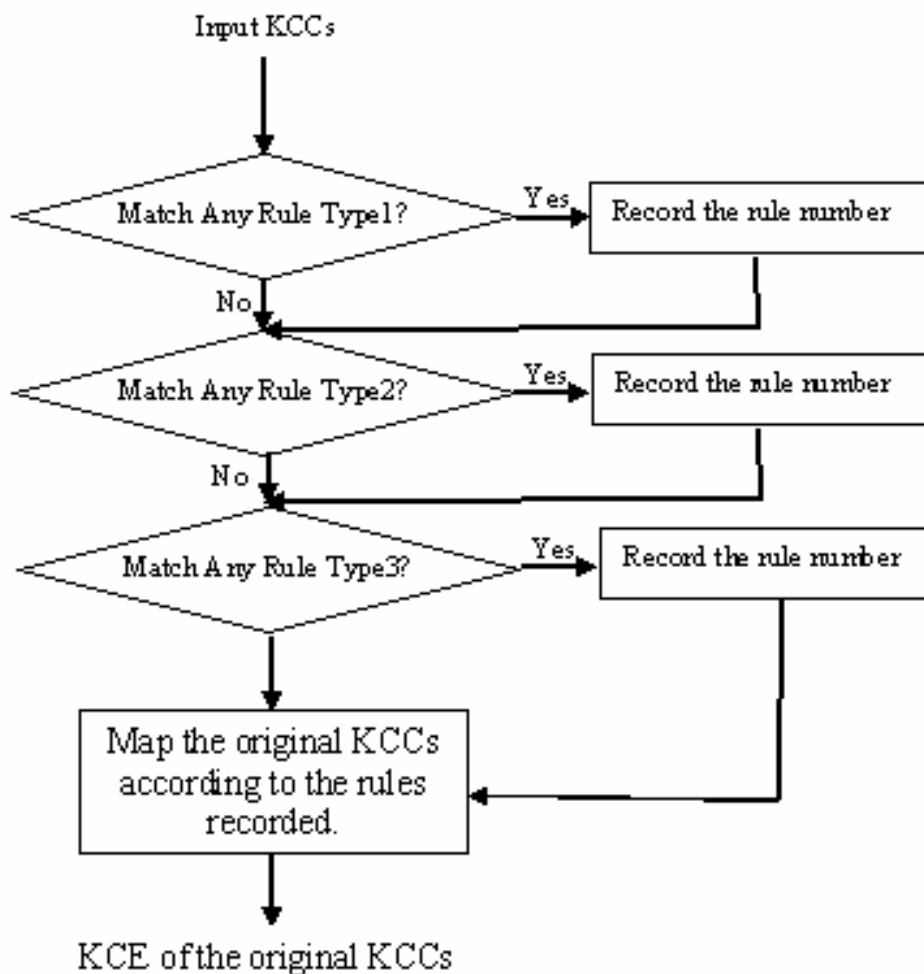
→ One KCC:

<C | I> + [<Robat | Regshift>] + {COEUNG + <C + [Regshift] | I + [Regshift]>} + [[<ZWJ | ZWNJ>] + V] + {S} + [ZWJ + COEUNG + <C | I>]



# Khmer Common Expression

(KCE)





---

# Experimental Results



- Total number of words in the test set =8956
- Total Number of homophonous misspelling=436
- Number of error correctly detected = 403
- The recognition rate is : 92.43%



---

# Conclusion



- Khmer spell checker ease the correction task in document writing
- Khmer Spell Checker is a crucial key for the development of other NLP applications such as Information Retrieval, machine translation, OCR ...etc