

Language Processing Applications Bangla Spelling Checkers

Md. Abul Hasnat

Center for Research on Bangla Language Processing

BRAC University



Alphabets of Bangla script

Consonant	IPA
ক	/kɔ/
খ	/kʰɔ/
গ	/gɔ/
ঘ	/gʰɔ/

Vowel	Vowel sign with KA (ক)	IPA
অ	ক (none)	/kɔ/ and ko
আ	ক া = কা	ka
ই	ক ি = কি	ki
উ	ক ু = কু	ku

Consonant Cluster	Constituents
ক্ষ	ক + ্ + ষ
ঞ্ঞ	ঞ + ্ + চ
জ্ঞ	জ + ্ + ঞ
ল্ম	ল + ্ + ম

Vowel	11
Consonant	49
Consonant Cluster	More than 200



Outline

- Overview
- Challenges
- Phonetic encoding
- Implementation
- Conclusion



Overview: Bangla Spelling Checker

- Pure Java library to provide spell checking service for Bangla.
- Integrated with Jazzy (set of APIs that allow you to add spell checking functionality)
- Provides Web service API.
- Check spelling in Bangla Unicode (utf-8) documents.
- Used in other applications
 - Automated speech recognition
 - OCR post processor
 - email client
- Uses the combination of edit-distance and phonetic encoding algorithm for Bangla.



Challenges for Bangla spelling checker

- Complex orthographic rules
- Large gap between script and pronunciation in Bangla



Phonetic Encoding

- Encodes a word based on its pronunciation.
- Similar sounding words have same code.



Example

Dictionary Word List	Encoded Word List
অকালপক্ক /ɔkalpɔkko/	“okalpkk”
সকাল /ʃɔkal/	“skal”
পাষণ /paʃan/	“pasan”
দগ্ধ /dɔgdʱo/	“dgd”

Encoded Test word	Test Word
“skal”	শকাল /ʃɔkal/

Search the encoded misspelled word
 in the encoded word list rather
 than searching the misspelled word in
 the Dictionary word list



Phonetic Encoding

- Established phonetic encoding in English:
 - Soundex
 - Metaphone
 - Phonix
 - Double metaphone
- UzZaman and Khan's Bangla phonetic encoding:
 - Soundex (2004) and
 - Double Metaphone (2005).



Proposed Phonetic Encoding

- Double Metaphone phonetic encoding
- No of transformation: 108
- Includes all vowels, consonants, consonant clusters (named jukhtakhor in Bangla)



Sample Encoding Rules for

ক্ষ

**Soundex
Encoding**

ক	\u0995	“k”
্	\u0981	0 (zero)
ষ	\u09B7	"s"

Double Metaphone Encoding

ক্ষ	\u0995\u09CD\u09B7	“k”	@the beginning	ক্ষত
ক্ষ	\u0995\u09CD\u09B7	“kk”	@ middle/end	দক্ষ

Ranking the suggestion

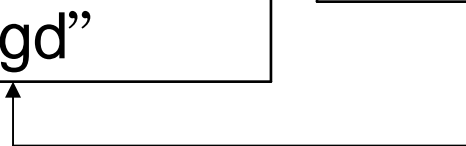
- Solution: Edit distance
- Combination of edit distance on phonetic encoding
 - Able to rank the suggestions phonetically and typographically



Example - generate phonetic code

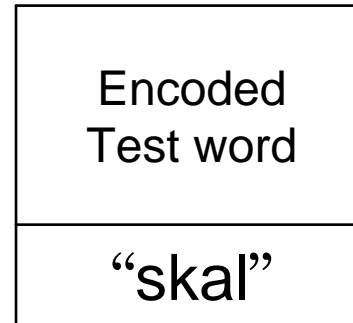
Dictionary Word List	Encoded Word List
অকালপক্ক /ɔkalpɔkko/	“okalpkk”
সকাল /ʃɔkal/	“skal”
পাষণ /paʃan/	“pasan”
দগ্ধ /dɔgd̪hɔ/	“dgd”

Encoded Test word	Test Word
“skal”	শকাল /ʃɔkal/



Example - edit distance with codes

Encoded Word	Edit distance
“okalpkk”	4
“skal”	0
“pasan”	4
“dgd”	4



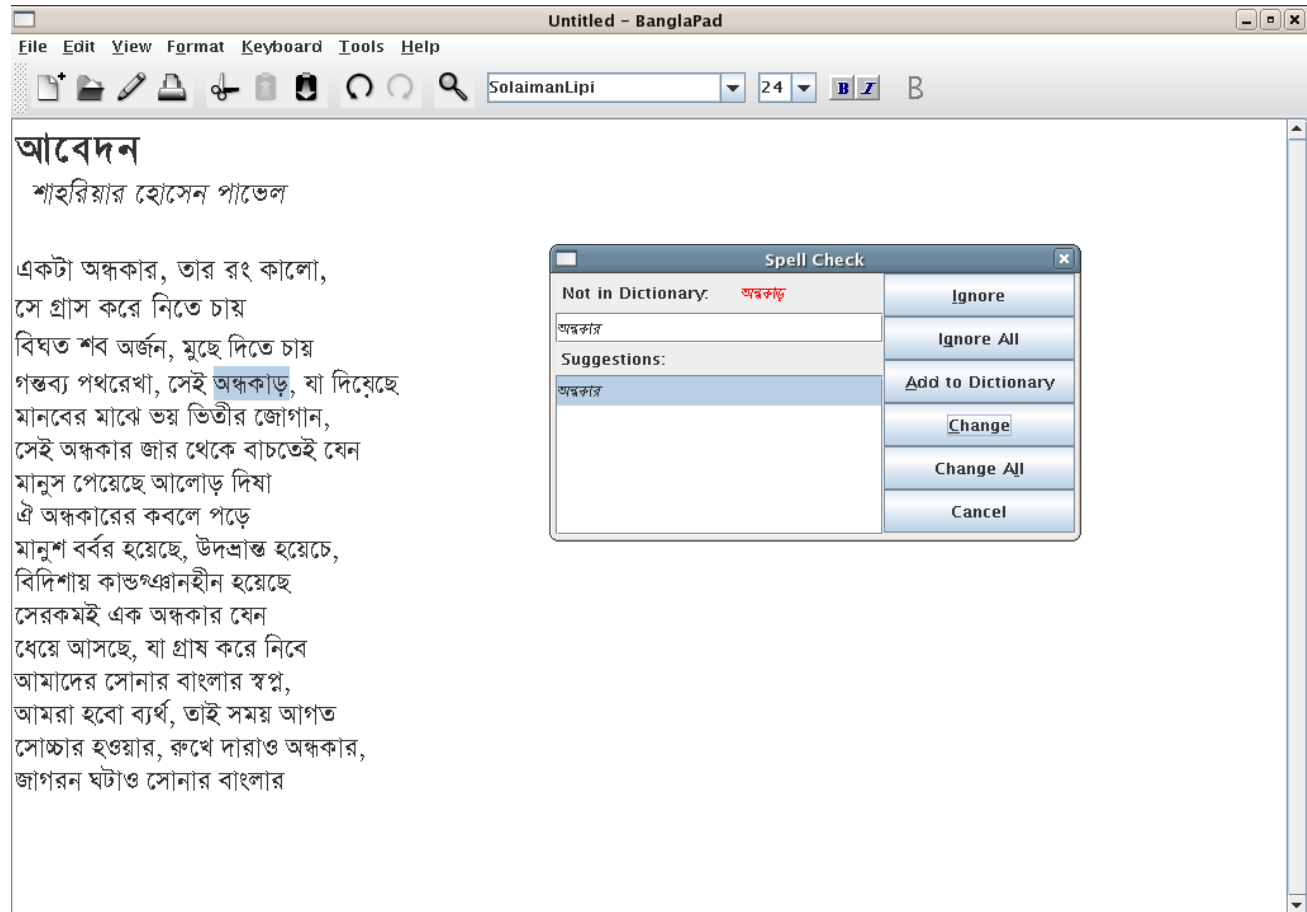
Example - sort according to distance

- Suggestions for শকাল /ʃɔkal/ 𑒧
 1. সকাল /ʃɔkal/
 2. অকালপক্ক /ɔkalpɔkko/
 3. পাষণ /paʃan/
 4. দক্ষ /dɔgdʰo/



Implementation

○ Spellchecker on BanglaPad



Result

- Accuracy rate 91.67%



Conclusion

- Overview of Bangla Spellchecker.
- Usage of the phonetic encoding.
- Encoding rules.
- Suggestion ranking.
- Implementation & Result.



Thank You.



Edit distance

- **Definition:** The smallest number of insertions, deletions, and substitutions required to change one *string* into another.

		k	i	t	t	e	n
	0	1	2	3	4	5	6
s	1	1	2	3	4	5	6
i	2	2	1	2	3	4	5
t	3	3	2	1	2	3	4
t	4	4	3	2	1	2	3
i	5	5	4	3	2	2	3
n	6	6	5	4	3	3	2
g	7	7	6	5	4	4	3

		S	a	t	u	r	d	a	y
	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
u	2	1	1	2	2	3	4	5	6
n	3	2	2	2	3	3	4	5	6
d	4	3	3	3	3	4	3	4	5
a	5	4	3	4	4	4	4	3	4
y	6	5	4	4	5	5	5	4	3

