

# A Double Metaphone Encoding for Bangla and its Application in Spelling Checker

Naushad UzZaman and Mumit Khan

Center for Research on Bangla Language Processing

BRAC University, Bangladesh

**2005 IEEE International Conference on Natural  
Language Processing and Knowledge Engineering**

**Oct 31, 2005**

Donghu Hotel, Wuhan, China

# [ Topics to be covered ]

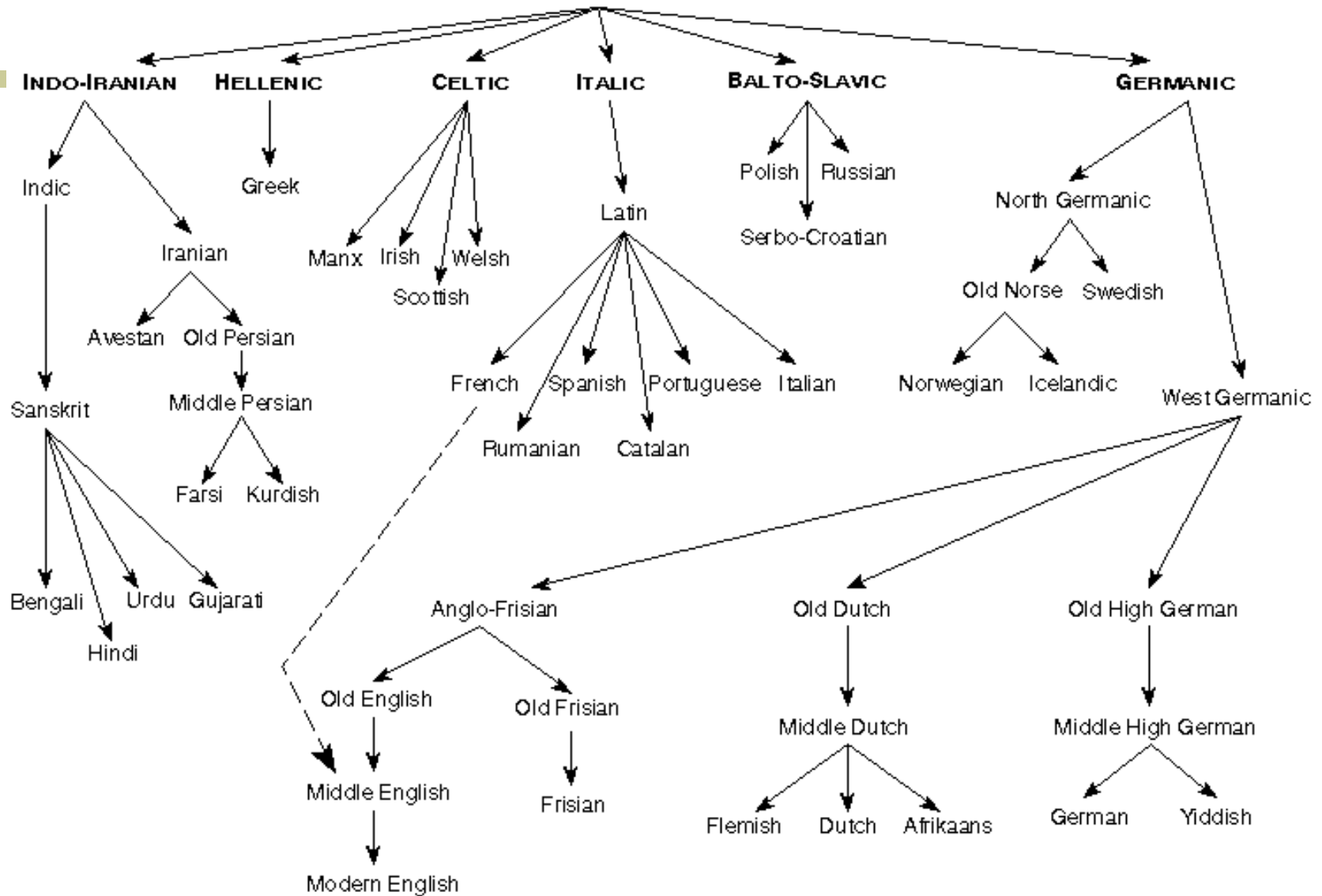
- About Bangla / Bengali language
- Motivation for phonetic encoding
- Phonetic encoding
- Performance of phonetic encoding in spelling checker
- Conclusion

# Background of Bengali / Bangla

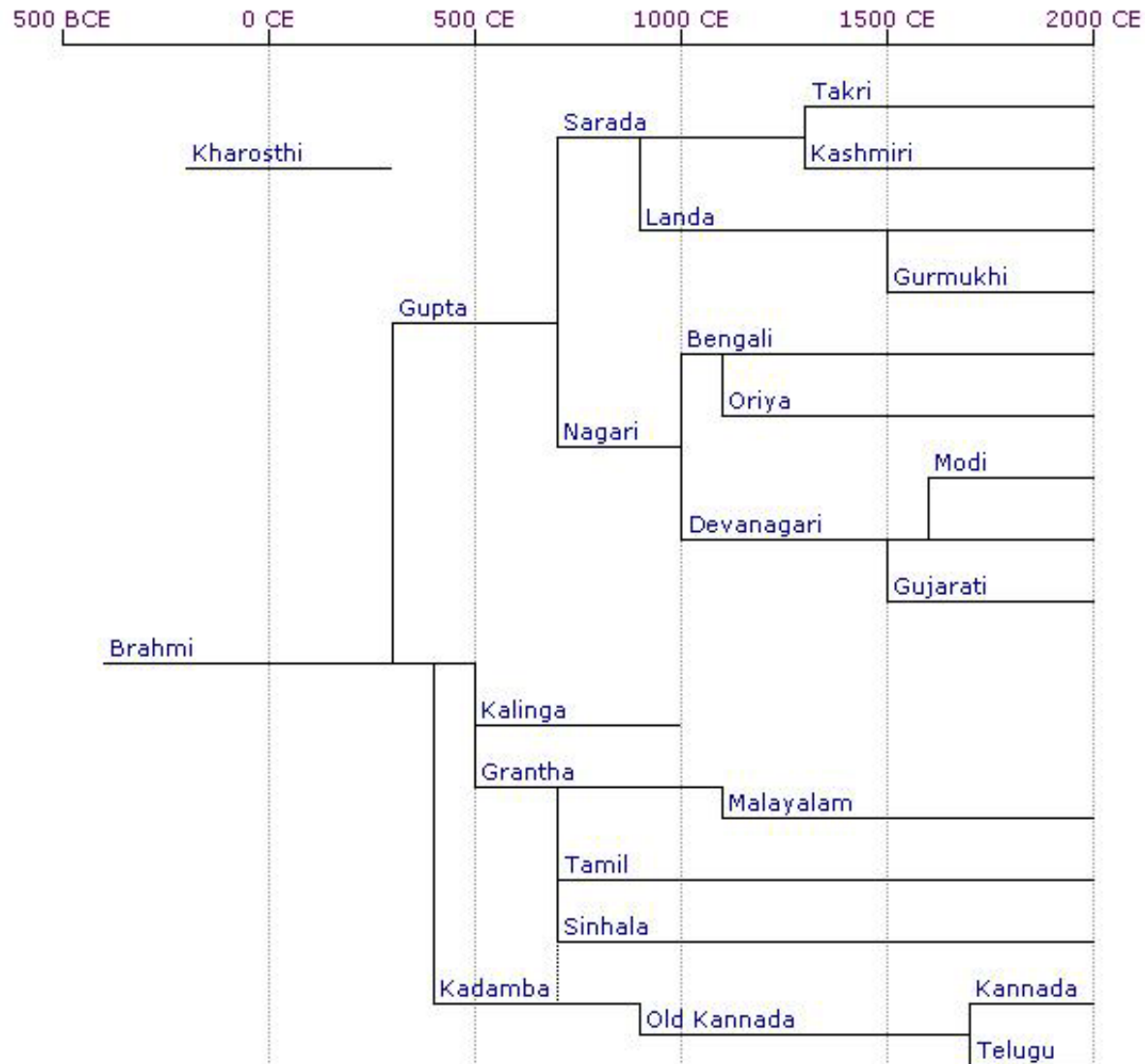
- Spoken mainly in
  - Bangladesh, Indian states of West Bengal, Tripura, Assam
- Native speakers
  - More than 200 million
  - 4th most widely spoken native language by Ethnologue survey
- Bengali/Bangla
  - *Bengali* is the exonym
  - *Bangla* (বাংলা) is the ethnonym

# Generic Classification of Bangla language

## Proto-Indo-European



# Generic Classification of Bangla script



# Example 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
ঞ্চ	ঞ+্+চ
জ্ঞ	জ+্+ঞ
ল্ম	ল+্+ম

Vowel	11
Consonant	49
Consonant Cluster	More than 200

# [ Motivation ]

---

- 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 of Spell Checking Using Encoding

Dictionary	Encoded		
Word List	Word List		
অকালপক্ক /ɔkalpɔkko/	“okalpkk”		
সকাল /ʃɔkal/	“skal”		
পাষণ /paʃan/	“pasan”		
দগ্ধ /dɔgd <sup>ho</sup> /	“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 in English ]

- Established phonetic encoding in English:
  - Soundex
  - Metaphone
  - Phonix
  - Double metaphone

# Key concepts from English phonetic encoding

- Soundex: groups the letter of similar pronunciation and give them same code
  - Real**ize** – 6004020 – 6420
  - Real**ise** – 6004020 – 6420
- Metaphone & Phonix: also considers the context of a letter to encode it
  - **Knigh**t – NT
  - **Nite** – NT

# Key concepts from English phonetic encoding

- Double metaphone: gives multiple codes to same word, if it is pronounced in more than two ways
  - Basinger is pronounced in both way as “Basin-gger” or “Basin-ger”
  - Basinger - BSNJR
  - Basin-gger - BSNKR
  - Basin-ger - BSNJR

# [ Existing Encoding in Bangla ]

- Hoque and Kaykobad's encoding, 2002
- UzZaman and Khan's encoding, 2004

# Hoque and Kaykobad's phonetic encoding Table

Name	Group Member
1	ক, খ, গ, ঘ, ঙ্গ
2	চ, ছ, জ, ঝ, য
3	ট, ঠ, ড, ঢ
4	ত, থ, দ, ধ, ত্
5	প, ফ, ব, ভ
6	ঙ, ঞ, ং
7	শ, স, ষ
8	র, ড়, ঢ়, ঝ
9	ন, ণ
α	ম
β	ল

- For example, কৰ্ম /kɔ̃rmo/ will be converted to a 4 element code “α8a0”, with zero padding.

# UzZaman and Khan's encoding, 2004

Code	Group members
0	্, ো, ঁ
“a”	আ, া
“i”	ই, ঈ, ি, িী
“u”	উ, ঊ, ু, ূ
“e”	এ, ে, ঐ, ঐে
“o”	অ, ও, ঔ, ৌ
“k”	ক, খ
“g”	গ, ঘ
“m”	ম, ঙ, ং
“c”	চ, ছ
“j”	য, জ, ঝ

# Example of UzZaman and Khan's soundex

Misspelled word	Correct word	Encoding
খুমাড় /k <sup>h</sup> umar/	কুমার /kumar/	kumar
পাসান /paʃan/	পাষণ /paʃan/	pasan
দগধ /dɔgd <sup>h</sup> o/	দক্ষ (দগ ্ধ) /dɔgd <sup>h</sup> o/	dgd

# Limitation of existing encodings

- Different pronunciation of constituents in consonant cluster context.
- let us consider ক্ষ .
  - ক্ষ = ক /kɔ/ + ঞ্ + ষ /ʃɔ/;
  - ক্ষ pronounced as /k<sup>h</sup>/
  - ক্ষত /k<sup>h</sup>ɔt̪o/ is pronounced as খত /k<sup>h</sup>ɔt̪o/, where ষ /ʃɔ/ is silent

# Limitation of existing encodings

- Different pronunciation of letters or consonant clusters in different contexts: consider again ক্ষ .
  - At the beginning of a word
    - (ক্ষত → খত /k<sup>h</sup>ɔt̪o/);
  - In the middle or at the end of a word
    - (দক্ষ → দকখ /dɔkk<sup>h</sup>o/).
- Multiple pronunciations of some letters in the same context, such as হ with ব:
  - আহ্বান → আওভান /aovan/.
  - আহ্বান → আহভান /ahob<sup>h</sup>an/

# [ 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 ক

## Soundex Encoding

"k"	ক	KA	\u0995
0 (zero)	্	Virama/Hasant	\u0981
"s"	ষ	SSA	\u09B7

## Double Metaphone Encoding

ক	\u0995\u09CD\u09B7	"k"	@the beginning	কত
ক	\u0995\u09CD\u09B7	"kk"	@ middle/end	দক

# Performance in spelling checker

No of words	1607*
Correct (Edit Distance 0)	1473
Error	134
Rate of accuracy	91.67%
Rate of error	8.33%

\*Source of words: Bangla Banan Obhidhan, Dr. Khurshid Alam, Mirnava, Dhaka, Bangladesh.

# Performance in spelling checker

No of words	1607
Correct (Edit Distance 0)	1473
Error	134
Rate of accuracy	91.67%
Rate of error	8.33%

Error	134	8.33%
Edit Distance 1	107	6.65%
Edit Distance 2	27	1.68%

# Summary and Conclusion

- Proposed a double metaphone phonetic encoding for Bangla
- Handles the complexity of Bangla orthographical rules
- Used the encoding in Spelling checker
- 92% accuracy in spelling checker with just the phonetic encoding
- For our particular sample set we get a 100% accuracy by using phonetic encoding and 2 edit distance.

# [ Question ]

---

- ?

# The Summer Institute for Linguistics (SIL) Ethnologue Survey (1999)

- **(SIL) Ethnologue Survey (1999)** lists the following as the top languages by population: □ number of native speakers in parentheses
  - **1.Chinese\*** (937,132,000)
  - **2.Spanish** (332,000,000)
  - **3.English** (322,000,000)
  - **4.Bengali** (189,000,000)
  - **5.Hindi/Urdu** (182,000,000)
  - **6.Arabic\*** (174,950,000)
  - **7.Portuguese** (170,000,000)
  - **8.Russian** (170,000,000)
  - **9.Japanese** (125,000,000)
  - **10.German** (98,000,000)
  - **11.French\*** (79,572,000)

# Dr. Bernard Comrie's article for the *Encarta Encyclopedia* (1998)

- The following list is from **Dr. Bernard Comrie's** article for the *Encarta Encyclopedia* (1998): (number of native speakers in parentheses)
  - **1.Mandarin Chinese** (836 million)
  - **2.Hindi** (333 million)
  - **3.Spanish** (332 million)
  - **4.English** (322 million)
  - **5.Bengali** (189 million)
  - **6.Arabic** (186 million)
  - **7.Russian** (170 million)
  - **8.Portuguese** (170 million)
  - **9.Japanese** (125 million)
  - **10.German** (98 million)
  - **11.French** (72 million)

<http://www.aneki.com/languages.html>

Source: University of Washington

Rank	Language	No of Speaker
1	Chinese (Mandarin)	1,000,000,000 +
2	English	508,000,000
3	Hindustani (Hindi and Urdu)	497,000,000
4	Spanish	392,000,000
5	Russian	277,000,000
6	Arabic	246,000,000
7	Bengali	211,000,000
8	Portuguese	191,000,000
9	Malay-Indonesian	159,000,000
10	French	129,000,000

# Found in 15,162,317 words

Error zone length (in no. of char.)	% of word
1	41.36
2	32.94
3	16.58
4	7.10
5	1.78
6	0.24

B. B. Chaudhuri, “Reversed word dictionary and phonetically similar word grouping based spell-checker to Bangla text”, *Proc. LESAL Workshop*, Mumbai, 2001.