
A Double Metaphone Encoding for Approximate Name Searching and Matching in Bangla

Naushad UzZaman and Mumit Khan
Center for Research on Bangla Language
Processing
BRAC University, Bangladesh

**The Fourth IASTED International
Conference on COMPUTATIONAL
INTELLIGENCE**

**July 4, 2005
Calgary, Alberta, Canada**

Topics to be covered

- Motivation for name searching
- Name searching in English
- Phonetic encoding
- Background of Bangla
- Challenges in Bangla name searching
- Name searching in Bangla
- Proposed phonetic encoding for Bangla
- Application to name searching
- Ranking suggestions
- Conclusion

Motivation for name searching

- Applications
 - Land registry
 - Census
 - Educational institutes
 - Criminal record search
 - Health sector
 - Industries
 - etc

Name searching in English

- Solution ?
 - Phonetic encoding
 - Approximate string matching algorithm
 - Levenshtein edit distance
 - Longest common subsequences
 - Etc..

Phonetic encoding

- Encodes a word or name based on how it is pronounced
- Same names have the same phonetic code
- Search the codes, not the names

Phonetic encoding in English

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

Key concepts from English phonetic encodings

- Soundex: groups the letter of same pronunciation and give them same code
 - Brian - 16005 - 165
 - Bryan - 16005 - 165
- Metaphone & Phonix: also considers the context of a letter to encode it
 - **Knight** – NT
 - Nite – NT

Key concepts from English phonetic encodings...

- 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

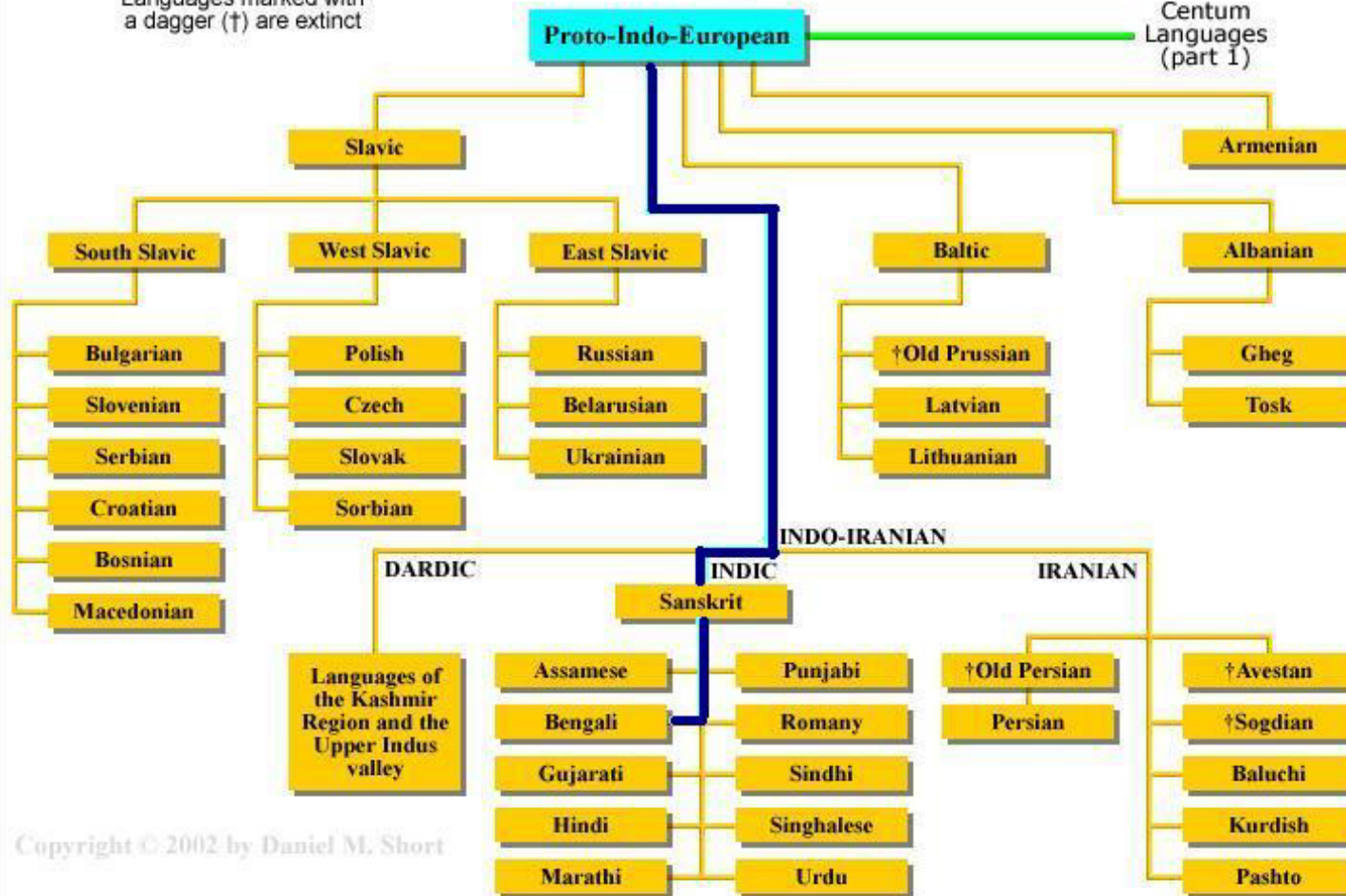
Background of Bangla / Bengali

Indo-European Language Tree

Part 2: Satem Languages

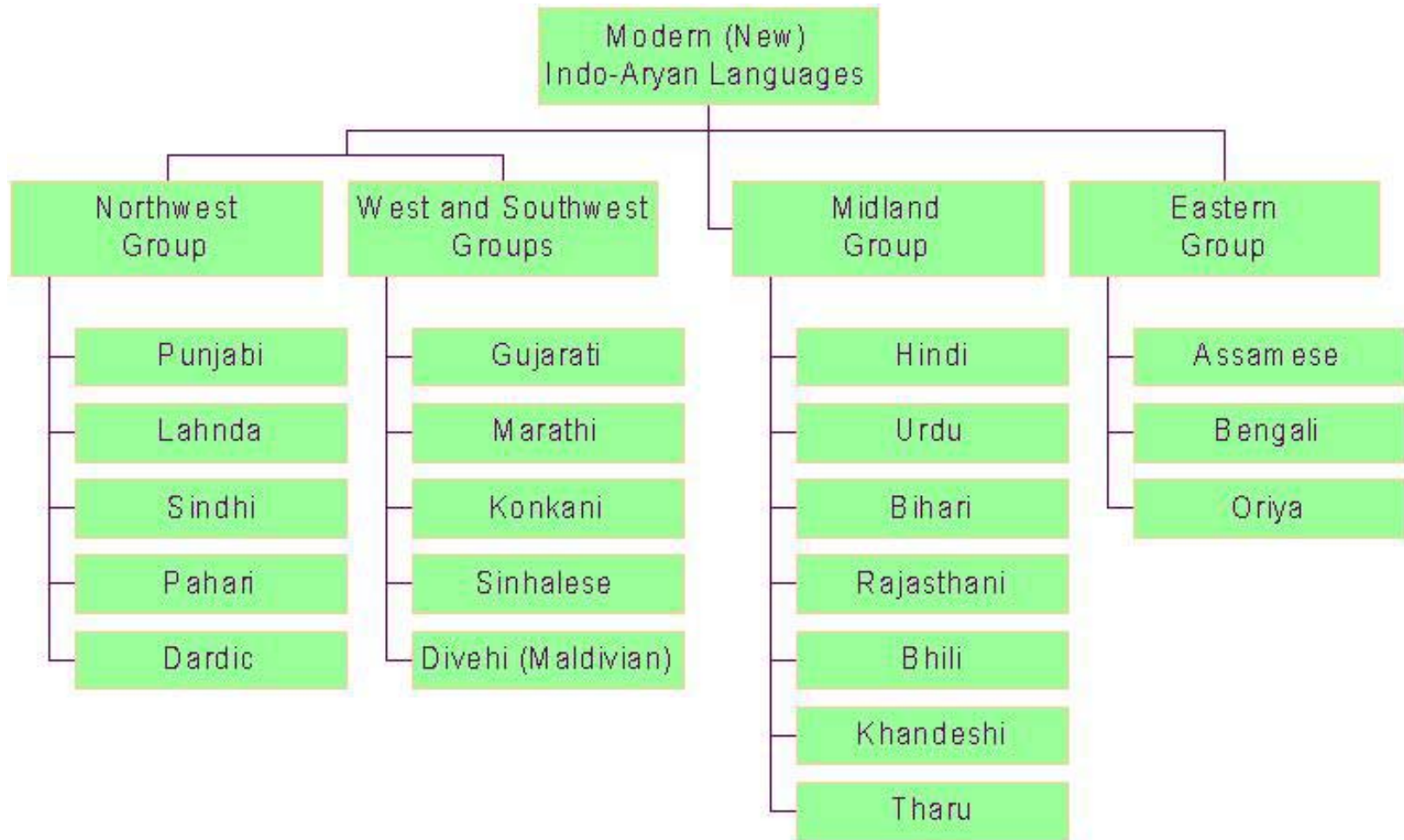
Languages marked with a dagger (†) are extinct

Centum Languages (part 1)



Copyright © 2002 by Daniel M. Short

Background of Bangla / Bengali



Background of Bangla / Bengali

- More than 200 million people speaking in Bangla, 4th most widely spoken language in the world
- Native language of Bangladesh, Indian state of West Bengal.
- Significant Bangla speaking community in the Indian state of Assam and Tripura.

Challenges in Bangla name searching

- Any word can be a name (*complex orthographic rules, large gap between script and pronunciation in Bangla*)
- Different origins of names (*significant changes in both spelling and pronunciation from original as it evolves*)
 - ❑ Sanskrit
 - ❑ Perso-arabic languages
 - ❑ Portuguese and other western languages

Challenges for Bangla words

- Bangla has many consonant clusters or *juktakkhor* with unusual pronunciations (i.e., ক্ষ, ক্ষা, etc.)
 - ক্ষ = ক /kɔ/ + ্ + ষ /ʃɔ/; ক্ষত /kʰɔ̃tɔ/ is pronounced as খত /kʰɔ̃tɔ/, where ষ /ʃ/ does not have any sound.
- Different pronunciation of letters or conjuncts in different contexts: consider again ক্ষ.
 - At the beginning of word /kʰ/
 - (ক্ষত → খত /kʰɔ̃tɔ/)
 - In the middle or at the end of a word /kkʰ/
 - (দক্ষ → দকখ /dɔkkʰo/)

Challenges for Bangla words

- Multiple pronunciations of some letters in the same context, such as শ /s ~ ʃ/ in প্রশ্ন
 - প্রশ্ন /prosno/
 - প্রশ্ন /proʃno/

Different manifestation of imported names

- মোহাম্মদ /mohammɔd/ from Arabic
- We use this name as
 - মোহাম্মদ /mohammɔd/
 - মুহাম্মদ /muhammɔd/
 - মোহাম্মেদ /mohammed/
 - মুহাম্মেদ /muhammed/
 - মোহাম্মাদ /mohammad/
 - মুহাম্মাদ /muhammad/

Proposed phonetic encoding for Bangla

- Double metaphone phonetic encoding for Bangla
- No of transformations: 108
- Includes all vowels, consonants, consonant clusters (called *Juktakkhor* in Bangla)

Sample Encoding Rules for য /j/, জ/j/ and ঝ/j^h/

Soundex Encoding

“j”	য	YA	“09AF”
	জ	JA	“099C”
	ঝ	JHA	“099D”

Double Metaphone Encoding

য	YA as fola	x”09CD”“09AF”	“e”	@ the beginning as YA fola	ব্যথিত, ব্যক্ত, ন্যস্ত
		...xy”09CD”z”09CD”“09AF”	Not Coded	@ middle/end with jukhtakhor	সন্ধ্যা, মর্ত্য
		...xy”09CD”“09AF”	Doubles: yy	@ middle/end	অদ্য, মধ্য
য	YA	“09AF”	“j”		
জ	JA	“099C”	“j”		
ঝ	JHA	“099D”	“j”		

Encoding examples...

Abbr eviation	Elaborated form	Encoding
মোঃ	মোহাম্মদ /mohammɔd/	“mmmD”
ডঃ	ডক্টর /dɔktor/	“DkTr”
ডাঃ	ডাক্তার /daktar/	“DkTr”
এডঃ	এডভোকেট /advokæt/	“DbkT”

- মোঃ is the same as মোহাম্মদ /mohammɔd/
- one-to-one transformations are used before encoding process
- So, to encode মোঃ we will first transform it to মোহাম্মদ before the final encoding

Application to name searching

Name list	Encoded name list
মরতুজা /mɔɾtuʒa/	“mrtj”
নাইম /naim/	“nm”
নাহলীন /nahleen/	“nl̩n”
পুষ্প /pushpɔ/	“psp”

Encoded query name	Query name
“mrtj”	মুরতোজা /murtoʒa/



Ranking the suggestions

- Need to consider
 - Edit distance between codes
 - Edit distance between names
 - Considering both generate a score
 - Rank the suggestion using the score

Algorithm for name searching

1. Encode the name to search for: মর্তুজা /mɔrtuʒa/ → mrtj
2. Compute the phonetic edit-distance, using the encoded versions
3. Compute the phonetic edit distance score from PED:
$$\text{PEDscr} = (\text{maxLen}(s1, s2) - \text{ED}) / \text{maxLen}(s1, s2)$$
4. Compute the edit-distance between the candidate name and each of the names from list
5. Compute the edit distance score between the two strings s1 and s2 from ED:
$$\text{EDscr} = (\text{maxLen}(s1, s2) - \text{ED}) / \text{maxLen}(s1, s2)$$
6. The figure of merit (FOM) is the weighted sum of PEDscr and Edscr, with PEDscr as the dominant factor:
$$(\text{PEDscr} + \text{Edscr}/10) / 1.1$$
 and value ranges from 0 to 1

Generate suggestions for name searching

Names	Encoding	ED	EDscr	PED	PEDscr	FOM
রশিদ /rɔʃid/	"rsd"	5	0.167	4	0	0.02
মুকসিত /mukʃit/	"mkst"	5	0.167	3	.25	0.24
মরতুজা /mɔrtuʃa/	"mrtj"	0	1	0	1	1
মুরতোজা /murtoʃa/	"mrtj"	2	0.714	0	1	0.97
মরতোজা /mɔrtoʃa/	"mrtj"	1	0.833	0	1	0.98
মোরতুজা /mortuʃa/	"mrtj"	1	0.857	0	1	0.99

Final suggestion for মরতুজা /mɔɾtuʒa/

মরতুজা /mɔɾtuʒa/

মোরতুজা /mɔɾtuʒa/

মরতোজা /mɔɾtoʒa/

মুরতোজা /murtoʒa/

মুকসিত /mukʃit/

রশিদ /ɾɔʃid/

Conclusion

- We proposed a phonetic encoding that encodes a Bangla name based on its pronunciation
- Used the phonetic encoding in name searching application
- Used edit distance to rank the suggestion

Questions?

Levenshtein Edit distance

- The edit distance of two strings, s_1 and s_2 , is defined as the minimum number of point mutations required to change s_1 into s_2 , where a point mutation is one of:
 - Replace a letter,
 - Insert a letter,
 - Delete a letter,
 - Transpose consecutive letters

Example of Edit distance

- $e(\text{"Virginia"}, \text{"Vermont"}) = 5$
- Virginia
- Verginia
- Verminia
- Vermonia
- Vermonta
- Vermont

Soundex table

<i>Code</i>	<i>Letters</i>
0 (not coded)	A, E, I, O, U, H, W, Y
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Metaphone transformation

- B -> B unless at the end of a word after "m" as in "dumb"
- C -> X (sh) if -cia- or -ch-
- S if -ci-, -ce- or -cy-
- K otherwise, including -sch-
- D -> J if in -dge-, -dgy- or -dgi-
- T otherwise
- F -> F
- G -> **silent if in -gh-** and not at end or before a vowel
- in -gn- or -gned- (also see dge etc. above)
- J if before i or e or y if not double gg
- K otherwise
- H -> silent if after vowel and no vowel follows
- H otherwise
- J -> J
- K -> silent if after "c"
- K otherwise
- L -> L
- M -> M
- N -> N

- P -> F if before "h"
- P otherwise
- Q -> K
- R -> R
- S -> X (sh) if before "h" or in -sio- or -sia-
- S otherwise
- T -> X (sh) if -tia- or -tio-
- 0 (th) if before "h"
- silent if in -tch-
- T otherwise
- V -> F
- W -> silent if not followed by a vowel
- W if followed by a vowel
- X -> KS
- Y -> silent if not followed by a vowel
- Y if followed by a vowel
- Z -> S
-
- Initial Letter Exceptions
-
- Initial **kn-**, gn- pn, ac- or wr- -> **drop first letter**
- Initial x- -> change to "s"
- Initial wh- -> change to "w"

Sample Encoding Rules for ক্ৰ

Soundex Encoding

"k"	ক	KA	"0995"
0 (zero)	্	Virama/Hasant	"0981"
"s"	ষ	SSA	"09B7"

Double Metaphone Encoding

ক্ৰ	"0995""09CD""09B7"	"k"	@the beginning	ক্ৰত
ক্ৰ	"0995""09CD""09B7"	"kk"	@ middle/end	দক্ৰ

Bangla / Bengali

- Bangla is the ethnonym, our name for our language
- Bengali is the exonym, the name in English for our language