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

Naushad UzZaman
Center for Research on Bangla Language Processing
BRAC University
Dhaka, Bangladesh
naushad@bracuniversity.net

Mumit Khan
Center for Research on Bangla Language Processing
BRAC University
Dhaka, Bangladesh
mumit@bracuniversity.net

***Abstract*- We present a Double Metaphone encoding for Bangla that can be used by spelling checkers to improve the quality of suggestions for misspelled words. The complex rules of Bangla spelling present a significant challenge in producing suggestions for a misspelled word when employing the traditional edit-distance methods; one must take phonetic similarity into account for the suggested alternatives to be reasonably accurate. We propose a Double Metaphone encoding for Bangla, taking into account the various context-sensitive rules, including those involving the large repertoire of consonant clusters in Bangla, and present a comparison with the traditional edit-distance based methods in producing suggestions for misspelled words.**

Keywords
Bangla, Bengali, Phonetic Encoding, Double Metaphone, Spelling suggestions, Spelling Checker.

I. INTRODUCTION

Bangla, also known as Bengali, is the language of approximately 189 million native speaker, the majority of whom live in Bangladesh and the Indian state of West Bengal, making it the 4th most widely spoken language in the world [1]. It belongs to the eastern group of Indo-Aryan languages, and is written in the Brahmi derived Bangla script. Bangla underwent a period of vigorous Sanskritization that started in the 12th century and continued throughout the middle ages, resulting in the vast gap between the script and the pronunciation [2]. Bangla lexicon today consists of *tatsama* (Sanskrit words that have changed pronunciation, but retained the original spelling), *tadbhava* (Sanskrit words that have changed at least twice in the process of becoming Bangla), and a fairly large number of "loan-words" from Persian, Arabic, Portugese, English, and other languages. There are also a large number of words of unknown etymology, which may have originated from Dravidian, Austric or Sino-Tibetan languages. All of these contribute to the complexity of the Bangla spelling rules, with the Sanskritization process as the largest contributor. An additional factor is the large number of consonant clusters or *juktakkhors* in Bangla. One impact of this complexity can be seen in the observation that two of the most common reasons for misspelling are (i) phonetic similarity of Bangla characters, and (ii) the difference between the grapheme representation and the phonetic utterances [3]. Methods based on traditional edit-distance algorithms are not able to produce "good" suggestions for misspelled Bangla words unless phonetic similarity is taken into account. While there has been significant research into "fuzzy" string matching algorithms for English and other Western languages [4-7], similar work for Bangla has barely begun [9-13]. These efforts are mostly based on Soundex or other *ad-hoc* methods, which cannot handle the complexity of Bangla spelling rules. This is the primary motivation for creating a Double Metaphone encoding that can handle such complexity.

We begin by examining some of the complex orthographic rules in Bangla in Section II, which illustrate the challenges in creating an effective phonetic encoding for it. We then describe the proposed Double Metaphone encoding in Section III, along with the rationale for the mapping rules. In Section IV, we then demonstrate the use of this phonetic encoding in a spelling checker, and provide some comparisons with spelling checkers that use other phonetic encodings such as Soundex or only use traditional edit distance methods.

II. CHALLENGES IN CREATING A PHONETIC ENCODING FOR BANGLA

The complex orthographic rules in Bangla pose a challenge when creating a phonetic encoding for it. Some of the common cases illustrating these spelling rules, ones that a candidate encoding must be able to handle, are shown below:

1. There are groups of phonetically similar characters in Bangla; for example, NA (ন) and NNA (ণ), SA (শ), SHA (স), SSA (ষ) etc. The contrast between long and short vowels in the scripts is also in the modern version of spoken language.
2. Bangla has many consonant clusters or conjuncts with unusual pronunciations (i.e., ক্ষ, ক্ষা, etc.): consider ক্ষ. ক্ষ = ক+্+ষ; ক্ষত /kʰɔt̪o/ is pronounced as খত /kʰɔt̪o/[1], where ষ is silent and ক /k/ and ষ /ʃ/ becomes খ /kh/ in initial position.
3. Bangla has different uses of *Phalaa's*[2]*, the cluster for of the semi-vowels in Bangla* (i.e., BA *phalaa*, MA *phalaa*, YA *phalaa*, RA *phalaa*, LA *phalaa*), which are represented using the distinct sign form. BA phalaa for example has a distinct pronunciation from a BA in any other position in a cluster or in a standalone configuration.

[1] Pronunciation of Bangla words is given in IPA (International Phonetic Alphabet). IPA of each word is kept inside slashes, e.g. /bengali/

[2] A phalaa is an allograph, which originally denoted contracted forms of consonants. However, in Bangla the pronunciation has changed to a great extent, whereby the phalaa doubles as an allograph and a diacritic or may be silent altogether.

4. Different pronunciation of letters or conjuncts in different contexts: consider again ক্ষ. In the initial position of a word of word, it is pronounced as খ (ক্ষত → খত); in the middle or at the end of a word, it is pronounced as কখ /kk^h/, দক্ষ /d̪okkho/ → দকখ.
5. Multiple pronunciations of some letters in the same context, such as হ /h/ with ব /b/: According to Bangla phonological rules, হ /h/ should be pronounced as ও /o/ or উ /u/ and ব /b/ should be pronounced as /v/**:** আহ্বান → আওভান /aovan/. However, most native speakers pronounce these words the same way as it is written. For example, আহ্বান is usually pronounced as আওভান /ahobhan/. Both pronunciations are considered correct.

Previous efforts in creating phonetic encoding for Bangla are based on the Soundex algorithm [4]. Soundex partitions the letters into disjoint sets, assuming the letters within the same set have similar sound. It works on a letter-by-letter basis, and cannot handle context-sensitive rules, such as those illustrated above. A recently published encoding for Bangla [9] based on Soundex is able to handle most of the trivial cases, and those involving some of the conjuncts, but it falls far short of producing suggestions for a large majority of the complex misspelled words. Metaphone encoding [5] does consider the context, so it is able to handle all but the last case above, which requires that the encoding be able to produce multiple encoded forms of the same character sequence. Double Metaphone [6] remedies that problem of Metaphone of not being able to produce multiple encoding from same string. These limitations in part led us to create a Double Metaphone encoding for Bangla that does not suffer from the problems listed above, and in addition, is able to handle the full complexity of Bangla spelling rules.

III. DOUBLE METAPHONE ENCODING FOR BANGLA

We had proposed a Double Metaphone encoding for Bangla, which is available in [8]. In this paper, we have given the rationale for the mapping rules. When describing the rationale for the various mappings, we omit the cases handled in [9], as those have remained the same. As in [9], we assume that the Bangla text is encoded using Unicode Normalization Form C (NFC) [14].

There are a total of **107 transformations** in our proposed encoding, which includes vowels, consonants, and conjuncts in all-different contexts.

We encode the letters based on how the letters and conjuncts are pronounced in different contexts, based on rules found in [15-17].

For each rationale in the mapping, we give the Bangla letter, followed by the names of that letter used in the Unicode chart [14], then the Unicode code point of that letter, and finally the pronunciation of that letter in IPA.

Examples of our proposed phonetic encoding are kept inside angle braces, e.g. <oi> for /oi/

ঐ AI \u0990 IPA: /oi/
ৈ SIGN AI \u09C8 IPA: /oi/
Code: <oi>
ঔ AU \u0994 IPA: /ou/
ৌ SIGN AU \u09CC IPA: /ou/
Code: <ou>
ক্ষ = ক ◌্ ষ \u0995 \u09CD \u09B7

Case 1: In the initial position of a word of a word, it is pronounced as খ /k^h/. So it is given the same code as খ, which is <k>.

ক্ষত /k^{h}ɔt̪o/ → <k̲t>

Case 2: In the middle or at the end of words, it is similar to কখ /kk^h/, so it is encoded as <kk>.

দক্ষ /d̪okk^{h}o/ → dk̲k̲

Exception: তৎক্ষনাত্ /t̪ɔt̪k^{h}ɔnat̪/ According to its pronunciation it should be encoded as <ttknat>, but it is instead encoded as <ttk̲k̲nat>.

ঙ NGA \u0999 IPA: /ŋ/
ং ANUSVARA \u0982 IPA: /ŋ/

ঙ and ং sounds like ŋ, so it is encoded as <ng>.

বাঙলা /baŋla/ → <ba<u>ng</u>la>

বাংলা /baŋla/ → <ba<u>ng</u>la>

য YA \u09AF IPA: /ɟ/
য as *phalaa*

Case 1: If the য **/ɟ/** phalaa is positioned after the initial consonant and is followed by an আ /a/ or another consonant then it is pronounced as /æ/.

Example: ব্যক্ত /bækt̪o/, ধ্যান /d̪^{h}æn/

If the য **/ɟ/** phalaa is positioned after the initial consonant and is followed by an ই /i/, then it is pronounced as এ /e/.

Example: ব্যক্তি /bekt̪i/

We encode both /æ/ and /e/ as <e>.

ব্যক্ত /bækt̪o/ → <b<u>e</u>kt>

ধ্যান /d̪^{h}æn/ → <d<u>e</u>n>

ব্যক্তি /bekt̪i/ → <b<u>e</u>kti>

Case 2: If the য **/ɟ/** phalaa is attached to a medial consonant cluster it is usually silent, so it is *Not coded.*

সন্ধ্যা /ʃond^{h}ha/ → সন্ধা → <snda>

স্বাস্থ্য /ʃast̪^{h}o/ → সাস্থ → <sast>

Case 3: When attached to medial consonant or in the terminal position the য **/ɟ/** phalaa acts as a diacritic as the letter it attaches to is geminated, so it has the same code as the previous code.

অদ্য /od̪d̪o/ → অদ্দ → <od<u>d</u>>

মধ্য /mɔd̪d̪^{h}o/ → মদ্ধ → <md<u>d</u>>

Case 4: Otherwise when it is used the full form of the phalaa or allograph is the grapheme য **/ɟ/**, hence it is encoded as <j>

ঞ NYA \u099E IPA: /ɲ/

Case 1: Usually in conjuncts if a ঞ **/ɲ/** is added before a চ /c/, ছ /c^{h}/, জ /ɟ/, ঝ **/ɟ^{h}/**, or after a চ /c/ then it is pronounced as ন /n/; in these cases it is encoded as <n>.

Before চ: অঞ্চল /ɔnc͡ɔl/ → <oncl>

Before ছ: বাঞ্ছা /banc^{h}a/ → <banca>

Before জ: অঞ্জলি /ɔnɟoli/ → <onjli>

Before ঝ: যাচ্ঞা /ɟ^{h}ɔnɟ^{h}a/ → <jnja>

After চ: যাচ্ঞা /ɟacna/ → <jacna>

Case 2: If আ-কার /a/ and ই-কার /i/ is added after a ঞ, then it is pronounced as a nasalized /j̃/. মিঞা /miɲa/. However, since in our encoding nasal sounds are *Not Coded*, it is also *Not Coded*.

মিঞা/miɲa/ → <mia>

নাঞি /naɲi/ → <nai>

Case 3: In conjuncts after জ /ɟ/, ঞ /ɲ/ it sounds as গ /g̃/ in the initial position and it geminate /gg̃/ in the medial and final position. Again, if আ (া)-কার is positioned after the জ্ঞ in initial position then আ (া)-কার is pronounced as <æ>, which is encoded as <e>.

In the initial position of a word:

জ্ঞাত /g̃æt̪o/ → <get>

জ্ঞান /g̃æn/ → <gen>

In the medial or final position:

বিজ্ঞান /bigg̃æn/ → <biggan>

বিজ্ঞ /bigg̃o/ → <bigg>

Exception: সংজ্ঞা /ʃɔŋga/ should be encoded as <sngga> but it is instead encoded as <snggga>.

Case 4: Otherwise, if there is a VIRAMA/Hasant after it, then it is simply encoded as <oi>.

নঞ /noɲ/ → <noi>

নঞর্থক /nɔɟɔrt̪^{h}ok/ → <noirtk>

ঋ VOCALIC R \u098B IPA: /ri/

Case 1: In the initial position of a word Vocalic R ঋ /ri/ and Sign Vocalic R ৃ are encoded as <ri>.

ঋতু /rit̪u/ → <ritu>

Case 2: It geminate the sound of the attached letter if it is in the medial or final position. However, since people usually pronounce it as /ri/ in such cases as well, it is encoded as both codes.

বিকৃত /bikkrit̪o/ → <bikkrit>

বিকৃত /bikrit̪o/ → <bikrit>

র RA \u09B0 IPA: /r/

র as *phalaa*

Case 1: If the র **/r/** *phalaa* is positioned after the initial consonant then it is pronounced as /r/ so it is encoded as <r>.

প্রকাশ /prokaʃ/ → <prkas>

প্রনাম /pronam/ → <prnam>

Case 2: When attached to medial consonant or in the terminal position the র **/r/** phalaa acts as a diacritic as the letter it attaches to is geminated, so the code is doubled as well.

But if we consider the pronunciation of these words, it is also pronounced as only র /r/. As a solution, we again encode it both the codes using the Double Metaphone approach.

রাত্রি /rat̪̪ri/ → <rattri>

রাত্রি /rat̪ri/ → <ratri>

ছাত্র /chat̪̪ro/ → <cattr>

ছাত্র /chat̪ro/ → <catr>

Case 3: Otherwise র is encoded as <r>.

ব BA \u09AC IPA: /b/

ব as *phalaa*

Case 1: If the ব **/b/** *phalaa* is positioned after the initial consonant then it doesn't have any sound. So it is *Not Coded*.

স্বাধিকার /ʃad̪^{h}ikar/ → <sadikar>

স্বদেশ /ʃɔd̪eʃ/ → <sdes>

জ্বালা /ɟala/ → <jala>

Case 2: In the medial or final position ব /b/ *phalaa* with ব /b/, ম /m/ and গ /g/ keeps it sound. So it is encoded as <b>.

ব: তিব্বত /t̪ibbɔt̪/ → <tibbt>

সাব্বাশ /ʃabbaʃ/ → <sabbas>

ম: লম্ব /lɔmbo/ → <lmb>

সর্ম্বধনা /ʃɔmbord̪^{h}ona/ → <smbrdna>

গ: দিগ্বিদিক /d̪igbidik/ → <digbidik>

Case 3: If the ব **/b/** phalaa is positioned after the উদ্ at initial position then ব **/b/** keeps it sounds. So it is encoded as <b>.

দ that is derived from উদ্:

উদ্বেগ /ud̪beg/ → <udbeg>

উদ্বোধন /ud̪bod^{h}on/ → <udbdn>

Case 4: If the ব **/b/** phalaa is attached to a medial consonant cluster it is usually silent, so it is *Not coded.*

তত্ত্ব /t̪ɔt̪̪o/ → <ttt>

উজ্জ্বল /uɟɟɔl/ → <ujjl>

উচ্ছ্বাস /ucc^{h}aʃ/ → <uccas>

Case 5: When attached to medial consonant or in the terminal position the ব **/b/** phalaa acts as a diacritic as the letter it attaches to is geminated, so it has the same code as the previous code.

দ্বিত্ব /dit̪̪o/ → <ditt>

বিশ্ব /biʃʃo/ → <biss>

ম MA \u09AE IPA: /m/

ম as *phalaa*

Case 1: If the ম **/m/** *phalaa* is positioned after the initial consonant then it doesn't have any sound. So it is *Not Coded*.

স্মরণ /ʃɔroɳ/ → <srn>

স্মশান /ʃɔʃan/ → <ssan>

Case 2: If the ম /m/ phalaa is attached to a medial consonant cluster it is usually silent, so it is *Not coded.*

সূক্ষ্ম /ʃukkʰõ/ → <sukk>

লক্ষ্মণ /lɔkkʰon/ → <lkkn>

Case 3: In the medial or final position ম *phalaa* with ক /k/, গ /g/, ঙ /ŋ/, ট /t/, ণ /n/, ন /n/, ম /m/, ল /l/, স /s, ʃ/, ষ /ʃ/, শ /s, ʃ/ keep its sound. So it is simply coded to <m>.

ক: রুক্মিনী /rukmini/ → <rukmini>

গ: বাগ্মী /bagmi/ → <bagmi>

যুগ্ম /ɟugmo/ → <jugm>

ঙ: বাঙ্ময় /baŋmɔj/ → <bangmy>

বাঙ্মুখ /baŋmukʰ/ → <bangmuk>

ট: কুট্মল /kutmɔl/ → <kuTml>

কুট্মলিত /kutmoliṭo/ → <kuTmlit>

ণ: হিরণ্ময় /hirɔnmɔj/ → <hirnmy>

মৃণ্ময় /mrinmɔj/ → <mrinmy>

ন: উন্মাদ /unmaḍ/ → <unmad>

জন্ম /ɟɔnmo/ → <jnm>

ম: সম্মান /ʃɔmman/ → <smman>

সম্মতি /ʃɔmmoṭi/ → <smmti>

গ: গুল্ম /gulmo/ → <gulm>

বল্মীক /bolmik/ → <blmik>

স: সুস্মিতা /ʃuʃmiṭa/ → <susmita>

ষ: কুষ্মান্ড /kuʃmando/ → <kusmand>

শ: কাশ্মীর /kaʃmir/ → <kasmir>

Case 4: Otherwise when attached to medial consonant or in the terminal position the ম /m/ phalaa acts as a diacritic as the letter it attaches to is geminated, so it has the same code as the previous code.

ছদ্ম /cʰɔḍḍõ/ → <cdd>

পদ্ম /pɔḍḍõ/ → <pdd>

হ HA \u09B9 IPA: /h/

হ with ঋ: When combining with ঋ /ri/, হ /h/ loses it sound as a full consonant and marks aspiration on ঋ /ri/. So, it is *Not Coded*.

হ্ঋদয় /rʰiḍoj/ → <ridy>

হ্ঋতপিন্ড /rʰiḍpindo/ → <ritpinD>

হ with র: When combining with র /r/, হ /h/ loses it sound as a full consonant and marks aspiration on র /r/. So, it is *Not Coded*.

হ্রদ /rʰɔḍ/ → <rd>

হ্রাস /rʰaʃ/ → <ras>

হ with ণ/ন: When হ /h/ combines with ণ/ন /n/ the /h/ is pronounced after /n/, নহ /nh/ or acts as aspiration on /n/, /nʰ/. So, it is encoded as <n>.

পূর্বাহ্ন /purbannho/ → <purbann>

চিহ্ন /chinnho/ → <cinn>

প্রাহ্ণ /prannho/ → <prann>

হ with ম: When হ /h/ combines with ম /m/ the /h/ is pronounced after /m/, মহ /mh/ or acts as aspiration on /m/, /mʰ/. So, it is encoded as <m>.

ব্রহ্মা /bromma/ → <brmma>

ব্রাহ্ম /brammo/ → <brамm>

হ with য: When হ /h/ combines with the allograph of য /ɟ/, the resultant sound is an aspirated geminate. So, it is encoded as the same as য /ɟ/, which is <j>.

উহ্য /uɟɟʰo/ → <ujj>

ঐতিহ্য/oiṭiɟɟʰo/ → <oitijj>

হ with ল: হ doesn't have any sound in conjuncts with ল i n the initial position of a word. So, it is *Not Coded.*

হ্লাদ /lhad/ → <lad>

When হ /h/ combines with ল /l/ in the medial or final position the /h/ is pronounced after /l/, লহ /lh/. So, it is encoded as <l>.

আহ্লাদ /allhad/ → <allad>

হ with ব: According to grammatical rules হ should be sounded like ও or উ and ব should be sounded as ভ.

আহ্বান /aovan/ → আওভান

However, most native speakers pronounce these words the same way as it is written. For example, আহ্বান is usually pronounced as আহভান /ahobʰan/, so we encode it to two different codes for the two different pronunciations.

আহ্বান → আওভান /aovan/ → <aoban>

আহ্বান → আহভান /ahobʰan/ → <ahban>

ঃ VISARGA \u0983

Case 1: In the medial position ঃ acts as a diacritic, geminating the sound of the consonant it follows. So, it gets the code of next character.

দুঃসময় /ḍuʃʃomoi/ → <dussmy>

দুঃখ /ḍukkho/ → <dukk>

Case 2: In the final positions with word length 2 or 3 ঃ acts as alphabetic হ /h/. So it is encoded as <h>.

উঃ /uh/ → <uh>

বাঃ /bah/ → <bah>

Case 3: In the final position with word length greater than 3, ঃ acts as alphabetic ও /o/. However, since ও /o/ is *Not Coded* in our encoding, in this case ঃ is *Not Coded* as well.

পুনঃ /puno/ → <pun>

অধঃ /odho/ → <od>

IV. PERFORMANCE IN SPELLING CHECKER APPLICATION

Table 1 shows the performance of this encoding when it is used on 1607 commonly misspelled words found in [18]. We first apply our encoding to both the correct and misspelled words, then compute the phonetic edit distance between the two encoded versions using the algorithm in [19]. It is considered correct if the edit distance is 0. In our case 134 out of 1607 words do not produce an edit distance of 0 with the correct word, which are termed as error, resulting in an accuracy of 91.37%.

Table 1. *Encoding performance*

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

The number of unmatched words fall to 107 and 27 if we consider edit distances of 1 and 2 respectively, as shown in Table 2.

Table 2. *Error distribution*

Error	134
Edit Distance 1	107
Edit Distance 2	27

Misspelled words with an edit distance of 1 can be easily handled using existing techniques, while those with an edit distance of 2 can also be handled with only slightly higher complexity. Table 3 shows a performance comparison of spelling checkers using three different methods: (i) traditional edit distance algorithm [19], (ii) Soundex encoding described in [9], and (iii) our proposed encoding. For the Soundex and Double Metaphone methods, the error (denoted by *E* in the table) is calculated from the phonetic edit distance between the encoded versions. The results clearly show that the proposed encoding performs much better than the other existing methods for the sample chosen.

Table 3. *Performance comparison*

Edit Distance			**Soundex**			**Double Metaphone**		
Misspelled Word	Correct Word	E	Misspelled Word	Correct Word	E	Misspelled Word	Correct Word	E
কসট /kɔʃto/	কষ্ট /kɔʃto/	2	<ksT>	<ksT>	0	<ksT>	<ksT>	0
দুকখ /d̪ukkʰo/	দুঃখ /d̪ukkʰo/	1	<dukk>	<duhk>	1	<dukk>	<dukk>	0
ষামি /ʃami/	স্বামী /ʃami/	2	<sami>	<sbami>	1	<sami>	<sami>	0
অততান্ত /ot̪t̪anṭo/	অত্যন্ত /ot̪t̪ɔnṭo/	2	<ottant>	<otjnt>	2	<ottant>	<ottnt>	1
রিদয় /riḍoj/	হৃদয় /rʰiḍoj/	2	<ridy>	<hrdy>	2	<ridy>	<ridy>	0
বিসশো /biʃʃo/	বিশ্ব /biʃʃo/	2	<biss>	<bisb>	1	<biss>	<biss>	0
চাদ /caḍ/	চাঁদ /cãḍ/	1	<cad>	<cad>	0	<cad>	<cad>	0
অস্তমান /ɔst̪oman/	অস্তায়মান /ɔst̪ajoman/	2	<ostman>	<ostayman>	2	<ostman>	<ostayman>	2
জরাজীরনো /ɟɔraɟirno/	জরার্জীর্ণ /ɟɔraɟirno/	4	<jbrajirn>	<jrajirn>	1	<jrajirn>	<jrajirn>	0
তরংগ /t̪ɔrɔŋgo/	তরঙ্গ /t̪ɔroŋgo/	2	<trmg>	<trmg>	0	<trngg>	<trngg>	0
কনা /kɔna/	কণা /kɔna/	1	<kna>	<kna>	0	<kna>	<kna>	0
নিন্দ্যনিয় /ninḍɔnijo/	নিন্দনীয় /nindonijo/	3	<nindjniy>	<nindniy>	1	<nindniy>	<nindniy>	0
পদদ /pɔḍḍo/	পদ্ম /pɔḍḍõ/	2	<pdd>	<pdm>	1	<pdd>	<pdd>	0
নিচ /nic/	নীচ /nic/	1	<nic>	<nic>	0	<nic>	<nic>	0

V. CONCLUSION

We present a Double Metaphone encoding for Bangla, tailored for spelling checking application. This encoding encapsulates the complex spelling rules for Bangla, and in addition, takes into account some of the dialectic pronunciation differences that are not possible to handle otherwise. The performance results show that it easily outperforms the current state of the art Bangla spelling checkers in producing appropriate suggestions for not only the commonly misspelled words, but also for the large number of "corner" cases which are currently beyond the reach of the other existing methods.

ACKNOWLEDGMENT

This work has been supported by the PAN Localization Project (www.PANL10n.net) grant from the International Development Research Center, Ottawa, Canada, administered through Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Pakistan. Special thanks to Naira Khan for correcting the IPA symbols in this paper.

REFERENCES

[1] The Summer Institute of Linguistics (SIL) Ethnologue Survey 1999, available online at http://www2.ignatius.edu/faculty/turner/languages.htm.

[2] Facts about the World's Languages: an Encyclopedia of the World's Major Languages, Past and Present, Jane Garry and Carl Rubino (ed.), New York/Dublin: H.W. Wilson Press, 2001.

[3] P. Kundu and B.B. Chaudhuri, "Error Pattern in Bangla Text", *International Journal of Dravidian Linguistics*, 28(2), 1999.

[4] The Soundex Algorithm, available online at http://www.archives.gov/research_room/genealogy/census/soundex.html .

[5] Lawrence Phillips, "Hanging on the Metaphone", *Computer Language*, 7(12), 1990.

[6] Lawrence Phillips, "The Double Metaphone Search Algorithm", *C/C++ Users Journal*, 18(6), June, 2000.

[7] Hodge, Victoria J. and Austin, Jim. (2001a). An Evaluation of Phonetic Spell Checkers, Technical Report YCS 338. Department of Computer Science, University of York

[8] Naushad UzZaman, "Phonetic Encoding for Bangla and its Application to Spelling checker, Name searching, Transliteration and Cross language information retrieval", Undergraduate thesis (Computer Science), BRAC University, May 2005

[9] Naushad UzZaman and Mumit Khan, "A Bangla Phonetic Encoding for Better Spelling Suggestion", *Proc. 7th International Conference on Computer and Information Technology*, Dhaka, December, 2004.

[10] Md. Tamjidul Haque and M. Kaykobad, "Quantitative Approaches for Bangla Spell Checker", *Proc. 6th International Conference on Computer and Information Technology*, Dhaka, December, 2003.

[11] Md. Tamjidul Haque and M. Kaykobad, "Use of Phonetic Similarity for Bangla Spell Checker", *Proc. 5th International Conference on Computer and Information Technology*, Dhaka, December, 2002.

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

[13] Arif Billah Al-Mahmud Abdullah and Ashfaq Rahman, "A Different Approach in Spell Checking for South Asian Languages", *Proc. 2nd International Conference on Information Technology for Applications (ICITA)*, China, 2004.

[14] The Unicode Consortium, *The Unicode Standard, Version 4.0*, Addison-Wesley, 2003.

[15] Bangla Uccharon Obidhan, Bangla Academy, Dhaka, Bangladesh.

[16] Bangla Banan Obidhan, Bangla Academy, Dhaka Bangladesh.

[17] R. Ishida's Bengali script notes [Draft], available online at http://people.w3.org/rishida/scripts/bengali/bengali-script/.

[18] Bangla Banan Obhidhan, Dr. Khurshid Alam, Mirnava, Dhaka, Bangladesh.

[19] Levenshtein edit distance algorithm, available online at http://www.nist.gov/dads/HTML/Levenshtein.htm