Interdisciplinary Journal of Linguistics
Volume [9] 2016, Pp.51-66

Designing a Digital Pronunciation Dictionary in Bangla

Niladri Sekhar Dash*

Abstract

It will be a nice learning experience for the Bangla language learners if an on-line Bangla education system is supported with a Digital Bangla Pronunciation Dictionary (DBPD), which may be accessed in classroom and at home, as the case may be, as one of the most useful reference guides for learning standard and acceptable pronunciation of Bangla words. With direct utilization of the DBPD, the learners will learn how Bangla words should be pronounced in standard or acceptable mode. In the era of on-line education a digital resource of this kind has the potential to improve on the traditional methods of language teaching where learners get opportunities to learn standard or acceptable pronunciation with direct utilization of modern computer technology in an interactive fashion with indirect assistance of language teachers. Through activation of a dialogue-based interactive user interface, learners will know how words in Bangla are pronounced in acceptable manner when the orthography of words hardly matches with pronunciation. This dictionary will also provide opportunities to the learners to understand how variations in pronunciation of words are caused due to variations in part-of-speech and meanings of similar orthographic forms. Within on-line education, this dictionary can be highly useful for non-native and foreign learners as well as for the speakers of different Bangla dialects and regional varieties – because they will get good opportunities and exposure to learn pronunciation of words considered standard and acceptable. Further application of this dictionary may be visualized in speech recognition, digital lexicography, text-to-speech conversion, language description, and language planning. Keeping such applications in mind, in this paper, I have tried to present the strategies and methods we have adopted to develop a DBPD in Bangla as a part of the digitization process of Bangla language education system. Our strategies and methods can easily be adopted for any of the Indian languages for developing such a highly useful digital resource for the service of its people.

Keywords: Pronunciation, Orthography, Parts-of-speech, Meaning, Bangla, Dictionary

1. Introduction

This paper presents an outline of the methods and strategies we have adopted for designing a Digital Bangla Pronunciation Dictionary (DBPD) in Bangla, which is being developed with a lexical database of nearly hundred thousand words obtained from a digital corpus of modern Bangla written texts as well as from various other digital lexical sources. This dictionary is different from other dictionaries in the sense that it aims at providing information of pronunciation of words in orthographic and audio output (Dash 2007b). The pronunciation output is produced in standard Bangla orthography, in Indic Roman, and in International Phonetic Alphabet (IPA) – and all these are available in machine-readable form. In addition, one can listen to audio output of pronunciation of words from this dictionary.

The primary objective behind this attempt is to address the need of a DBPD for Bangla, which may be directly used in on-line teaching, computer assisted language teaching, pronunciation teaching, text-to-speech system development, and similar other works in Bangla. Since there is no such digital resource developed for Bangla language as yet, we are developing this resource keeping in view its multiple utilities in various domains of language education, speech technology, applied linguistics, language description and language planning.

* Linguistic Research Unit, Indian Statistical Institute, Kolkata

The inspiration behind designing this DBPD is the lack of such a resource in digital form in the language. Although there are few printed pronunciation dictionaries in the language (Bhattacharya 1993, Choudhury 2009), these are not up-to-date to capture and represent of the pronunciation of words as we find today. For instance, we can refer to the *Saṃsad Bāṅglā Uccāraṇ Abhidhān* (Bhattacharya 1993). It was published nearly two decades ago and it has lost much of its referential relevance due to its inability to reflect on the changes taken place in pronunciation of Bangla words over the years as observed in Standard Colloquial Bangla (SCB). The other printed dictionary, which is published just a few years ago, is the *Śabda Saṃket* (Choudhury 2009), which has also several limitations in representation of actual pronunciation of words in both IPA and in Bangla orthography, as well as in presentation of other relevant information of the words.

Because of the limitations of printed dictionaries as well as because of the lack of digital pronunciation dictionaries in Bangla, we have started developing the DBPD with as much as load of lexical information for words as possible. The present DBPD includes – with a happy user-friendly interface – the following layers of information for the words included in the dictionary (Table 1). The audio output of pronunciation of words is available both in sentence-free and sentence-bound contexts.

No	Item	Example
(1)	Entry word in Bangla orthography	অক্ষি
(2)	Part-of-speech of the word	Noun
(3)	Word in Indic Roman script with diacritic	akṣi
(4)	Pronunciation of word in Bangla orthography	[ওক্.খি]
(5)	Pronunciation of the word in IPA	[o^{k}khi]
(6)	Audio output of pronunciation of words	[o^{k}khi]
(7)	Meaning of the word in Bangla	চোখ, নয়ন, আঁখি, লোচন, নেত্র
(8)	Meaning of the word in English	Eye
(9)	Usage in a Bangla sentence	তার অক্ষিকমল অতি সুন্দর
(10)	Translation of Bangla sentence into English	Her eyes are very Beautiful

Table 1: Presentation of lexicographic information in the dictionary

In our view the layers of information presented for each entry word in the DBPD will be enough to serve the requirements of the Bangla speech community as well as the non-native Bangla language learners. And due to this service this DBPD will be treated as a highly useful digital linguistic resource that is able to serve various linguistic needs of the learners with proper representation of actual empirical information of words of the language.

Although development of such a digital resource is a real technological challenge (as it asks for a perfect equilibrium of knowledge and expertise between the computer science and the linguistics), we believe that a happy collaboration of experts of the two fields will deliver a potentially robust system for the language and its people (Dash 2010b).

After justifying the rationale behind the development of the DBPD in Bangla (Section 1), I have referred to the some of the complexities involved in pronunciation of Bangla words (Section 2); defined the methods of selection of words for the dictionary (Section 3); justified the selection of particular spelling of words for the dictionary (Section 4); focussed on the methods of representation of pronunciation of the words in the dictionary (Section 5); investigated the issues of pronunciation of words with reference to their parts-of-speech and meanings (Section 6); reported about the present state of development of the dictionary (Section 7); and finally identified the potential areas of utilization of the dictionary in different domains of language and linguistics.

2. Pronunciation of Bangla Words

Variation of pronunciation of words has been one of the crucial issues in Bangla from the early days when the language was gradually evolving as an independent variety. This has been manifested in the orthographic representation of words in old Bangla texts like the *Caryāpadas* (Ali 1964, Das 1997, Nath 1980, Nath 1987, Nath 2011).

In modern Bangla written texts it is found that many words show striking differences between their surface forms and their pronunciations (Chaudhury 1990), because there is hardly 1: 1 mapping between orthography and pronunciation of characters used in formation of words (Dash, Chowdhury, and Sarkar 2011). Gradually, the language has become famous (or notorious!) for having a large number of words, which exhibit striking differences between their surface orthographic forms and their pronunciations as the following sample list of words shows (Table 2).

In general, a Bangla word, as a complete lexical unit, exhibits just one pronunciation. And in most cases, this is an accepted pronunciation and usually treated as a standard one. For instance, words like ঘর (ghar) [gɦɔr] "home", হাত (hāt) [ɦat] "hand", গাড়ি (gāṛi) [gaɽi] "car", বালিকা (bālikā) [balika] "girl", রাস্তা (rāstā) [rasta] "road", etc. are words where there is no variation in pronunciation of the characters used in formation of the words. Perhaps, one- third of the total words available in the language comes under this category.

On the other hand, for a large number of Bangla words, the phenomenon of variation in pronunciation is a distinct phonological feature (Dash 2006). This demands for special attention to trace the actual intended pronunciation of the words based on their form, part-of-speech, meaning, and context of their occurrence in texts. In a simple count, there are nearly twenty five thousand Bangla words, which fall within this category. Such words need to be included in a lexical database as they ask for special treatment in the proposed DBPD.

Bangla Word	In Roman with Diacritic	Pronunciation in Bangla orthography	Pronunciation in IPA
অগ্নি	agni	অ্হদৃা	[ogni]
অক্ষি	akṣi	অ্হদৃ্হ	[o^{k}khi]
ক্ষতি	kṣati	চ্হর্ভৃ	[khoti]
চঞ্চু	cañcu	চঁঞ্জহ্লহ্লৈ	[concu]
জ্ঞান	jñān	ঃঞ্জহ্ল	[gæn]
চিহ্ন	cihna	দৌঁহ্লচ্ভি	[cinɦo]
বিশ্ব	biśva	দৃ্হহ্লচ্হঙ্ক	[biʃ̍ʃo]
রঞ্জু	rañju	চ্ঞহ্লএ॰	[roɲju]
বহ্নি	bahni	চ্ঞহ্লদৃ্ি	[bonɦi]
বন্দী	bandī	চ্ঞহ্লদৃ্দ্	[bondi]
বাক্য	bākya	ঙ্খহ্লচ্খঙ্ক	[ba^{k}ko]
লক্ষ্মী	lakṣmī	চ্ক্ষহ্লদৃ্হ	[lo^{k}khi]
ঋত্বিক	ṛttvik	দৃ্লদৃ্ঞহ্ল	[ri^{t}tik]
বিল্ব	bilva	দৃ্হপ্হচ্হঙ্ক	[bi^{l}lo]
স্বামী	svāmī	হঙ্কদৃ্হ	[ʃami]
ব্রহ্ম	brahma	চ্ছঙ্করহ্লচ্ভি	[bro^{m}ɦo]
মন্ত্রী	mantrī	চ্ঞহ্লদৃ্ঃ	[montri]
স্মৃতি	smṛti	দৃহ্হর্ভৃ্	[ʃr̃ĩti]

Table 2: Some Bangla words that show disparity between orthography and pronunciation

Based on such observations, we have classified those words that show pronunciation variation into three broad types:

(a) Non-inflected two-letter words,
(b) Non-inflected three letter-words, and
(c) Inflected and affixed words.

Some examples from each type are presented in three different tables (Table 3, Table 4, and Table 5) below.

Word	POS	Meaning	Pronunciation
কর (kårå)	Noun	tax or hand or sun beam	[kɔr]
কর (kårå)	Finite Verb	you do (impolite)	[kɔr]
কর (kårå)	Finite Verb	you do (polite)	[kɔro]
বল (bålå)	Noun	ball	[bɔl]
বল (bålå)	Finite Verb	you say (impolite)	[bɔl]
বল (bålå)	Finite Verb	you say (polite)	[bɔlo]
বন্ধ (båndhå)	Adjective	closed	[bɔndɦo]
বন্ধ (båndhå)	Noun	strike	[bɔndɦ]
মত (måtå)	Noun	opinion	[mɔt]
মত (måtå)	Adverb	like or such as	[mɔto]
সর (sårå)	Noun	thin layer over boiled milk	[ʃɔr]
সর (sårå)	Finite Verb	you move (impolite)	[ʃɔr]
সর (sårå)	Finite Verb	you move (polite)	[ʃɔro]
হত (håtå)	Finite Verb	would have been	[ɦoto]
হত (håtå)	Adjective (PPL)	killed	[ɦɔto]
হল (hålå)	Noun	hall	[ɦɔl]
হল (hålå)	Finite Verb	was/became	[ɦolo]

Type 3: Multiple pronunciations of non-inflected two letter words

Word	**Part-of-Speech**	**Meaning**	**IPA**
কমল (kåmålå)	Noun	lotus	[kɔmɔl]
কমল (kåmålå)	Finite Verb	reduced	[komlo]
পরত (påråtå)	Noun	layer	[pɔrɔt]
পরত (påråtå)	Finite Verb	use to wear	[porto]
পরব (påråbå)	Noun	festival	[pɔrɔb]
পরব (påråbå)	Finite Verb	will wear	[porbo]
বসত (båsåtå)	Noun	place of habitation	[bɔsot]
বসত (båsåtå)	Finite Verb	used to seat	[bosto]
সরব (såråbå)	Adjective	vocal	[ʃɔrɔb]
সরব (såråbå)	Finite Verb	will move	[ʃorbo]
সরল (sårålå)	Adjective	straight	[ʃɔrɔl]
সরল (sårålå)	Finite Verb	moved	[ʃorlo]

Table 4: Multiple pronunciations of non-inflected three-letter words

Word	**Part-of-speech**	**Meaning**	**Pronunciation**
করে (kåre)	Noun	in tax, in hand	[kɔre]
করে (kåre)	Finite Verb	(s)he does / they do	[kɔre]
করে (kåre)	No-finite verb	doing	[kore]

করে (kåre)	Indeclinable	by	[kore]
খেলে (khele)	Finite verb	they play / (s)he plays	[khæle]
খেলে (khele)	Finite verb	you ate	[khele]
খেলে (khele)	Non-Finite verb	playing	[khele]
গড়ে (gåṛe)	Noun	in the fort/castle	[gɔɽe]
গড়ে (gåṛe)	Non-finite Verb	constructing	[goɽe]
গড়ে (gåṛe)	Finite verb	constructs	[gɔɽe]
গড়ে (gåṛe)	Adverb	in an average	[gɔɽe]
দেখে (dekhe)	Finite Verb	(s)he sees	[dækhe]
দেখে (dekhe)	Non-Finite Verb	seeing	[dekhe]
বদলে (bådåle)	Adverb	in exchange of	[bɔdɔle]
বদলে (bådåle)	Non-Finite Verb	changing	[bodle]
মানসিক (mānåsik)	Noun	committed sacrifice	[manʃik]
মানসিক (mānåsik)	Adjective	mental	[manoʃik]
মেলা (melā)	Noun	fair	[mela]
মেলা (melā)	Infinitive	to unfurl	[mæla]
মেলে (mele)	Infinitive	to meet or match	[mele]
মেলে (mele)	Finite Verb	unfurl	[mæle]
সুরমা (suråmā)	Noun	person name	[ʃurɔma]
সুরমা (suråmā)	Noun	collyrium of eyes	[ʃurma]
হলে (håle)	Noun	in hall	[ɦɔle]
হলে (håle)	Non-finite Verb	being	[ɦole]
হলে (håle)	Finite Verb	were	[ɦole]

Table 5: Multiple pronunciations of inflected or affixed words

We have observed that the implicit reasons behind the phenomenon of pronunciation variation of words (as shown in Table 3, Table 4, and Table 5) are part-of-speech, meanings, and usages, besides other factors such as lexical ambiguity, polysemy, homonymy, etc. Therefore, keeping these factors in view it is very much possible to show how a Bangla word is differently pronounced based on its parts-of-speech and meaning as the following examples and the diagram show (Fig. 1)

(1) তখনকার দিনে গড়ে অনেক সৈন্য থাকত।
(tåkhånkār dine gåṛe anek sainya thāktå) (gåṛe) : Noun: [gɔɽe]
"In those days, many soldiers used to stay in the fort/castle.

(2) সে প্রতি বছর ওই চালাঘরটা ভাঙ্গে আর গড়ে।
(se pråti båchår oi cālāghårṭā bhāṅge ār gåṛe) (gåṛe) : FV: [gɔɽe]
"Every year he breaks and constructs that thatched house)

(3) অনেক কষ্টে সে এই হাসপাতালটা গড়ে তুলেছে।
(anek kåṣṭe se ei hāspātālṭā gåṛe tuleche) (gåṛe) : NFV: [goɽe]
"With much effort he has constructed this hospital)

(4) প্রতিদিন গড়ে প্রায় ৫০০ জন রোগী আসে।
(pråtidin gåṛe prāy 500 jån rogī āse) (gåṛe) : ADV: [gɔɽe]
"Everyday, in an average, nearly 500 patients come)

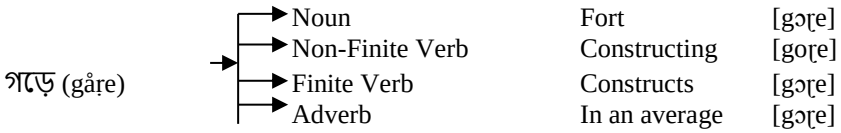

Fig. 1: A word is differently pronounced based on POS and meaning

What is understood from the discussion and examples given above is that some Bangla characters and letters – when used in formation of words – deviate from the standard norm of their sound representation in speech. Moreover, words change their pronunciation based on differences in parts-of-speech and meanings. Furthermore, various lexico-semantic factors are responsible for differences in pronunciation of words in the language (Dash, Chowdhury, and Sarkar 2011).

3. Selection of Words for DBPD

The purpose of developing the proposed DBPD it is necessary to collect a well-balanced and properly representative lexical database from the language. For our purposes, we have used four different resources:

(a) The TDIL corpus of modern Bangla prose texts (1981-1995)
(b) The modern Bangla Newspaper Corpus used in Bangla WordNet
(c) The digital lexical database of Spelling Dictionary of Bangla Akademi, Kolkata
(d) Some Bangla prose texts available in digital form

First, a large number of words are collected from the TDIL corpus of modern Bangla written prose texts that contains more than five million words (Dash 2007a). Second, words are collected from a modern Bangla Newspaper Corpus that contains nearly one million words. This corpus is developed for the purpose of sense validation of the Synsets used in the Bangla WordNet (Dash 2012). Third, a good number of words are collected from the lexical stock of the Spelling Dictionary of *Paschimbanga Bangla Akademi,* Kolkata, which is freely available in the internet. Finally, some words are collected from the digital prose texts freely found in the internet.

Since our target was to develop a DBPD of hundred thousand words, it was a challenge to select words from the large list of millions of words assembled by us. We had adopted the following criteria for selection of lexical items for the DBPD from the main lexical database.

(a) Single word units which show grapheme-phoneme disparities.
(b) Compound words which exhibit grapheme-to-phoneme disparities.
(c) English words which are nativised into Bangla with naturalized pronunciation.
(d) Prefixed and suffixed words (both single words and compound words)
(e) Inflected words are lemmatized before inclusion

We understand that the proposed scheme of lexical selection has some limitations. It is noted that a large majority of words in Bangla show grapheme-to-phoneme disparities as there is hardly 1:1 mapping relationship between a phoneme and a grapheme. These words, by virtue of their pronunciation uniqueness, are entitled to be included in the DBPD. Moreover, the words which exhibit grapheme-phoneme similarities also need to be included in the DBPD so that their pronunciation, usage, part-of-speech, and other lexicographical information are available to end

users. Therefore, it was decided that not only words with grapheme-phoneme disparities but also words with grapheme-phoneme similarities should be included in the DBPD.

The same argument stands valid for compound words. This policy was adopted because if we have not considered the compound words, then a large number of compound words would have been removed from the database. As a result of this a great number of compound words that constitute a major part of the Bangla vocabulary would have been eliminated from the dictionary. Since we did not want to lose such a huge amount of lexical stock of the language, we have kept compound words of both types (i.e., grapheme-phoneme disparities and grapheme-phoneme similarities) in the final database.

Following the strategies mentioned above, in total, one hundred thousand words are collected from all the corpora and lexical databases available to us. All the words in the final database are normalized and their spellings are corrected. However, the highly sanskritized complex words, which are not usually found in modern Bangla texts, are removed. The entire stock of words is put in the DBPD in simple alphabetically order.

The selected words belong to almost all parts-of-speech (i.e., noun, pronoun, non-finite verb, finite verb, adjective, adverb, postposition, and indeclinable) although their frequency of usage is not always uniform in all parts-of-speech. Besides, in some cases, particles and pre-conjuncts, such as, নাকি (nāki), তাই (tāi), সুতরাং (sutarāṃ), বলে (bale), যদি (ýadi), বরং (baraṃ), তো (to), etc. – due to their striking pronunciation variation – are also included in the word list selected for the dictionary.

With regard to etymology, it is observed that the final list of words includes Tatsama words, Tadbhava words, Deshi words and Local words. Moreover, some words belonging to different etymological antiquities (e.g., *English, Persian, French, Arabic, Dutch*, and others) are also taken into the list of the lexical stock. The words of foreign origin have been selected as suitable candidates, because these have become a part of the Bangla vocabulary over the decades and have been naturalised in the regular patterns of standard pronunciation of the language (Dash, Chowdhury and Sarkar 2009).

In the DBPD all words are stored in Unicode although their surface forms are shown in regular Bangla font as well as in Indic Roman script for better orthographic representation, robust computation, and simplified user accessibility. Also, each word is converted into a string made with symbols of IPA to make the system user-friendly for those people who do not know the Bangla script.

Each headword of the lexical database used in the proposed dictionary is alphabetically sorted out and stored without any discrimination made for single words, compound words, affixed forms, derived forms, content words, and functions words. Thus, the lexical database of the DBPD is rendered into Unicode-based Bangla script for better computation and easy access by both man and machine.

4. Issue of Spelling of Words

The modern Bangla texts record a large number of words that exhibit variation in spelling (Dash 2006). For elucidation, a sample list of words is given below (Table 6) which shows how a particular spelling, out of many variants, is selected for the present dictionary

Multiple Spelling Variations of a Word	Selected Spelling
রাণী (rāṇī), রাণি (rāṇi), রানী (rānī), রানি (rāni) “queen”	রানি (rāni)
হল (hålå), হলো (halo), হোলো (holo), হোল (holå) “becaome”	হল (hålå)
রঙ্গিন (råṅgin), রঙিন (råṅin), রঙ্গীন (råṅgīn), রঙীন (råṅīn) “colourful”	রঙিন (råṅin)
ঘুমানো (ghumāno), ঘুমুনো (ghumuno), ঘুমন (ghumånå), ঘুমনো (ghumåno) “to seep”	ঘুমানো (ghumāno)

কলিকাতা (kålikātā), কোলকাতা (kolkātā), কলকাতা (kålkātā) "Calcutta"	কলকাতা (kålkātā)
বাঙ্গালা (bāṅgālā), বাঙলা (bāṅlā), বাংলা (bāṃlā) "Bangla"	বাংলা (bāṃlā)
খ্রীষ্টাব্দ (khrīṣṭābdå), খ্রিষ্টাব্দ (khriṣṭābdå), খ্রিস্টাব্দ (khriṣṭābdå), খৃষ্টাব্দ (khṛṣṭābdå) "Christian era"	খ্রিষ্টাব্দ (khriṣṭābdå)

Table 6: Selection of a particular spelling for the dictionary

For the present work this issue is settled in a sensible manner for the precision of the task planed for the proposed dictionary. To avoid disputes and debates of all kinds, the spellings proposed in the spelling dictionary published by the *Pashchimbanga Bangla Akademi, Kolkata* (Sarkar, Mukhopadhyay and Dasgupta 2003, 2005) has been accepted due to the following reasons:

(a) This spelling system has greater acceptability among the native Bangla speakers of the state of West Bengal, India.
(b) Spellings proposed in this dictionary are in a process of naturalization in the language.
(c) Spellings proposed in this dictionary are now used to a great extent in preparing text books, dictionaries, grammar books, teaching aids, and other reference materials used in schools and colleges of the state.
(d) Government and its affiliated bodies of the state (e.g., *West Bengal Board of Primary Education, West Bengal Board of Secondary Education, West Bengal Council of Higher Secondary Education,* colleges and universities of the state, etc.) are using the spellings proposed in this dictionary.

Because of the factors stated above the spelling system proposed by the *Paścimbaṅga Bāṅglā Akademi, Kolkata* (2003 and 2005) is accepted and used in this dictionary. However, at certain situations, we have also consulted dictionaries produced by other agencies, such as, *Śabda Saṃket* (Chaudhury 2009), *Saṃsad Bāṅglā Abhidhān* (Biswas 2012), *Saral Bangla Abhidhān* (Mitra 2009), etc.

5. Representation of Pronunciation

Since the primary goal of the present dictionary is to be maximally authentic in representation of pronunciation of words as approved in SCB, all the words are reproduced in the present dictionary in three different manners:

(a) Pronunciation of a word is represented in regular Bangla orthography. For instance, the word বন্দী is represented as [®h¡eÚ¢c], the word কবি is represented as [কোবি], and the word মন is represented as [মোন্], etc. Through this kind of representation the native language users, who do not know Indic Roman or IPA, will be able to know how these words are to be pronounced with direct usage of known orthography.
(b) Pronunciation is also presented in IPA for non-Bangla speakers who do not know the Bangla script but are acquainted with IPA. With IPA, they will be able to easily use this dictionary for learning pronunciation of Bangla words. Pronunciation presented in IPA will also be helpful for those scholars who want to use this dictionary for text-to-speech research and machine learning.
(c) Pronunciation of each word is also presented in the form of speech output in audio format keeping in mind the need of those visually challenged language learners who are interested to hear pronunciation of Bangla words in spoken output form.

It can be argued that '3-Type representation' of pronunciation of Bangla words is an appreciable facility for the foreign learners, native learners, and less educated people as they can – with the

help of this dictionary – read, hear, and imitate pronunciation of the words to acclimatise and adopt the standard patterns of pronunciation of the Bangla words.

6. Complexities in Pronunciation Representation

A highly complicate feature of the Bangla language is that the pronunciation of a word in the SCB may vary due to variation in parts-of-speech and meaning of the word. However, the most striking feature is that, due to these factors, variation in pronunciation of words differs in three broad types, as shown below:

(a) Type-I: A word has just one part-of-speech, one meaning, and one pronunciation.
(b) Type-II: A word has two parts-of-speech, two meanings, and one pronunciation.
(c) Type-III: A word has two parts-of-speech, two meanings, and two pronunciations.

Such new and striking findings, which are retrieved from the Bangla lexical database, contribute heavily to classify the headwords into three broad categories based on parts-of-speech and meanings. Moreover, these have inspired us to treat the headwords in three different manners in the DBPD as the following sub-sections elucidate.

6.1 Pronunciation: Type-I

Most of the Bangla words, which show difference between their forms and pronunciations, belong to Type-I. Here a single word denotes one part-of-speech, one meaning, and just one pronunciation, although there is hardly any 1:1 parity between its form and its pronunciation. Consider, for instance, the underlined word in the following example:

(5) লোকটি তোমার ক্ষতি করতে পারে।
(lokṭi tomār kṣåti kårte pāre)
"The man can do harm to you."

In the above example (5) the word ক্ষতি (kṣåti) has just one part-of-speech (i.e., noun), one meaning (i.e., harm), and has one pronunciation (খোতি) [khoti]. However, there is hardly any parity between its surface form and its pronunciation because the word-initial consonant grapheme cluster ক্ষ (kṣ) is normally pronounced as [kh] in place of original [kʃ]. Thus the word is pronounced as (খোতি) [khoti], and not as [kʃoti] (Dash 2009). There are large numbers of words of this type in Bangla texts and these words require an elegant strategy for representing their pronunciation in the DBPD. We have used a simple technique to represent the pronunciation information of these words in the following manner (Fig. 2).

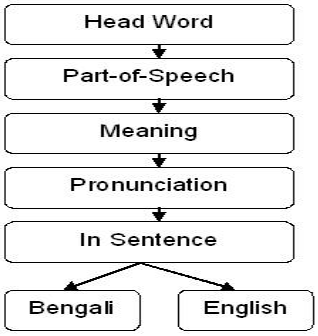

Fig. 2: Representation of words with Type-I pronunciation

The method of information storage for such headwords in the DBPD is as the followings: each headword is first alphabetically sorted out and then stored in MS Excel file in standard Bangla script compatible to Unicode. The part-of-speech information of the headword is stored in the second row to track its dictionarial part-of-speech. The third row displays the word in Indic Roman script with diacritic. The fourth row displays the pronunciation of the word in Bangla orthography. The fifth row represents the pronunciation of the word in IPA. The sixth row stores the audio output of pronunciation of the word. The Bangla meaning of the word is provided in the seventh row for sense disambiguation. The eighth row carries the English meaning of the word for sense understanding. The ninth row carries information of usage of the word in a well-formed Bangla sentence, while the tenth row contains an English translation of the Bangla sentence. The usage of the word is shown in last two rows, where the word is presented in sentence-bound context to provide ideas how the word has to be pronounced within a sentence taking into consideration its form, part-of-speech, and meaning into consideration. Moreover, the Bangla sentence is produced in regular English translation for better representation of information related to the headword. In essence, the entire load of information of the headword is represented in the following manner:

Row 1	Head word in Bangla and Unicode	ক্ষতি
Row 2	Part-of-speech of the word	Noun
Row 3	Display in Indic Roman Script	(kṣåti)
Row 4	Pronunciation in Bangla Script	[খোতি]
Row 5	Pronunciation in IPA	[khoti]
Row 6	Audio output in speech form	[khoti]
Row 7	Meaning of the word in Bangla	লোকসান
Row 8	Meaning of the word in English	loss
Row 9	Usage in a Bangla Sentence	লোকটি তোমার ক্ষতি করতে পারে।
Row 10	Translation in English	“The man can do harm to you”

Table 6: Information representation of Type-I headword

6.2 Pronunciation: Type-II

There are also good numbers of words in Bangla, which belong to the Type-II pronunciation. Here a single surface form of a word shows two parts-of-speech, two meanings, but just one pronunciation. Consider, for instance, the words like করে (kåre), বলে (båle), পড়ে (på̀re), দিয়ে (diye), etc. which belong to this type. For elucidation, consider the underlined words in the following examples:

(6) তুমি ওকে বইটা দিয়ে বাড়ি যাবে।
(tumi oke baiṭā diye bāṛi ýābe)
“You go home after giving him the book”

(7) আমার কথাটা মন দিয়ে শোন।
(āmār kåthāṭā mån diye śonå)
“Listen to my words carefully.”

The word দিয়ে (diye) in the above two sentences (6 and 7), represents one orthographic form, two different parts-of-speech, two different meanings, and just one pronunciation. In the first sentence (6) the words is used as a non-finite verb (NFV) with a meaning something like “after giving”, but in the second sentence (7), it is used as an indeclinable (IND) in the sense of “fixing

attention". Strikingly, in spite of differences in part-of-speech, meaning, and usage, the word represents just one pronunciation [dije].

The method of lexical information presentation for such headwords is almost same to the technique used for the headwords belonging to Type-I. The only notable difference is that although such words are marked with two different parts-of-speech, two different meanings, and two different usages, only one pronunciation is provided for these headwords in the DBPD, as the following diagram shows (Fig. 3). The method we have adopted for such headwords have been so useful and elegant that almost all the headwords belonging to this type are properly represented in the proposed dictionary.

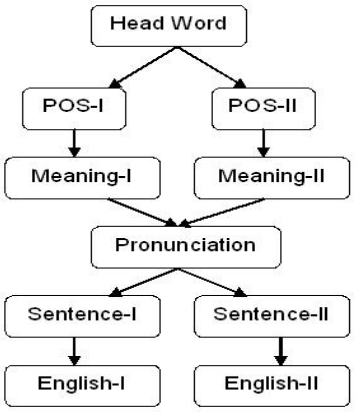

Fig. 3: Representation of words with Type-II pronunciation

6.3 Pronunciation: Type-III

In case of headwords belonging to Type-III, it is highly necessary to understand how part-of-speech and meaning of headwords can have strong effect on their pronunciations. Consider, for instance, the word বদলে (bådåle), which is pronounced in two different ways in SCB based on its part-of-speech and meaning to represent its two different lexical identities as the following examples show:

(8) সে নিজেকে অনেকটাই বদলে ফেলেছে ।
(se nijeke anekṭāi bådle pheleche)
"He has changed himself a lot".

(9) বিয়েতে গয়নার বদলে ক্যাশ টাকা দিতে পারেন।
(biyete gåynār bådåle kyāś ṭākā dite pāren)
"In the marriage you can give cash in place of ornaments".

The morphological information (given below) shows that the headword বদলে (bådåle) exhibits two different pronunciations due to variation in parts-of-speech and meaning.

Information Set: 1

Surface form of the word : বদলে (bådåle)
Base form of the word : বদল (bådål)

Suffix part tagged with word : -ে (-e)
Part-of-speech of the word : Non-finite verb (NFV)
Meaning of the word : Exchanging/changing
Pronunciation of the word : ®h¡cÚ®m [bodle]

Information Set: 2

Surface form of the word : বদলে (bådåle)
Base form of the word : বদলে (bådåle)
Suffix part tagged with word : Ø
Part-of-speech of the word : Postposition (PP)
Meaning of the word : In exchange of
Pronunciation of the word : বদলে [bɔdɔle]

The part-of-speech and semantic information presented above show that the word বদলে (bådåle), in a context-free situation, can exhibit two different pronunciations. When it acts as a non-finite verb (NFV), it is usually pronounced as ®h¡cÚ®m [bodle], and when it acts as a postposition [PP], it is pronounced as বদলে [bɔdɔle]. Other examples of this type are ধরে (dhåre) [dɦɔre] and ধরে (dhåre) [dɦore], বিকৃত (bikṛtå) [bikrito] and বিক্রিত (bikritå) [bi^{k}krito], পরে (påre) [pɔre] and পরে (påre) [pore], etc. (see Table 5 given above).
Furthermore, this interface provides two useful alternatives, which are highly beneficial for recognizing those homographic words, which have one orthographic form, two parts-of-speech, two meanings, two pronunciations, and two usages. For elucidation, let us consider the word হলে (håle), which is a homographic word having two different parts-of-speech vis-à-vis two different meanings, two pronunciations and two usages.

(10) ভালো সিনেমা হলে$_{[NFV/NN]}$ যেতে পারি।
(bhāla sinemā håle$_{[NFV/NN]}$ ýete pāri)

Meaning 1 : If the film is good, I can go.
Meaning 2 : If the cinema hall is good, I can go.

The word হলে (håle) in the above sentence (10) is a homographic form because it has one orthographic form, but two different parts-of-speech, two different meanings, two different pronunciations, and two different usages. Which of the two pronunciations will be triggered in the DBPD is interrelated to part-of-speech and meaning it is able to denote in the sentence.
The pronunciation of the word will be হোলে [ɦole] if the word is identified as a non-finite verb (NFV) that denotes a meaning something like "if something is being…". On the other hand, it will be pronounced as হলে [ɦɔle] if it is identified as a noun (NN) meaning "hall". Thus it implies that the word is pronounced in the first manner if the first part-of-speech and meaning is intended; and on the other hand, it is pronounced in the second manner, if the second part-of-speech and meaning is desired.
For those headwords which have one orthographic form, two different parts-of-speech, two different meanings, two different pronunciations, and two different usage variations, the interface for representing their pronunciation is a more complex process as the following diagram shows (Fig. 4).

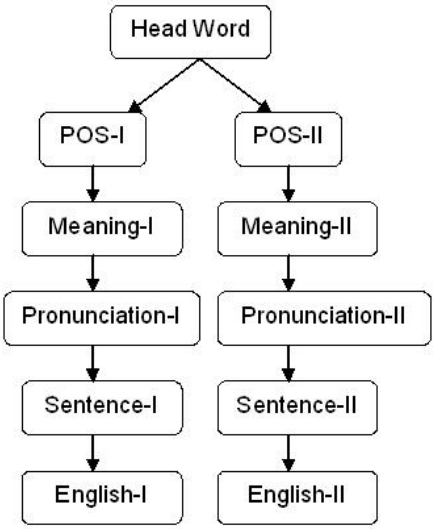

Fig.4: Representation of words with Type-III pronunciation

The total set of linguistic information for such headwords are more or less same to the headwords with other two types. Thus this DBPD, at the first stage, identifies if a headword has two parts-of-speech, and if it finds so, it produces two different alternative forms for generating two different meanings, two different pronunciations, as well as two different usages. That means it first determines if a headword has two parts-of-speech and then it tries to treat the headword in two different manners. Eventually, it properly identifies the parts-of-speech and different meanings and pronunciations. This technique has been greatly useful to deal with a large number of headwords, which fall in the category of Type-III pronunciation.

7. Present State of the Resource

The proposed Bangla DBPD is in the process of completion and it is expected that it will soon be made available to the end users. It is being developed in such a manner that it becomes maximally capable in capturing unique pronunciation aspects of the Bangla words with due reference to their parts-of-speech, meanings, and usages. Thus the dictionary will be quite competent to address various kinds of linguistic needs of the language and its people.

The lexical database is stored in the dictionary in an elegant manner with efficient user-interface of the database to help target users in accessing the dictionary in a seamless manner. Furthermore, the operational interface is also developed in such a way that users face little trouble to access relevant lexical database and information on-line without technical snag or linguistic complexity of any kind.

Since each headword is adorned with the Indic Roman script marked with diacritic marks, those who do not know the Bangla script but know the Indic Roman script can easily use it. Also, the pronunciation of the headwords is presented in the standard Bangla script for those who know the Bangla script but do not have any knowledge of the IPA. Pronunciation is presented in IPA also for those people who know IPA but have no idea about Bangla script. In fact, IPA is provided mainly for those people who are learning Bangla language at different universities and institutes in India and abroad. Audio speech output of pronunciation of the headwords is also provided for the visually challenged people and others.

The meaning for each entry word is provided with an English equivalent for sense disambiguation of words, which is particularly useful for the homographic and homophonous homonyms – the words which have similar orthographic forms or pronunciation but different meanings.

The dictionary access technique first looks at a headword as a lexical entity with Type-I pronunciation, and accordingly provides part-of-speech, meaning, pronunciation, and usage of the word in a sentential form. On the other hand, when it finds a headword with two or more parts-of-speech, it immediately provides two different meanings and pronunciations along with their usages in two or more different sentences. The necessary information for discrimination in part-of-speech, meaning, pronunciation, and usage is previously stored in the central lexical database of the dictionary so that it can be directly accessed to address various word-specific queries of the target dictionary users (Fig. 5)

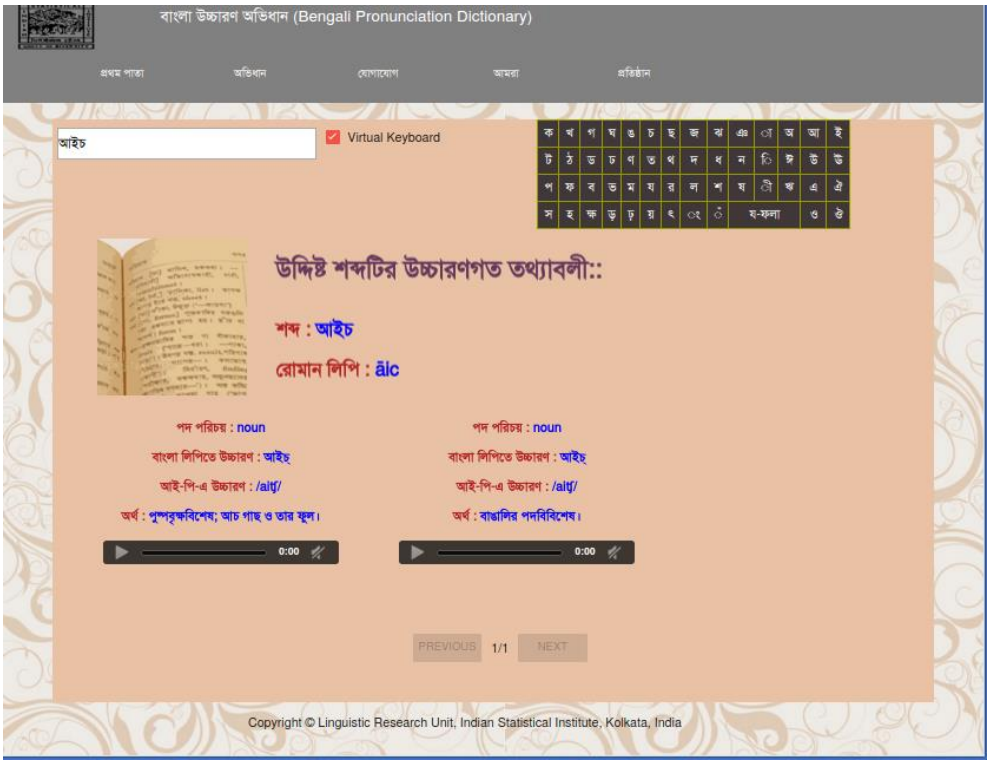

Fig. 5: A Screen shot of the present DBPD in Bangla

8. Conclusion: Application Relevance

This DBPD is, perhaps, the first of its kind in Bangla and in other Indian languages. It is fully computer-assisted with multimedia interface facilities for regular scope for data up-gradation, information augmentation, and system modification (Dash 2011). It has facilities for speech output, which can be effectively used in teaching Bangla as first or second language, on-line language teaching, text-to-speech conversion, language recognition, word recognition, machine learning, machine aided translation, Bangla to English parallel sentence generation, lexicography, word-sense disambiguation, and E-governance. It can also be used to train the linguistically impaired people in recognition and production of standard Bangla speech.

The immediate beneficiaries of this dictionary are the native Bangla learners, foreign Bangla learners, Bangla teachers, text-to-speech system developers, machine translation system designers, computational lexicographers, language planners, speech pathologists, and descriptive and cognitive linguists.

In the present era of information technology, a DBPD is considered as an indispensable resource in Computer Assisted Language Learning (CALL). Language teachers and course designers often incorporate this resource into an on-line and/or off-line language education system, where teaching learners about the pronunciation of words is an important part of the system (Jones 1986, Warschauer and Healey 1998, Bax 2003). Such applicational potentials act as a motivation behind the development of the resource for the Bangla language.

This DBPD, if compared with pronunciation dictionaries available in printed and digital format in Indian and foreign languages, invariably differs in its form content, treatment, and interface as depicted above. However, the striking deficiency of the present DBPD is that it does not relate to the pronunciation style used in Bangladesh, because it has not been possible for us to get access of this pronunciation type.

References

Ali, Saiyad Murtaja. 1964. Caryapader bhasa. *Sahitya Patrika.* 7(2): 86-106.

Bax, Stephen. 2003. Computer Assisted Language Learning (CALL): Past, Present and Future. *System*, 31(1): 13-28.

Bhattacharya, Subhas. 1993. *Samsad Bangla Uccharan Abhidhan (Samsad Bangla Pronunciation Dictionary)*. Kolkata: Sahitya Samsad.

Biswas, Shailendra. 2012. *Samsad Bangla Abhidhan (Samsad Bangla Dictionary).* 5th Edition. Kolkata: Sahitya Samsad.

Chaudhury, Jamil. 1990. *Banan o Uccharan (Spelling and pronunciation).* Dhaka: Bangla Academy Press.

Chaudhury, Jamil. 2009. *Sabda Sanket (A Bangla Dictionary with Pronunciation in IPA).* Kolkata: Dey's Publishing.

Das, Nirmal. 1997. *Caryagiti Parikrama (A Textual Criticism of the Old Bangla Carya lyrics).* Kolkata: Dey's Publishing.

Dash, Niladri Sekhar. 2006. *Bahurupi Bangla Banan (Multifaceted Bangla Spellings).* Kolkata: Daksha Bharati.

Dash, Niladri Sekhar. 2007a. Indian scenario in language corpus generation. In, Dash, Niladri Sekhar, Probal Dasgupta, and Pabitra Sarkar (Eds.) *Rainbow of Linguistics: Vol. I.* Kolkata: T. Media Publication. Pp. 129-162.

Dash, Niladri Sekhar. 2007b. The art of lexicography. In, V. Muhvic-Dimanovski and L. Sočanac (Eds.) *Encyclopaedia of Life Support Systems*, Oxford: EOLSS Publishers. Pp. 225-276.

Dash, Niladri Sekhar. 2009. Exploring the patterns of usage and utterance variations of Bangla consonant clusters for linguistic applications. Presented in the *31st All India Conference of Linguists (AICL-2009), Central University of Hyderabad,* Hyderabad, 15th-17th December. Pp. 171-172.

Dash, Niladri Sekhar. 2010. Utilization of language corpora in compilation of digital dictionaries for Indic languages. Presented in the *International Seminar on Tamil Computing*, 24th-26th February, 2010, Linguistic Studies Unit, Dept. of Tamil Language, Madras University, Chennai, India.

Dash, Niladri Sekhar. 2011. Some physical advantages of an electronic dictionary. *Indian Linguistics*. 71(1-4): 93-102.

Dash, Niladri Sekhar. 2012. Language Specific Synsets in Bangla: Some Empirical Explorations. *Journal of Advanced Linguistic Studies*. 1(1-2): 189-207.

Dash, Niladri Sekhar, Payel Dutta Chowdhury and Abhisek Sarkar. 2011. Digital Pronunciation Dictionary for Bangla: A Tool of the Time". In, Sharma, Dipti Misra, Rajeev Sangal and Sobha L. (Eds.) *Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011),* Pp. 117-124, Anna University, Chennai, India, 16th – 19th December 2011.

Dash, Niladri Sekhar, Payel Dutta Chowdhury and Abhisek Sarkar. 2009. Naturalization of English words in modern Bangla: a corpus-based empirical study, *Language Forum.* 35(2): 127-142.

Jones, George. 1986. Computer simulations in language teaching-the KINGDOM experiment. *System*. 14(2): 179-186.

Mitra, Subal Chandra. 2009. *Saral Bangala Abhidhan (Easy Bangla Dictionary).* 9th Edition. Kolkata: New Bengal Press.

Nath, Mrinal. 1980. Caryapader Bhasa: ekti anya bhabana. *Shabdakarna*. 2(2): 1-12.

Nath, Mrinal. 1987. Caryapader bhasar punarmulyayan. Bhasa. 4-5: 68-80.

Nath, Mrinal. 2011. *Charyapd: Bhasa, Patha, Rupantar (Caryapadas, Controversy about its language, texts along with renderings in Bangla).* Kolkata: Ebang Mushayera.

Sarkar, Pabitra, Amitabha Mukhopadhyay and Prasanta Kumar Dasgupta. 2003. *Akademi Banan Abhidhan* (Academy Spelling Dictionary), 4th Edition, Kolkata: Paschimbanga Bangla Akademi.

Sarkar, Pabitra, Amitabha Mukhopadhyay, and Prashanta Kumar Dasgupta. 2005. *Akademi Banan Abhidhan (Akademi Spelling Dictionary).* 5th Edition. Kolkata: Pashchimbanga Bangla Akademi.

Warschauer, Mark and Deborah Healey. 1998. Computers and language learning: an overview, *Language Teaching*, 31(1): 57-71.

❐❐❐