Evolution, Optimization, and Language Change: The Case of Bengali Verb Inflections

Monojit Choudhury[1], **Vaibhav Jalan**[2], **Sudeshna Sarkar**[1], **Anupam Basu**[1]
[1] Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India
{monojit,sudeshna,anupam}@cse.iitkgp.ernet.in
[2] Department of Computer Engineering
Malaviya National Institute of Technology, Jaipur, India
vaibhavjalan.mnit@gmail.com

Abstract

The verb inflections of Bengali underwent a series of phonological change between 10^{th} and 18^{th} centuries, which gave rise to several modern dialects of the language. In this paper, we offer a functional explanation for this change by quantifying the functional pressures of ease of articulation, perceptual contrast and learnability through objective functions or constraints, or both. The multi-objective and multi-constraint optimization problem has been solved through genetic algorithm, whereby we have observed the emergence of Pareto-optimal dialects in the system that closely resemble some of the real ones.

1 Introduction

Numerous theories have been proposed to explain the phenomenon of *linguistic change*, which, of late, are also being supported by allied mathematical or computational models. See (Steels, 1997; Perfors, 2002) for surveys on computational models of language evolution, and (Wang et al., 2005; Niyogi, 2006) for reviews of works on language change. The aim of these models is to explain why and how languages change under specific socio-cognitive assumptions. Although computational modeling is a useful tool in exploring linguistic change (Cangelosi and Parisi, 2002), due to the inherent complexities of our linguistic and social structures, modeling of real language change turns out to be extremely hard. Consequently, with the exception of a few (e.g., Hare and Elman (1995); Dras et al. (2003); Ke et al. (2003); Choudhury et al. (2006b)), all the mathematical and computational models developed for explaining language change are built for artificial toy languages. This has led several researchers to cast a doubt on the validity of the current computational models as well as the general applicability of computational techniques in diachronic explanations (Hauser et al., 2002; Poibeau, 2006).

In this paper, we offer a *functional explanation*[1] of a real world language change – the morpho-phonological change affecting the Bengali verb inflections (BVI). We model the problem as a multi-objective and multi-constraint optimization and solve the same using Multi-Objective Genetic Algorithm[2] (MOGA). We show that the different forms of the BVIs, as found in the several modern dialects, automatically emerge in the MOGA framework under suitable modeling of the objective and constraint functions. The model also predicts several

[1]Functionalist accounts of language change invoke the basic function of language, i.e. communication, as the driving force behind linguistic change (Boersma, 1998). Stated differently, languages change in a way to optimize their function, such that speakers can communicate maximum information with minimum effort (ease of articulation) and ambiguity (perceptual contrast). Often, ease of learnability is also considered a functional benefit. For an overview of different explanations in diachronic linguistics see (Kroch, 2001) and Ch. 3 of (Blevins, 2004).

[2]Genetic algorithm was initially proposed by Holland (1975) as a self-organizing adaptation process mimicking the biological evolution. They are also used for optimization and machine learning purposes, especially when the nature of the solution space is unknown or there are more than one objective functions. See Goldberg (1989) for an accessible introduction to single and multi-objective Genetic algorithms. Note that in case of a multi-objective optimization problem, MOGA gives a set of Pareto-optimal solutions rather than a single optimum. The concept of Pareto-optimality is defined later.

Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 65–74,
Prague, June 2007.

other possible dialectal forms of Bengali that seems linguistically plausible and might exist or have existed in the past, present or future. Note that the evolutionary algorithm (i.e., MOGA) has been used here as a tool for optimization, and has no relevance to the evolution of the dialects as such.

Previously, Redford et al. (2001) has modeled the emergence of syllable systems in a multi-constraint and multi-objective framework using Genetic algorithms. Since the model fuses the individual objectives into a single objective function through a weighted linear combination, it is not a multi-objective optimization in its true sense and neither does it use MOGA for the optimization process. Nevertheless, the present work draws heavily from the quantitative formulation of the objectives and constraints described in (Redford, 1999; Redford and Diehl, 1999; Redford et al., 2001). Ke et al. (2003) has demonstrated the applicability and advantages of MOGA in the context of the vowel and tonal systems, but the model is not explicit about the process of change that could give rise to the optimal vowel systems. As we shall see that the conception of the *genotype*, which is arguably the most important part of any MOGA model, is a novel and significant contribution of this work. The present formulation of the genotype not only captures a snapshot of the linguistic system, but also explicitly models the course of change that has given rise to the particular system. Thus, we believe that the current model is more suitable in explaining a case of linguistic change.

The paper is organized as follows: Sec. 2 introduces the problem of historical change affecting the BVIs and presents a mathematical formulation of the same; Sec. 3 describes the MOGA model; Sec. 4 reports the experiments, observations and their interpretations; Sec. 5 concludes the paper by summarizing the contributions. In this paper, Bengali graphemes are represented in Roman script following the ITRANS notation (Chopde, 2001). Since Bengali uses a phonemic orthography, the phonemes are also transcribed using ITRANS within two /s.

2 The Problem

Bengali is an *agglutinative language*. There are more than 150 different inflected forms of a single verb root in Bengali, which are obtained through affixation of one of the 52 inflectional suffixes, optionally followed by the emphasizers. The suffixes mark for the tense, aspect, modality, person and polarity information (Bhattacharya et al., 2005). The origin of modern Bengali can be traced back to Vedic Sanskrit (circa 1500 BC 600 BC), which during the middle Indo-Aryan period gave rise to the dialects like $M\bar{a}gadh\bar{i}$, and $Ardham\bar{a}gadh\bar{i}$ (circa 600 BC 200 AD), followed by the $M\bar{a}gadh\bar{i} - apabhramsha$, and finally crystallizing to Bengali (circa 10th century AD) (Chatterji, 1926). The verbal inflections underwent a series of phonological changes during the middle Bengali period (1200 - 1800 AD), which gave rise to the several dialectal forms of Bengali, including the standard form – the Standard Colloquial Bengali (SCB).

Attributes	Classical (Λ_0)	SCB	ACB	Sylheti
PrS1	*kari*	*kori*	*kori*	*kori*
PrS2	*kara*	*karo*	*kara*	*kara*
PrS3	*kare*	*kare*	*kare*	*kare*
PrSF	*karen*	*karen*	*karen*	*karoin*
PrC1	*kariteChi*	*korChi*	*kartAsi*	*koirtAsi*
PrC2	*kariteCha*	*korCho*	*kartAsa*	*koirtAsae*
PrC3	*kariteChe*	*korChe*	*kartAse*	*koirtAse*
PrCF	*kariteChen*	*korChen*	*kartAsen*	*kortAsoin*
PrP1	*kariAChi*	*koreChi*	*korsi*	*koirsi*
PrP2	*kariACha*	*koreCho*	*karsa*	*koirsae*
PrP3	*kariAChe*	*koreChe*	*karse*	*koirse*
PrPF	*kariAChen*	*koreChen*	*karsen*	*korsoin*

Table 1: The different inflected verb forms of Classical Bengali and three other modern dialects. All the forms are in the phonetic forms and for the verb root kar. Legend: (tense) Pr – present; (aspects) S – simple, C – continuous, P – perfect, ; (person) 1 – first, 2 – second normal, 3 – third, F – formal in second and third persons. See (Bhattacharya et al., 2005) for list of all the forms.

The Bengali literature of the 19^{th} century was written in the Classical Bengali dialect or the $s\bar{a}dhubh\bar{a}sh\bar{a}$ that used the older verb forms and drew heavily from the Sanskrit vocabulary, even though the forms had disappeared from the spoken dialects by 17^{th} century. Here, we shall take the liberty to use the terms "classical forms" and "Classical Bengali" to refer to the dialectal forms of middle Bengali and not Classical Bengali of the 19^{th} century literature. Table 1 enlists some of the corresponding verb forms of classical Bengali and SCB. Table 3 shows the derivation of some of the current verb inflections of SCB from its classical counterparts as reported in (Chatterji, 1926).

2.1 Dialect Data

Presently, there are several dialects of Bengali that vary mainly in terms of the verb inflections and intonation, but rarely over syntax or semantics. We do not know of any previous study, during which the different dialectal forms for BVI were collected and systematically listed. Therefore, we have collected dialectal data for the following three modern dialects of Bengali by enquiring the näive informants.

- *Standard Colloquial Bengali* (SCB) spoken in a region around Kolkata, the capital of West Bengal,
- *Agartala Colloquial Bengali* (ACB) spoken in and around Agartala, the capital of Tripura, and
- *Sylheti*, the dialect of the Sylhet region of Bangladesh.

Some of the dialectal forms are listed in Table 1. The scope of the current study is restricted to 28 inflected forms (12 present tense forms + 12 past tense forms + 4 forms of habitual past) of a single verb root, i.e., kar.

2.2 Problem Formulation

Choudhury et al. (2006a) has shown that a sequence of simple phonological changes, which we shall call the *Atomic Phonological Operators* or APO for short, when applied to the classical Bengali lexicon, gives rise to the modern dialects. We conceive of four basic types of APOs, namely Del or deletion, Met or metathesis, Asm or assimilation, and Mut or mutation. The complete specification of an APO includes specification of its type, the phoneme(s) that is(are) affected by the operation and the left and right context of application of the operator specified as regular expressions on phonemes. The semantics of the basic APOs in terms of rewrite rules are shown in Table 2.2. Since Bengali features assimilation only with respect to vowel height, here we shall interpret $Asm(p, LC, RC)$ as the height assimilation of the vowel p in the context of LC or RC. Also, we do not consider *epenthesis* or insertion as an APO, because epenthesis is not observed for the case of the change affecting BVI.

APO	Semantics
$Del(p, LC, RC)$	$p \rightarrow \phi / LC—RC$
$Met(p_i p_j, LC, RC)$	$p_i p_j \rightarrow p_j p_i / LC—RC$
$Asm(p, LC, RC)$	$p \rightarrow p' / LC—RC$
$Mut(p, p', LC, RC)$	$p \rightarrow p' / LC—RC$

Table 2: Semantics of the basic APOs in terms of rewrite rules. LC and RC are regular expressions specifying the left and right contexts respectively. p, p', p_i and p_j represent phonemes.

Rule No.	APO	Example Derivations		
		$kar - iteChe$	$kar - iten$	$kar - iAChi$
1	$Del(e, \phi, Ch)$	$kar - itChe$	NA	NA
2	$Del(t, \phi, Ch)$	$kar - iChe$	NA	NA
3	$Met(ri, \phi, \phi)$	$kair - Che$	$kair - ten$	$kair - AChi$
5	$Mut(A, e, \phi, Ch)$	NA	NA	kair-eChi
6	$Asm(a, i, \phi, \phi)$	$koir - Che$	$koir - ten$	$koir - eChi$
7	$Del(i, o, \phi)$	$kor - Che$	$kor - ten$	$kor - eChi$

Table 3: Derivations of the verb forms of SCB from classical Bengali using APOs. "NA" means the rule is not applicable for the form. See (Choudhury et al., 2006a) for the complete list of APOs involved in the derivation of SCB and ACB forms

The motivation behind defining APOs rather than representing the change in terms of rewrite rules is as follows. Rewrite rules are quite expressive and therefore, it is possible to represent complex phonological changes using a single rewrite rule. On the other hand, APOs are simple phonological changes that can be explained independently in terms of phonetic factors (Ohala, 1993). In fact, there are also computational models satisfactorily accounting for cases of vowel deletion (Choudhury et al., 2004; Choudhury et al., 2006b) and assimilation (Dras et al., 2003).

Table 3 shows the derivation of the SCB verb forms from classical Bengali in terms of APOs. The derivations are constructed based on the data provided in (Chatterji, 1926).

2.3 Functional Explanation for Change of BVI

Let Λ_0 be the lexicon of classical Bengali verb forms. Let $\Theta : \theta_1, \theta_2, \cdots \theta_r$ be a sequence of r APOs. Application of an APO on a lexicon implies the application of the operator on every word of the lexicon. The sequence of operators Θ, thus, represent a dialect obtained through the process of change from Λ_0, which can be represented as follows.

$$\Theta(\Lambda_0) = \theta_r(\cdots\theta_2(\theta_1(\Lambda_0))\cdots) = \Lambda_d$$

The derivation of the dialect Λ_d from Λ_0 can be constructed by following the APOs in the sequence of their application.

We propose the following functional explanation for the change of BVI.

A sequence of APOs, Θ is preferred if $\Theta(\Lambda_0)$ has some functional benefit over Λ_0. Thus, the modern Bengali dialects are those, which have some functional advantage over the classical dialect.

We would like to emphasize the word "some" in the aforementioned statements, because the modern dialects are *not better* than the classical one (i.e., the ancestor language) in an absolute sense. Rather, the classical dialect is suboptimal compared to the modern dialects only with respect to "some" of the functional forces and is better than the them with respect to "some other" forces. Stated differently, we expect both the classical as well as the modern dialects of Bengali to be Pareto-optimal[3] with respect to the set of functional forces.

In order to validate the aforementioned hypothesis, we carry out a multi-objective and multi-constraint optimization over the possible dialectal forms of Bengali, thereby obtaining the Pareto-optimal set, which has been achieved through MOGA.

3 The MOGA Model

Specification of a problem within the MOGA framework requires the definition of the *genotype*, *phenotype* and genotype-to-phenotype mapping plus the objective functions and constraints. In this section, we discuss the design choices explored for the problem of BVI.

[3]Consider an optimization problem with n objective functions f_1 to f_n, where we want to minimize all the objectives. Let S be the solution space, representing the set of all possible solutions. A soulution $sinS$ is said to be Pareto-optimal with respect to the objective functions f_1 to f_n, if and only if there does not exist any other solution $s' \in S$ such that $f_i(s') \leq f_i(s)$ for all $1 \leq i \leq n$ and $f_i(s') < f_i(s)$ for at least one i.

3.1 Phenotype and Genotype

We define the *phenotype* of a dialect d to be the lexicon of the dialect, Λ_d, consisting of the 28 inflected forms of the root verb kar. This choice of phenotype is justified because, at the end of the optimization process, we would like to obtain the Pareto-optimal dialects of Bengali and compare them with their real counterparts.

The *genotype* of a dialect d could also be defined as Λ_d, where the word forms are the genes. However, for such a choice of genotype, crossover and mutation lead to counter-intuitive results. For example, mutation would affect only a single word in the lexicon, which is against the *regularity* principle of sound change (see Bhat (2001) for explanation). Similarly, exchanging a set of words between a pair of lexica, as crossover would lead to, seems insensible.

Therefore, considering the basic properties of sound change as well as the genetic operators used in MOGA, we define a chromosome (and thus the genotype) as a sequence of APOs. The salient features of the genotype are described below.

• *Gene*: A gene is defined as an APO. Since in order to implement the MOGA, every gene must be mapped to a number, we have chosen an 8-bit binary representation for a gene. This allows us to specify 256 distinct genes or APOs. However, for reasons described below, we use the first bit of a gene to denote whether the gene (i.e., the APO) is active (the bit is set to 1) or not. Thus, we are left with 128 distinct choices for APOs. Since the number of words in the lexicon is only 28, the APOs for Del, Asm and Met are limited, even after accounting for the various contexts in which an APO is applicable. Nevertheless, there are numerous choices for Mut. To restrain the possible repertoire of APOs to 128, we avoided any APO related to the mutation of consonants. This allowed us to design a comprehensive set of APOs that are applicable on the classical Bengali lexicon and its derivatives.

• *Chromosome*: A chromosome is a sequence of 15 genes. The number 15 has been arrived through experimentation, where we have observed that increasing the length of a chromosome beyond 15 does not yield richer results for the current choice of APOs and Λ_0. Since the probability of any gene

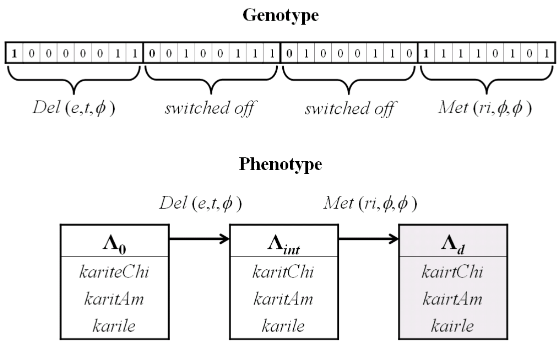

Figure 1: Schematic of genotype, phenotype and genotype-to-phenotype mapping.

being switched off (i.e., the first bit being 0) is 0.5, the expected number of active APOs on a chromosome with 15 genes is 7.5. It is interesting to note that this value is almost equal to the number of APOs required (7 to be precise) for derivation of the SCB verb forms.

• *Genotype to phenotype mapping*: Let for a given chromosome, the set of active APOs (whose first bit is 1) in sequence be $\theta_1, \theta_2, \cdots, \theta_r$. Then the phenotype corresponding to this chromosome is the lexicon $\Lambda_d = \theta_r(\cdots\theta_2(\theta_1(\Lambda_0))\cdots)$. In other words, the phenotype is the lexicon obtained by successive application of the active APOs on the chromosome on the lexicon of classical Bengali.

The concepts of gene, chromosome and the mapping from genotype to the phenotype are illustrated in Fig. 3.1. It is easy to see that the regularity hypothesis regarding the sound change holds good for the aforementioned choice of genotype. Furthermore, crossover in this context can be interpreted as a shift in the course of language change. Similarly, mutation of the first bit turns a gene on or off, and of the other bits changes the APO. Note that according to this formulation, a chromosome not only models a dialect, but also the steps of its evolution from the classical forms.

3.2 Objectives and Constraints

Formulation of the objective functions and constraints are crucial to the model, because the linguistic plausibility, computational tractability and the results of the model are overtly dependent on them. We shall define here three basic objectives of ease of articulation, perceptual contrast and learnability, which can be expressed as functions or constraints.

Several models have been proposed in the past for estimating the articulatory effort (Boersma (1998), Ch. 2, 5 and 7) and perceptual distance between phonemes and/or syllables (Boersma (1998), Ch. 3, 4 and 8). Nevertheless, as we are interested in modeling the effort and perceptual contrast of the whole lexicon rather than a syllable, we have chosen to work with simpler formulations of the objective functions. Due to paucity of space, we are not able to provide adequate details and justification for the choices made.

3.2.1 f_e: Articulatory Effort

Articulatory effort of a lexicon Λ is a positive real number that gives an estimate of the effort required to articulate the words in Λ in some unit. If f_e denotes the effort function, then

$$f_e(\Lambda) = \frac{1}{|\Lambda|}\sum_{w\in\Lambda} f_e(w) \qquad (1)$$

The term $f_e(w)$ depends on three parameters: 1) the length of w in terms of phonemes, 2) the structure of the syllables, and 3) the features of adjacent phonemes, as they control the effort spent in co-articulation. We define $f_e(w)$ to be a weighted sum of these three.

$$f_e(w) = \alpha_1 f_{e1}(w) + \alpha_2 f_{e2}(w) + \alpha_3 f_{e3}(w) \qquad (2)$$

where, $\alpha_1 = 1$, $\alpha_2 = 1$ and $\alpha_3 = 0.1$ are the relative weights.

The value of f_{e1} is simply the length of the word, that is

$$f_{e1}(w) = |w| \qquad (3)$$

Suppose $\psi = \sigma_1\sigma_2\cdots\sigma_k$ is the usual syllabification of w, where the usual or optimal syllabification for Bengali is defined similar to that of Hindi as described in (Choudhury et al., 2004). Then, f_{e2} is defined as follows.

$$f_{e2}(w) = \sum_{i=1}^{k} hr(\sigma_i) \qquad (4)$$

$hr(\sigma)$ measures the hardness of the syllable σ and is a function of the syllable structure (i.e. the CV pattern) of σ. The values of $hr(\sigma)$ for different syllable structures are taken from (Choudhury et al., 2004).

Since *vowel height assimilation* is the primary co-articulation phenomenon observed across the dialects of Bengali, we define f_{e3} so as to model only the effort required due to the difference in the heights of the adjacent vowels.

Let there be n vowels in w represented by V_i, where $1 \leq i \leq n$. Then f_{e3} is defined by the following equation.

$$f_{e3}(w) = \sum_{i=1}^{n-1} |ht(V_i) - ht(V_{i+1})| \qquad (5)$$

The function $ht(V_i)$ is the tongue height associated with the vowel V_i. The value of the function $ht(V_i)$ for the vowels $/A/$, $/a/$, $/E/$ $/o/$, $/e/$, $/i/$ and /u/ are 0, 1, 1, 2, 2, 3, and 3 respectively. Note that the values are indicative of the ordering of the vowels with respect to tongue height, and do not reflect the absolute height of the tongue in any sense.

3.2.2 f_d and C_d: Acoustic Distinctiveness

We define the acoustic distinctiveness between two words w_i and w_j as the edit distance between them, which is denoted as $ed(w_i, w_j)$. The cost of insertion and deletion of any phoneme is assumed to be 1; the cost of substitution of a vowel (consonant) for a vowel (consonant) is also 1, whereas that of a vowel (consonant) for a consonant (vowel) is 2, irrespective of the phonemes being compared. Since languages are expected to increase the acoustic distinctiveness between the words, we define a minimizing objective function f_d over a lexicon Λ as the sum of the inverse of the edit distance between all pair of words in Λ.

$$f_d(\Lambda) = \frac{2}{|\Lambda|(|\Lambda| - 1)} \sum_{ij, i \neq j} ed(w_i, w_j)^{-1} \qquad (6)$$

If for any pair of words w_i and w_j, $ed(w_i, w_j) = 0$, we redefine $ed(w_i, w_j)^{-1}$ as 20 (a large penalty).

We say that a lexicon Λ violates the acoustic distinctiveness constraint C_d, if there are more than two pairs of words in Λ, which are identical.

3.2.3 C_p: Phonotactic constraints

A lexicon Λ is said to violate the constraint C_p if any of the words in Λ violates the phonotactic constraints of Bengali. As described in (Choudhury et al., 2004), the PCs are defined at the level of syllable onsets and codas and therefore, syllabification is a preprocessing step before evaluation of C_p.

3.2.4 f_r and C_r: Regularity

Although learnability is a complex notion, one can safely equate the learnability of a system to the regularity of the patterns within the system. In fact, in the context of morphology, it has been observed that the so called *learning bottleneck* has a regularizing effect on the morphological structures, thereby leaving out only the most frequently used roots to behave irregularly (Hare and Elman, 1995; Kirby, 2001).

In the present context, we define the regularity of the verb forms in a lexicon as the predictability of the inflectional suffix on the basis of the morphological attributes. Brighton et al. (2005) discuss the use of Pearson correlation between phonological edit distance and semantic/morphological hamming distance measures as a metric for learnability. On a similar note, we define the regularity function f_r as follows. For two words $w_i, w_j \in \Lambda$, the (dis)similarity between them is given by $ed(w_i, w_j)$. Let $ma(w_i, w_j)$ be the number of morphological attributes shared by w_i and w_j. We define the regularity of Λ, $f_r(\Lambda)$, as the *Pearson correlation coefficient* between $ed(w_i, w_j)$ and $ma(w_i, w_j)$ for all pairs of words in Λ. Note that for a regular lexicon, $ed(w_i, w_j)$ decreases with an increase in $ma(w_i, w_j)$. Therefore, $f_r(\Lambda)$ is negative for a regular lexicon and 0 or positive for an irregular one. In other words, $f_r(\Lambda)$ is also a minimizing objective function.

We also define a regularity constraint C_r, such that a lexicon Λ violates C_r if $f_r(\Lambda) > -0.8$.

4 Experiments and Observations

In order to implement the MOGA model, we have used the Non-dominated Sorting GA-II or NSGA-II (Deb et al., 2002), which is a multi-objective, multi-constraint elitist GA. Different MOGA models have been incrementally constructed by introducing the different objectives and constraints. The motivation behind the incorporation of a new objective or constraint comes from the observations made on the emergent dialects of the previous models. For instance, with two objectives f_e and f_d, and no constraints, we obtain dialects that violate phonotactic constraints or/and are highly irregular. One such example of an emergent dialect[4] is Λ = { *kor, kara, kar, kore, korea, kore, karA, karAa, karA, *korAlm, *korl, korla, *koreAlm, korel, korela, *karAlm, karAl, karAla* }. The * marked forms violate the phonotactic constraints. Also note that the forms are quite indistinct or close to each other. These observations led to the formulation of the constraints C_p and C_d.

Through a series of similar experiments, finally we arrived at a model, where we could observe the emergence of dialects, some of which closely resemble the real dialects and others also seem linguistically plausible. In this final model, there are two objectives, f_e and f_d, and 3 constraints, C_p, C_d and C_r. Table 4 lists the corresponding forms of some of the emergent dialects, whose real counterparts are shown in Table 1.

Fig. 2 shows the Pareto-optimal front obtained for the aforementioned model after 500 generations, with a population size of 1000. Since the objectives are minimizing in nature, the area on the plot below and left of the Pareto-optimal front represents impossible languages, whereas the area to the right and top of the curve pertains to unstable or suboptimal languages. It is interesting to note that the four real dialects lie very close to the Pareto-optimal front. In fact, ACB and SCB lie on the front, whereas classical Bengali and Sylheti appears to be slightly suboptimal. Nevertheless, one should always be aware that *impossibility* and *suboptimality* are to be interpreted in the context of the model and any generalization or extrapolation of these concepts for the real languages is controversial and better avoided.

Several inferences can be drawn from the experiments with the MOGA models. We have observed that the Pareto-optimal fronts for all the MOGA Models look like rectangular hyperbola with a horizontal and vertical limb; the specific curve of Fig. 2 satisfies the equation:

$$f_d(\Lambda)^{0.3}(f_e(\Lambda) - 5.6) = 0.26 \qquad (7)$$

Several interesting facts, can be inferred from the above equation. First, the minimum value of f_e under the constraints C_r and C_d, and for the given

[4]Due to space constraints, we intentionally omit the corresponding classical forms.

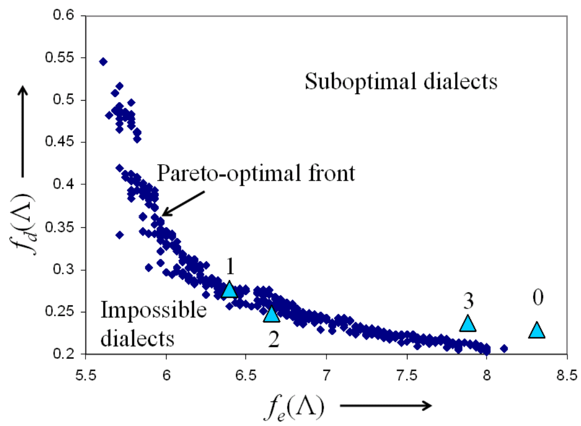

Figure 2: The Pareto-optimal front. The gray triangles (light blue in colored version available online) show the position of the real dialects: 0 – Classical Bengali, 1 – SCB, 2 – ACB, 3 – Sylheti. The top-most dot in the plot corresponds to the emergent dialect D0 shown in Table 4.

repertoire of APOs is 5.6. Second, at $f_e(\Lambda) = 6$, the slope of the front, i.e. df_d/df_e, is approximately -2, and the second derivative $d^2 f_d/df_e^2$ is around 20. This implies that there is sharp transition between the vertical and horizontal limbs at around $f_e(\Lambda) = 6$.

Interestingly, all the real dialects studied here lie on the horizontal limb of the Pareto-optimal front (i.e., $f_e(\Lambda) \geq 6$), classical Bengali being placed at the extreme right. We also note the negative correlation between the value of f_e for the real dialects, and the number of APOs invoked during derivation of these dialects from classical Bengali. These facts together imply that the natural direction of language change in the case of BVIs has been along the horizontal limb of the Pareto-optimal front, leading to the formation of dialects with higher and higher articulatory ease. Among the four dialects, SCB has the minimum value for $f_e(\Lambda)$ and it is positioned on the horizontal limb of the front just before the beginning of the vertical limb.

Therefore, it is natural to ask whether there are any real dialects of modern Bengali that lie on the vertical limb of the Pareto-optimal front; and if not, what may be the possible reasons behind their inexistence? In the absence of any comprehensive collection of Bengali dialects, we do not have a clear answer to the above questions. Nevertheless, it may

Attributes	D0	D1	D2	D3
PrS1	*kar*	*kor*	*kori*	*kori*
PrS2	*kara*	*kora*	*kora*	*kora*
PrS3	*kare*	*kore*	*kore*	*korA*
PrSF	*karen*	*koren*	*koren*	*koren*
PrC1	*kartA*	*karChi*	*karteChi*	*kairteChi*
PrC2	*kartAa*	*karCha*	*karteCha*	*kairteCha*
PrC3	*kartAe*	*karChe*	*karteChe*	*kairteChA*
PrCF	*kartAen*	*karChen*	*karteChen*	*kairteChen*
PrP1	*karA*	*korChi*	*koriChi*	*koriChAi*
PrP2	*karAa*	*korCha*	*koriCha*	*koriACha*
PrP3	*karAe*	*korChe*	*koriChe*	*koriAChA*
PrPF	*karAen*	*korChen*	*koriChen*	*koriAChen*

Table 4: Examples of emergent dialects in the MOGA model. Note that the dialects D1, D2 and D3 resemble SCB, ACB and Sylheti, whereas D0 seems to be linguistically implausible. For legends, refer to Table 1

be worthwhile to analyze the emergent dialects of the MOGA models that lie on the vertical limb. We have observed that the vertical limb consists of dialects similar to D0 – the one shown in the first column of Table 4. Besides poor distinctiveness, D0 also features a large number of diphthongs that might result in poorer perception or higher effort of articulation of the forms. Thus, in order to eliminate the emergence of such seemingly implausible cases in the model, the formulations of the objectives f_e and f_d require further refinements.

Similarly, it can also be argued that the structure of the whole lexicon, which has not been modeled here, has also a strong effect on the BVIs. This is because even though we have measured the acoustic distinctiveness f_d with respect to the 28 inflected forms of a single verb root kar, ideally f_d should be computed with respect to the entire lexicon. Thus, change in other lexical items (borrowing or extinction of words or change in the phonological structures) can trigger or restrain an event of change in the BVIs.

Furthermore, merging, extinction or appearence of morphological attributes can also have significant effects on the phonological change of inflections. It is interesting to note that while Vedic Sanskrit had different morphological markers for three numbers (singular, dual and plural) and no gender markers for the verbs, Hindi makes a distinction between the genders (masculine and feminine) as well as numbers (but only singular and plural), and Bengali has markers for neither gender nor number. Since both Hindi and Bengali are offshoots of Vedic Sanskrit, presumably the differences between the phonological structure of the verb inflections of these two languages must have also been affected by the loss or addition of morphological attributes. It would be interesting to study the precise nature of the interaction between the inflections and attributes within the current computational framework, which we deem to be a future extension of this work.

5 Conclusions

In this paper, we have described a MOGA based model for the morpho-phonological change of BVIs. The salient contributions of the work include: (1) the conception of the genotype as a sequence of APOs, whereby we have been able to capture not only the emergent dialects, but also the path towards their emergence, and (2) a plausible functional explanation for the morpho-phonological changes affecting the BVIs. Nevertheless, the results of the experiments with the MOGA models must be interpreted with caution. This is because, the results are very much dependent on the formulation of the fitness functions and the choice of the constraints. The set of APOs in the repertoire also play a major role in shaping the Pareto-optimal front of the model.

Before we conclude, we would like to reemphasize that the model proposed here is a functional one, and it does not tell us how the dialects of Bengali have self-organized themselves to strike a balance between the functional pressures, if at all this had been the case. The evolutionary algorithm (i.e., MOGA) has been used here as a tool for optimization, and has no relevance to the evolution of the dialects as such. Nevertheless, if it is possible to provide linguistically grounded accounts of the sources of *variation* and the process of *selection*, then the MOGA model could qualify as an evolutionary explanation of language change as well. Although such models have been proposed in the literature (Croft, 2000; Baxter et al., 2006), the fact, that global optimization can be an outcome of local interactions between the speakers (e.g., Kirby (1999), de

Boer (2001), Choudhury et al. (2006b)), alone provides sufficient ground to believe that there is also an underlying self-organizational model for the present functional explanation.

References

G. J. Baxter, R. A. Blythe, W. Croft, and A. J. McKane. 2006. Utterance selection model of language change. *Physical Review E*, 73(046118).

D.N.S. Bhat. 2001. *Sound Change*. Motilal Banarsidass, New Delhi.

S. Bhattacharya, M. Choudhury, S. Sarkar, and A. Basu. 2005. Inflectional morphology synthesis for bengali noun, pronoun and verb systems. In *Proc. of NCCPB*, pages 34–43, Dhaka.

Julia Blevins. 2004. *Evolutionary Phonology*. Cambridge University Press, Cambridge, MA.

P. Boersma. 1998. *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. Uitgave van Holland Academic Graphics, Hague.

Henry Brighton, Kenny Smith, and Simon Kirby. 2005. Language as an evolutionary system. *Physics of Life Reviews*, 2(3):177–226, September.

A. Cangelosi and D. Parisi. 2002. Comuputer simulation: A new scientific approach to the study of language evolution. In *Simulating the Evolution of Language*, pages 3–28. Springer Verlag, London.

S. K. Chatterji. 1926. *The Origin and Development of the Bengali Language*. Rupa and Co., New Delhi.

A. Chopde. 2001. Itrans version 5.30: A package for printing text in indian languages using english-encoded input. *http://www.aczoom.com/itrans/*.

M. Choudhury, A. Basu, and S. Sarkar. 2004. A diachronic approach for schwa deletion in indo-aryan languages. In *Proc. of ACL SIGPHON-04*, pages 20–26, Barcelona.

M. Choudhury, M. Alam, S. Sarkar, and A. Basu. 2006a. A rewrite rule based model of bangla morpho-phonological change. In *Proc. of ICCPB*, pages 64–71, Dhaka.

M. Choudhury, A. Basu, and S. Sarkar. 2006b. Multi-agent simulation of emergence of the schwa deletion pattern in hindi. *JASSS*, 9(2).

W. Croft. 2000. *Explaining Language Change: An Evolutionary Approach*. Longman Linguistic Library.

B. de Boer. 2001. *The Origins of Vowel Systems*. Oxford University Press.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6:182–197.

M. Dras, D. Harrison, and B. Kapicioglu. 2003. Emergent behavior in phonological pattern change. In *Artificial Life VIII*. MIT Press.

David E. Goldberg. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

M. Hare and J. L. Elman. 1995. Learning and morphological change. *Cognition*, 56(1):61–98, July.

M. D. Hauser, N. Chomsky, and W. T. Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298:1569–1579, 11.

John H. Holland. 1975. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor.

Jinyun Ke, Mieko Ogura, and William S-Y. Wang. 2003. Modeling evolution of sound systems with genetic algorithm. *Computational Linguistics*, 29(1):1–18.

S. Kirby. 1999. *Function, Selection and Innateness: the Emergence of Language Universals*. Oxford University Press. The full-text is only a sample (chapter 1: A Puzzle of Fit).

S. Kirby. 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Anthony Kroch. 2001. Syntactic change. In Mark baltin and Chris Collins, editors, *Handbook of Syntax*, pages 699–729. Blackwell.

P. Niyogi. 2006. *The Computational Nature of Language Learning and Evolution*. MIT Press, Cambridge, MA.

J. Ohala. 1993. The phonetics of sound change. In C. Jones, editor, *Historical linguistics: Problems and perspectives*, page 237278. Longman, London.

A. Perfors. 2002. Simulated evolution of language: a review of the field. *Journal of Artificial Societies and Social Simulation*, 5(2).

T. Poibeau. 2006. Linguistically grounded models of language change. In *Proc. of CogSci 2006*, pages 255–276.

Melissa A. Redford and R. L. Diehl. 1999. The relative perceptibility of syllable-initial and syllable-final consonants. *Journal of Acoustic Society of America*, 106:1555–1565.

Melissa A. Redford, Chun Chi Chen, and Risto Miikkulainen. 2001. Constrained emergence of universals and variation in syllable systems. *Language and Speech*, 44:27–56.

Melissa A. Redford. 1999. *An Articulatory Basis for the Syllable*. Ph.D. thesis, Psychology, University of Texas, Austin.

L. Steels. 1997. The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34.

W. S-Y. Wang, J. Ke, and J. W. Minett. 2005. Computational studies of language evolution. In *Computational Linguistics and Beyond: Perspectives at the beginning of the 21st Century, Frontiers in Linguistics 1. Language and Linguistics*.