International Journal of Humanities and Social Science Invention
ISSN (Online): 2319 – 7722, ISSN (Print): 2319 – 7714
www.ijhssi.org Volume 4 Issue 5 || May. 2015 || PP.21-42

Machine Generated Word Formation Framework: An Extension to BanglaGen.vbp

APARUPA DASGUPTA
Centre for Development of Advanced Computing, Pune, India

ABSTRACT: *This is an endeavor to extend my PhD thesis (awarded in the year 2001) "A Morphological Generator of Bangla: A Study on BanglaGen.vbp". My thesis was dedicated to a morphological generator for Bangla developed on Visual Basic 5.0 the then to synthesize the derivatives of the language. Bangla Derivative Generator was designed and executed to generate existing and possible derivatives of Bangla Word Formation process in terms of base / root forms with affixation on proposed level based matrix of:*

1) *(prefix) + prefix + root*
2) *(prefix) + prefix + root + suffix + (suffix) + (suffix)*
3) *root + suffix + (suffix) + (suffix)*

This is a derivational generative NLP utility of Bangla. This endeavor extends the present NLP generation framework into a syntactic and semantic layered feature propositions culminating to an annotated Bangla derivatives. Incrementally, these derivatives can be mapped to multilingual equivalents with associated information on: (i) syntactic and semantic category and (ii) language identification on the basis of morphological or syntactic classification. The extended framework can be implemented over analysis and generation of Bangla word formation process (WFP). The extended framework can be labeled with annotations tagged with [base] or [root] and [-affix-] segmented with morphological, syntactic and semantic information about (i) classifiers, (ii) case-markers or post-positions / parsargas, (iii) latent features of prefix and suffix (iv) suffix allomorphy, (v) POS tag and (vi) sense tag. Bangla WFP with an analytical segmentation for 'Analyzer' pre-empting to existing 'Generator'. This is an independent module developed to synthesize the derivates of Bangla. Induced concept of incremental lexicon is a framework packaging for Indian language technology on Machine Translation, Information Extraction & Retrieval or Sentiment Analysis or other aspects of NLP, an addition of dynamic component to derivational morphology. This Paper was presented in 36th International Conference of Linguistic Society of India (ICOLSI 36) during 1-4 December, 2014 in Thiruvananthapuram, India. In post-ICOLSI-36 duration, present paper is expanded for Incremental Lexicon within scope of the framework.

KEY WORDS: *PhD, VB, GUI, NLP, OPF, WFP, POS, MT, SA&DA, IE&IR, SA&DA, NLG, TAM, GNP, COSMAS, POS, KWAL, KWIC, HE, TE, SL, TL and ICOLSI.*

I. An Introduction to BanglaGen.VBP

Bangla Derivative Generator developed independently by the researcher with the supervision of Prof Udaya Narayana Singh for PhD (awarded in the year 2001) was created on window platform and was executed on Visual Basic 5.0 version. Initially, this was the part of Bangla-Hindi *morph* for Anusaaraka where existing rules of literature of traditional Bangla morphology tested for producing acceptable output matching with Anusaaraka proposals [*cf.* Dasgupta and Singh (1997)]. Eventually, BanglaGen.vbp developed into computational aspect on Bangla derivative generation, software that triggers actual and possible word formation process of the language.

Both regular word formation and neologism was the prime aspect of the derivative generator within a provided condition through matrix of Root and Affix mapping regulating the over or under generation. The concept of matrix for (i) prefix+root, (ii) root+suffix and (iii) prefix+root+suffix is a natural phenomenon of derivation morphology that has resulted into an incremental level of derivative generation. Following diagram depicts an overview of the link between the GUI and database of BanglaGen.vbp:

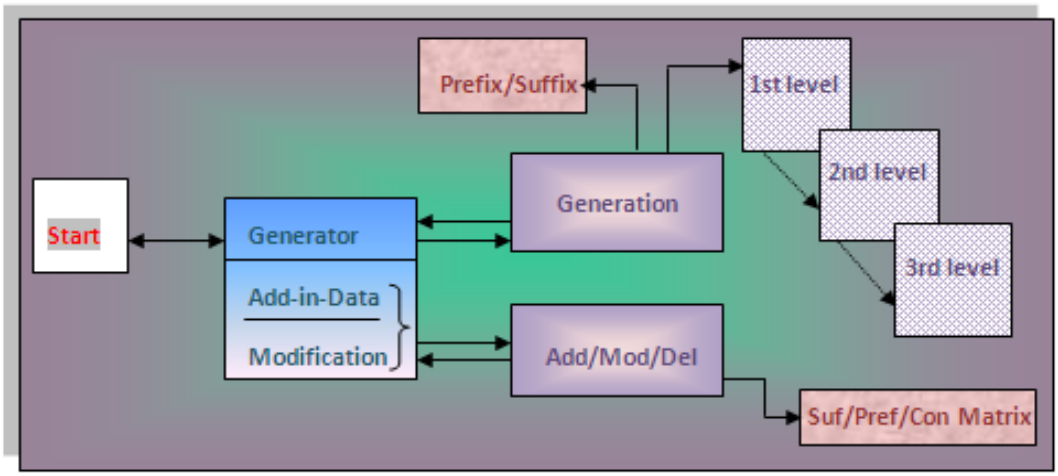

Figure 1: GUI and Database Link of BanglaGen.vbp

The above generator not only generates the Bangla derivatives but also governs the computing aspect of morphology of word formations rules. The concept of matrix as described earlier was supervised by conditions to control the over or under generation. Following example explains the affixation process of matrix in a best way possible for derivational morphology of Bangla. Let us consider an example:

Matrix for 1st level generation: (i) prefix+root, (ii) root+suffix and (iii) prefix+root+suffix
Condition: [+] actual/possible; [-] not actual/possible:

Table 1: Description of matrix, condition and database design of BanglaGen.vbp

Matrix	Root	Prefix	Suffix	Condition [±]	Derivative	Transcription	Gloss
(0)	পত্নি	0	0	0	0	patni	wife
(i)		বি	0	-	*বিপত্নি	-	-
(ii)		0	িক	-	*পত্নিক	-	-
(iii)		বি	িক	+	বিপত্নিক	bipatnik	widowed husband
(iv)		স	িক	+	সপত্নিক	saptnik	with wife

The three basic component of concept of matrix is brought within one word-group where (i) prefix matrix is conditioned with second level generation, (ii) root or stem is executed for noun, pronoun, adjective, verb etc. and (iii) suffix matrix conditioned with third level generation. The derivative generator culminates to word formation strategies for word group, generation of derivative forms, incremental classification and concept of matrix matching. Following is the login screen of BanglaGen.vbp:

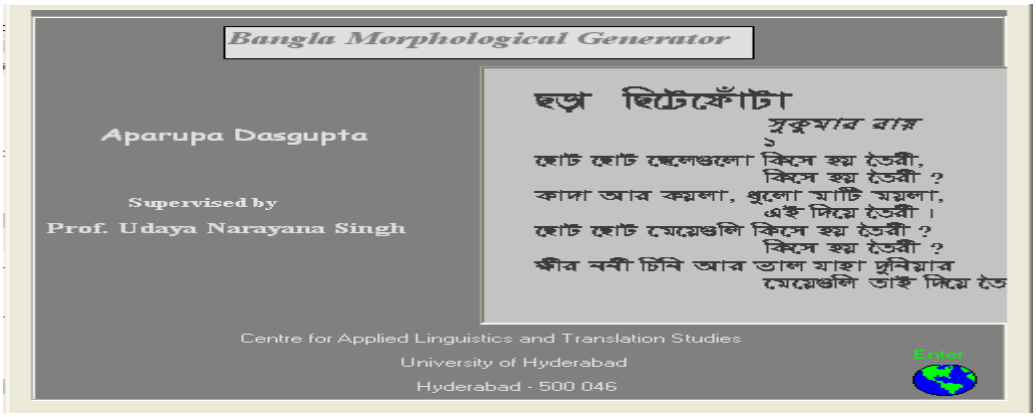

Figure 2: Login screen of BanglaGen.vbp

1.1 Concept of Matrix of BanglaGen.vbp

The verbal or non-verbal forms in natural language processing are bounded by an inter-relationship where synthesis or analysis of a morpho-syntactic structure within a defined context is regulated by various linguistic formalisms, theories and rules. In Indian language, synthesis of verbs is governed by kaaraka formalism (for example in Hindi) and non-verbals like adpositional words such as, adjective, post-position (parsarga) and particles (avyaya) are part of modified-modifier relationship [cf. Kellog, 1857]. The aspect of programming while computing the natural language for generated outputs correlates to the phonological, morphological, syntactic or semantic theory with various argument, agreement and features [cf. Sproat, 1991]. In Indian languages (from Paninian perspective), any generator (i.e., a multiplicative process) and analyzer (i.e., a deductive process) depends on the karaka assignments with demands (i.e. aakaankshaa) and merits (i.e. yogyataa) to words in a syntax. In matrix based concept for word level formation, the framework for derivative generator in BanglaGen.vbp will be explained in the following sub-sections. The nominal forms in Bangla can be understood as: *[Stem + Case Morpheme + (Classifier) + (Emphatic Marker)]*. Wherein, these forms occur on its own, with case morpheme and the later sequences are optional. Further the nominal forms of Bangla posses both syntactic and semantic features such as, [±animate], [±human], [±count], consonant ending [±low vowel], vowel ending and single syllable. Prevalently, Bangla is classifier based language invigorating various dialectal and, diglossic and literary forms both in written spoken varieties.

1.2 Pronominal Matrix (with GUI)

Classification for pronominal matrix in BanglaGen.vbp is as follows in tables 2(i) and 2(ii):

Table2 (i): Pronominal matrix of BanglGen.vbp

Root / Stem	Changed Root / effected Root[1]	Transcription	Gloss
আমি	আমা	aami~aama	I
তুমি	তোমা	tumi~tomaa	You (equal)
তুই	তো	tui~to	You (less / proximity / inferior)
আপনি	আপনা	aapni~aapanaa	You (honorific)
সে	তা	se~taa	He /she

Table 2(ii): Pronominal matrix of BanglGen.vbp

Suffix Forming Matrix	Transcription	Feature
কে	Ke	Accusative singular suffix
দেরকে	Derke	Accusative plural suffix
র	Ra	Genitive singular suffix
দের	Dera	Genitive plural suffix
তে	Te	Locative suffix

1.3 Other Matrices: Nominal, Adjectival, Gerundial and Non-Finite forms

Nominal matrix is executed for three levels of derivative generation. Suffix forming matrix is approximately 125 in number, whose few excerpts are provided in this section with examples and their descriptions:

Consider the 'Start up suffix' forming matrix as provided in the following table:

Table 3: Start up suffix matrix of BanglGen.vbp

Suffix Category	Examples with Transcription
Nominal forming suffix matrix	ক 'ka', ওয়ালা 'oyaalaa', মি 'mi', তা 'taa', ত্ব 'twa',
	পনা 'panaa', িত 'ita'.
Gerund forming suffix matrix	ানো 'aano', া 'aa'
Adjective forming suffix matrix	টে 'Te', চে 'che', তি 'ti', ীয় 'iiya', শীল 'shiila', ীন 'iina', িয়ে 'iye', াল 'aala', ুক 'uka', িক 'ika', টিত 'Tita',
	ময় 'maya', ঈয় 'iiya'

Suffix forming 1st increment matrix is provided in the following table:

[1] Base Root form for computing the affixation process (both for inflection and derivation)

Table 4: 1st increment matrix of BanglGen.vbp

Suffix Category	Examples with Transcription
Nominal forming suffix matrix	ক ‘ka’, ওয়ালা ‘oyaalaa’, মি ‘mi’, তা ’taa’, ত্ব ‘twa’,
	পনা ‘panaa’, িত ‘ita’.
Gerund forming suffix matrix	ানো ‘aano’, া ‘aa’
Adjective forming suffix matrix	টে ‘Te’, চে ‘che’, তি ‘ti’, ীয় ‘iiya’, শীল ‘shiila’, ীন ‘iina’, িয়ে ‘iye’, াল ‘aala’, ুক ‘uka’, িক ‘ika’, টিত ‘Tita’,
	ময় ‘maya’, ঈয় ‘iiya’

Suffix forming 2nd increment matrix is provided in the following table:

Table 5: 2nd increment matrix of BanglGen.vbp

Nominal	Transcription	Adjective/indeclinable	Transcription	Gerund/non-finite verb form	Transcription
1st Increment					
-ন্ত	anta			ানো	aano
		-ওয়া	oyaa	-া	aa
-মি, -পনা	mi, panaa				
		-শীল, -হীন	shiila, hiina		
		-শীল, -বান	shiila, baana		
2nd increment					
-ন , -ুনি	na, uni			-ানো	aano
-ানি,		-ো		-ানো	aano
3rd Increment					
-বাদ, মান	baada, maana	-শীল, -িত	shiila, ite		
-হানি, -দ	haani, da	-বান, -হীন	baana, hiina		
4th increment					
-মি, -মো	mi, mo	-টে	Te	-ানো	aano

Above few tables explained the concept of increment matrix in BanglaGen.vbp and actual database link in the generator are described for Root+Affix matrix designed with few features to accommodate data addition facility in the following GUI:

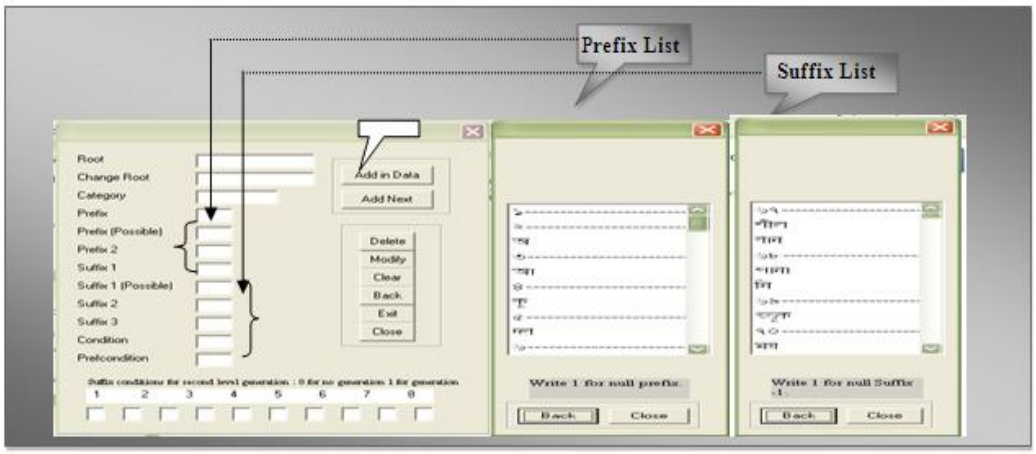

Figure 3: GUI display for Prefix and Suffix list and database connectivity of BanglaGen.vbp

Above set of affix matrix (i.e. incremental) are generated on the basis of three levels of generation according to Bangla derivation morphology:

Level 1 generation for *root+suffix*, consider the example:

1) root - বন্ধু ‘bandhu’; suffix – ত্ব ‘twa’; actual or existing derivative - বন্ধুত্ব ‘bandhutwa’, ‘friendship’
2) root - আসমান ‘aasmaana’; suffix - ী ‘ii’ ; actual or existing derivative – আসমানী ‘aasamaanii’, ‘sky like / skyish / sky’

Level 2 generation for *(prefix1)+(prefix2)+root+suffix1+suffix2*, consider the example:

1) Prefix 2- অন ‘ana’; Prefix1-উ ‘u’; Root–লঙ্ঘন ‘langhana’; suffix-ীয় ‘iiya’; actual or existing derivative – অনুলঙ্ঘনীয় ‘anulanghaniiya’, ‘not to disobey / no crosssing-over’
2) Root–কর্ম; suffix 1- শীল ; suffix2 - তা ; actual or existing derivative – কর্মশীলতা ‘business/fabulousness’

Level 3 generation for *(prefix1)+(prefix2)+root+suffix1+suffix2+suffix3*, consider the example:

1) Root-অংশ ; suffix1: ী ; suffix2: দার ; suffix3:ী; actual or existing derivative – অংশীদারী

Following description is of level 1 derivation through GUI of BanglaGen.vbp:

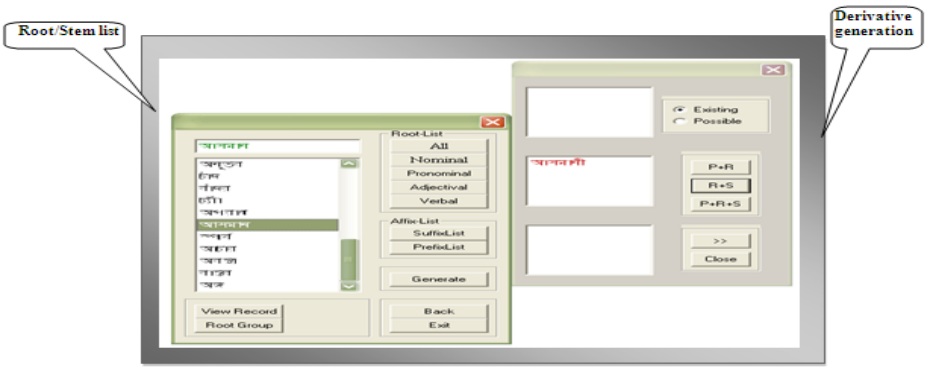

Figure 4: GUI display of Level 1 derivative generation of BanglaGen.vbp

1.4 GUI and Database of BanglaGen.vbp

The derivative generator ‘BanglaGen.vbp’ is independent software that can be executable for machine translation program or other NLP applications. Generated outputs of the software are conceptualized in a manner that both existing and possible word forms of modern or chaste Bangla usages can be interpolated with NLP program. Representation of the software is such that the incremental screens and display has connectivity (OPF – object, procedure and function) to the database designed for three levels of generation for verbal and non-verbal forms. Database is designed according to the concept of matrix for derivative generation. The interconnectivity, sequence, rule ordering and POS tagged data are thoughtfully developed with a flexible feature of data addition (for root/stem, prefix with 2 levels and suffix with 3 levels) option to BanglaGen.vbp. Root/stem, prefix list or suffix list display of BanglaGen.vbp are in a ‘list-box’ and the feature for data addition are in ‘text-box’ with direct type-enter option in Bangla font. A total of 125 suffixes are considered in the thesis. The matrix for derivative generation is accessed through *‘Prefix (P) – Root (R) - suffix (S)’* matrix internally mapped with conditions for Bangla derivational morphology. Maximally generated matrix is:

$$\left\{\begin{array}{l} \text{R+S} \\ \text{P+R(or changed Root)} \\ \text{P+R(or changed Root)+S+} \\ \text{P+R(or changed Root)+S1+S2} \\ \text{P1+P2+R(or changed Root)+S1+S2} \\ \text{P1+P2+R(or changed Root)+S1+S2+S3} \\ \text{R(or changed Root)+S1+S2+S3} \end{array}\right\}^{2}$$

[2] (a) R is main root or change root due to phonological process (b) P1, P2 are increment prefix and (c) S1, S2, S3 are increment suffix of the matrix of derivational morphology.

Above matrix generation are enabled in the database with field name, data type and field properties. Few examples of such properties in the database are displayed:

Table 6: Properties of Root/Stem and Matrix in the database of BanglaGen.vbp

(i) Properties of Root / Stem

Fields Name	Data Type	Field Properties	
MainRoot, ChangeRoot, Category	Text	Field Size	50
		Format	
		Allow Zero Length	Yes
		Input mask	
		Caption	
		Validation Rule	
		Validation Text	
		Required	No
		Default Value	
		Indexed	No(No Duplicates)

(ii) Properties of Matrix with Condition

Fields Name	Data Type	Field Properties	
Pref, Suff, Suftwo, Sufthree, PrefPos, PrefPot, SufPos, SufPot and Condition	Number	Field Size	Long Integer
		Format	
		Decimal Places	Auto
		Input mask	
		Caption	
		Validation Rule	
		Validation Text	
		Required	No
		Default Value	0
		Indexed	No

In-depth description in earlier sub section on data conceptualizing, designing and accessing with actual instances from the language are logically translated into algorithm to deal with computational and linguistic correlative cognition. The dynamic outputs of the derivatives are implemented with following viewpoints:

1) Interlink of Bangla root, prefix and suffix through database design.
2) Rules and Conditions bounded by Bangla derivational morphology through a framework.
3) Framework on derivative generation is executed through coding on Visual Basic 5.0.
4) Dynamic and relevant coding exuberating three levels of generations.
5) Three levels of generations are based on norms of productivity of Bangla Word Formation.
6) Open ended (flexible and dynamic) data and on-line addition for enriching the framework of Bangla Generator.
7) A practical system on Bangla Word Formation process correlating to computational aspects.

In the subsequent sections, the additional or extended factors on concept of derivative generation will be discussed. This is an endeavor to extend the existing concept and design of matrix and data of BanglaGen.vbp on parlance to Indian Language computing, as time progressed. Considering the various developments on the domain of analyzer and synthesizer in India (specifically on Bangla), with culmination of syntactic and semantic feature, an extended framework is put forward to overcome the factual problem of various parameters of NLP (be it decade old machine translation, information extraction and retrieval or recent program on sentiment or discourse analysis on Indian languages). [cf. Ekbal et al, 2007; Dandapat et al, 2007]. The *'analyzer cum generator'* for BanglaGen.vbp is discussed in subsequent sections.

II The Extended Framework of BanglaGen.VBP

Essentially, the extended framework is a *gap spotting* to reciprocate to the demand on South Asian research and study (specific to computational morphology) [cf. Alvesson et al, 2011]. This framework has syntactic and semantic layered feature propositions that are appended culminating to an annotated Bangla derivatives and inflections. In compliance to preceding generation, derivatives and inflections can be mapped to multilingual equivalents with associated information on: (a) syntactic and semantic category and (b) language class or group or family identification on the basis of morphological or syntactic classification. The extended framework can be labeled with annotations tagged with [base] or [root] and [-affix-] segmented with morphological, syntactic and semantic information about (i) classifiers, (ii) case-markers or post-positions/parsargas, (iii) latent features and stylistics of prefix and suffix, (iv) suffix allomorphy, (v) POS tag and (vi) Sense tag. In addition to analytical and generative dimensions, practically induced concept of incremental lexicon is a framework feature for Indian language technology on Machine Translation (MT), Information Extraction & Retrieval (IE&IR) or Sentiment and Discourse Analysis (SA&DA) or other aspects of Indian language processing.

Following chart represents the framework sequence of BanglaGen.vbp:

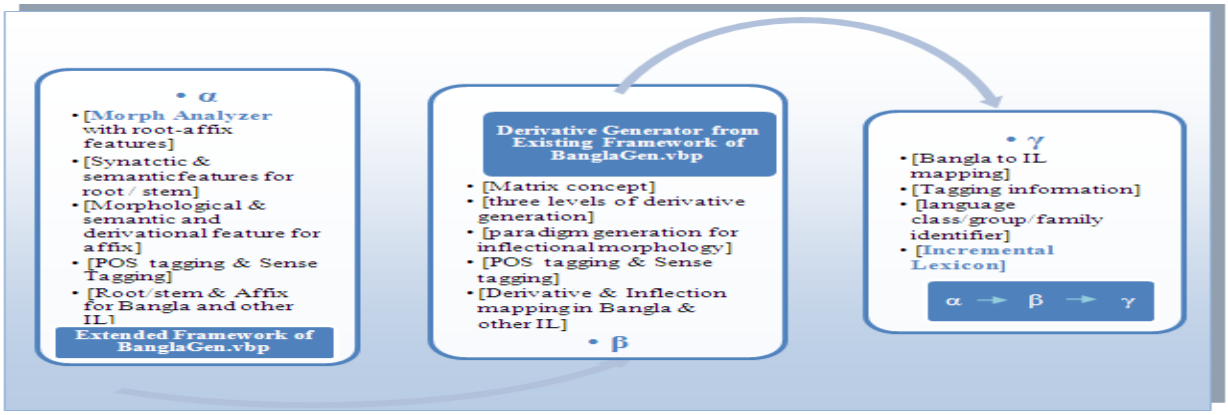

Figure 5: Proposed Framework of Bangla Analyzer cum Generator based on BanglaGen.vbp

Above sequence exposition can executed to the existing generative model in Bangla language processing. This can excel to (i) annotation (syntactic and semantic), (ii) morpho-syntactic labeling and (iii) lemmatization and affixation, (iv) derivative generation and derivative mapping with other Indian languages. This model will suffice to an incremental multilingual lexicon, thus completing all associative features implied to Natural Language Processing and Generation. Thus, a multilingual based NLP for *'Generators'*, [cf. Dasgupta & Dubey, 2010]. Various language processing based tools and collators for English are present as an open source also, (see COSMAS online accessible from 1995). NLP utilities on English morpho-syntactic base are available in standardized form for corpus concordance tool, tagger (pos, sense or emotion), sentence type identifier, stemmer, wordnet, semantic network (see framenet) etc. While processing diversified Indian languages in current era, it becomes obligatory to create a trend towards creating corpus driven language processor for annotations (i.e., corpus, discourse, prosodic etc.) to frequency count (i.e., KWAL and KWIC) , to chunker and disambiguator, to lemmatizer, so on and so forth [cf. J Sinclair, 1994 & 2004]. Complementary and contrastiveness of Indian language has underlying feature-rich trend that evolves out of the research and development of Indian languages [cf. Krishnamurti et al, 1986]. Assumptions and theories for evolved methodologies have culminated the gap spotting in extended framework of BanglaGen.vbp.

2.1 Workflow of the Extended Framework for Analysis and Generation

Morphological processing (considering general corpus) for Bangla can be implemented through the extended framework. Subsequent stages for implementing the framework are summarized in the sub-sections. Following is the sequence depicted for workflow of extended framework for Bangla mapping to other Indian languages:

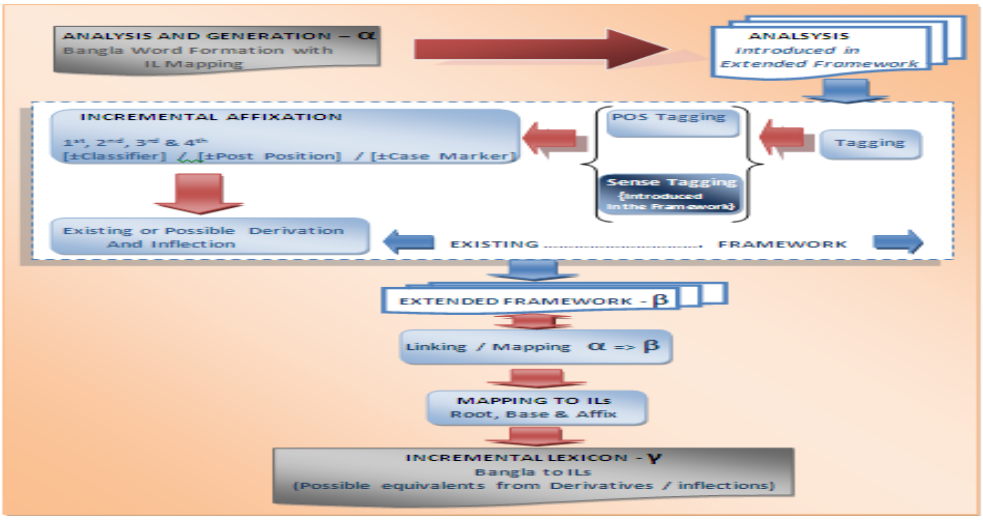

Figure 6: Workflow of extended Framework

2.1.1. Underlying Workflow for Morphological Analysis in Extended Framework

The stages of morphological analyzer shall experience concentrated deductive process with deep-level feature structure and annotation, such as:

1) Root / Stem identification,
2) Annotating (i.e. POS tag) the root and stem,
3) Paradigm generation for inflectional affix,
4) Marking the latent feature of prefix or suffix or the affix paradigm (i.e., intensifier or negation, classifier, post-position, case morpheme, GNP, TAM, transitivity, causativeness etc), (e) annotating the POS for derivational suffix or prefix,
5) Mapping Root / Stem from Bangla to other Indian languages, and
6) Mapping suffix and prefix from Bangla to other Indian languages.

Few examples with description are explained in next section of 3.0 (3.1 and 3.2).

2.1.2 Underlying Workflow for Derivative Generation in Extended Framework

The multiplicative process of derivative and inflection generator shall undergo matrix based synthesis of actual and possible Bangla word formation for:

1) Derivative generation according to matrix with condition,
2) Inflection generation according to paradigm,
3) Annotating the derivatives and inflection for POS and Sense tagging,
4) Specific identification for
   i. Poetic or dialectal form
   ii. Suffixal allomorphy,
5) Actual and possible generation of the language, and
6) Mapping the derivatives or inflections from Bangla to other Indian languages.

Few examples with description are explained in next 3.3 of 3.0 sections.

In addition to above workflow for extended framework, a concept of incremental lexicon is concatenated with: (a) generated and inflected forms mapping from Bangla to other Indian language mapping. Analytical data organization is explained on incremental lexicon in subsequent section 4.0.

2.2. Encoding of Tag Set and Features Vector in Extended Framework

Feature encoding and POS tagging is a vital parameter in language processing for context and discourse extraction and disambiguation. Corpus based approaches for NLP has yielded success in era of 21st century. For, any machine translation system (in example based, rule based or statistical based MT) morpho-syntactic and semantic tagging becomes essential for extracting disambiguated cross-lingual information. POS tagging is not only imperative for root/stem but also for affixes. Indian language tagset should transcend the morpho-syntactic and semantic tag set into integrated scenario of POS tag set. In such context, sense tagging is essential to explain the emotion of the pre-text. Consider an example from an article on POS tagging [cf. Dasgupta and Dubey, 2010].

1) khelaa dhulo karo tumi? 'Do you play?'
   RDP V PR
2) khelaa phelaa cheDhe ebaara paDhaate mona dao 'Stop playing now concentrate on studies'
   RDP V RB V N V
3) dayaa kore edike aashun 'Please come here'
   V RB V
4) dayaa karo, tumi aar ke~ndo naa 'Give a break don't cry more'
   V_DIST PR V V

In Bangla, *khelaa dhulo* and *khelaa phelaa* are reduplicated word meaning 'play' and distinction between these lexical forms are found in sense annotation where later connotes anger or frustration and former does not. Similarly, *daya kore* and *daya karo* are verbal forms meaning 'please'. These forms are disambiguated through sense marking. Former connotes request and later anger or disgust. In order to deal with such derivations in language, Ekman's six set of sense tagging can be introduced with morpho-syntactic tagging [cf. Ekman, 1993]. The six basic emotion types of Ekman's classification are: happiness, sadness, anger, fear, surprise and disgust.

Morpho-syntactic tagset are from Indian languages tag set [see. Bhaskaran, 2008] and IIIT Tagset guidelines (of Akshar Bharti). 12 categories that are identified as the universal categories for the Indian languages from the common tagset framework are introduced in this extended framework of BanglaGen.vbp. First level tag sets are:

Table 7 (a): Bhaskaran's Tagset

[N] Nouns	[V] Verbs	[V] Verbs	[JJ] Adjectives	[RB] Adverbs	[PL] Participles
[PP] Postpositions	[DM] Demonstratives	[QT] Quantifiers	[RP] Particles	[PU] Punctuations	[RD] Residual

After a first level annotation, one level deep tagging that has been designed by IIIT Hyderabad are introduced in this framework. These Tags are:

Table 7 (b): IIIT Tagset

[WQ] Question Words	[QC] Cardinals	[CL] Classifiers	[INTF] Intensifiers	[INJ] Interjections
[NEG] Negative	[C] Compounds	[RDP] Reduplication	[ECH] Echo words	

In continuation of above discussed tag sets, introduction of feature vector in extended framework is major aspect in BanglaGen.vbp. Due to induction of both analysis cum generation, it is complementary that both root and affixal disposition prevalent in a morpho-syntactic and semantic context are resolved. So that cross lingual mapping and extractions can be more descriptive and disambiguating in future. Basic classification of feature is distributed among: (i) prefix and suffix for morpho-syntactic and semantic feature and tagging, (ii) POS and Sense annotation and Feature vector for root/stem and derivatives. Consequent section 2.3 will describe the overall distribution of feature in general.

2.3 Overall Sketch of Extended Framework of BanglaGen.vbp

The salient aspects of feature vector in this framework are:

1) Feature for prefix denoting intensity and negation
2) POS information for derivational suffix
3) POS and sense tagging for derivatives or inflections
4) Feature vector for suffix denoting TAM, GNP, (±)finite form, (±) transitivity, ditransitivity, (±) causitivity, allomorphy, information on classifier, post-position and case morpheme

Examples are explained in subsequent sections from 3.0 to 3.4. Following is the overall glimpse of extended framework on annotation and feature vector induction to analysis cum generation in extended BanglaGen.vbp:

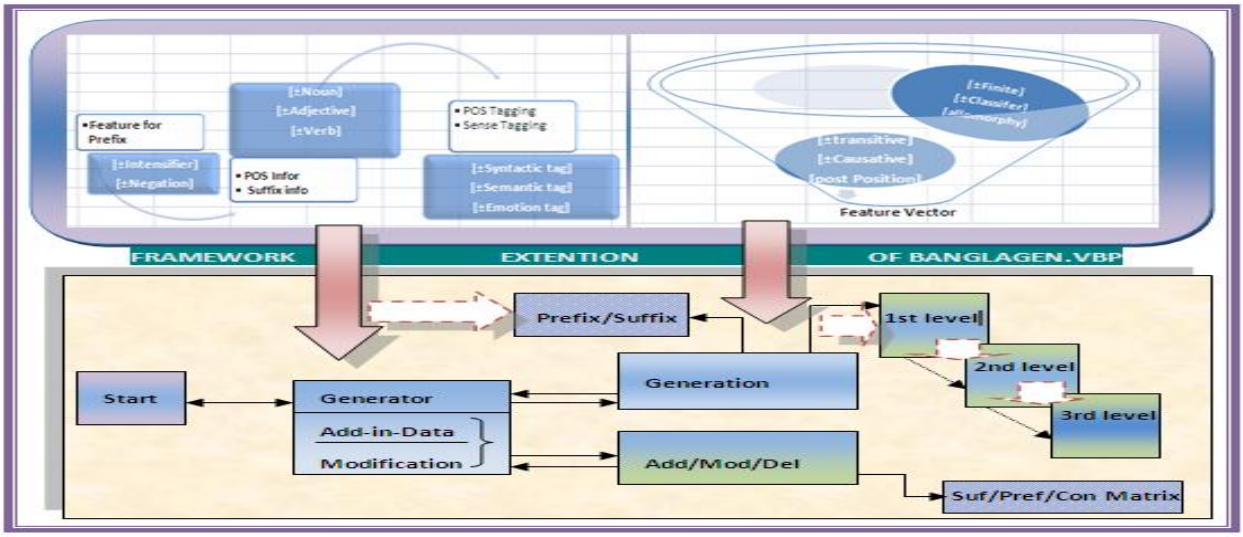

Figure 7: GUI of the proposed framework of extension to BanglaGen.vbp

III. Database Design and Mappings / Links for Multilingual BanglaGen.vbp

Extended framework of BanglaGen.vbp is representation of feature proposition prevailed in Indian languages. The Root + Affix matrix of BanglaGen.vbp is extended in one hand to feature matrix for derivative generation and in other hand to inflection generation. Basic three levels of feature matrix are induced to three levels of generation to the existing framework as:

1) Feature for Affix
2) POS (inflection and derivation) for Affix
3) Feature Vector (language divergence, stylistics and class/family) for Root / stem and Affix
4) Introduction of Sense (emotion) tagging
5) Language class / family identification
6) Mapping Indian languages (classifier based language and inflection based language)
7) Introduction of Collator for analysis and generation with links
8) Binding actual and possible word formation process in Indian languages

In concurrent sub-sections each of these characteristics of extended framework has been described along with excerpts from languages.

3.1 Nominal Analysis and Synthesis

Languages considered in multilingual frame are Bangla, Hindi and Odia. Bangla and Odia is classifier based language and Hindi is inflection based language. Links are established for affix mapping and derivative / inflection mapping. Consider the example with feature expansion in the following table:

Table 8: Affixation process for nominal

Example	Equivalent (H)[3]	Equivalent (O)[4]	POS	Sense	Classifier (B)[5]	Classifier Link (H)	Classifier Link (O)	Case ending (B)	Post-position Link (H)	Classifier Link (O)	Vibhakti	Derivative / Inflection (B)	Derivative /inflectionMapping(H)	Derivative / inflection Mapping (O)
ঘর (ghara)	घर (ghara)	ଘର (ghara)	NOUN	0	টা (Taa), টি (Ti)	०	ଟା (Taa), ଟି (Ti)	0	0	0	Nom_Sg	ঘরটা (gharaTaa)	घर	ଘରଟା (gharaTaa)
												ঘরটি (gharaTi)		ଘରଟି (gharaTi)
					গুলো (gulo), গুলি (guli)	ो, े (plural marker)	ଗୁଡ଼ିକ (guDhika)	0	0	0	Nom_Pl	ঘরগুলো (gharagulo)	घरों, घरें	ଘରଗୁଡ଼ିକ (gharaguDhika)
												ঘরগুলি (gharaguli)		

[3] (H) – Hindi equivalent, derivative, inflection, prefix or suffix
[4] (O) – Odia equivalent, derivative, inflection, prefix or suffix
[5] (B)- Bangla root /stem, derivative, inflection, prefix or suffix

								এর (er), ের (er), র (ra)	का, के, की	ର (ra)	Gen_Sg, Gen_Pl	ঘরের (ghar er)	घर का, घर के, घर की	ଘରର (ghar ara)
								এতে (ete), েতে (ete), তে (te), ে (e), এ (e)	में	ରେ (re)	Gen_Sg, Gen_Pl	घरे (ghar e)	घर में	ଘରେ (ghar e)

Examples are:

ঘর (in Bangla), घर (in Hindi), ଘର (in Odia) 'home'

টা/ টি/টে (in Bangla), ଟା/ ଟି (in Odia) 'classifier-singular or emphatic marker'

গুলো/ গুলি (in Bangla), ଗୁଡ଼ିକ (in Odia) 'plural marker'

का/ के /की (in Hindi) 'possessive marker'

में (in Hindi) 'post position'

3.2 Verb Stylistics

In 3.2 section, languages considered are Bangla, Hindi and Odia. Bangla, Odia and Hindi are convergent morphological process in root/stem forms irrespective of the divergence between classifier or inflection based language. Only Bangla experiences suffix allomorphy among other two languages. Links are established for affix mapping and inflection mapping. Consider few examples with feature expansion in the following table. Examples are:

Root / stem:

নে (in Bangla), ले (in Hindi), ନେ (in Odia) 'accept ,take'

যা (in Bangla), जा (in Hindi), ଯା (in Odia) 'go'

Suffix:

ছিলি/ছিলিস (in Bangla; suffix allomorphy) रहा था/ रही थी (in Hindi), ଇଛି (in Odia)

Table 9: Derivative / Inflection generation with affix allomorphy and verb stylistics

Example (B)	Equivalent (H)	Equivalent (O)	Category /Feature	Inflectional ending - (Allomorphy, TAM, GNP, Causativity)	Inflected Forms

Root (B)	**Changed Root - (B) (Vowel harmony)**	**Root (H)**	**Changed Root - (O) (Vowel harmony)**	**[±Finite], [±Trans/ Ditrans]**	**TAM-g[6] [allomorphy], [±cause]**	**TAM-g Link (H)- [±cause/ dicause], [GNP]**	**TAM-g Link (O) – [±cause], [GNP]**	**TAM-c[7] (B) – [±cause], [GNP]**	**TAM-c (H) – [±cause], [GNP]**	**TAM-c (O) – [±cause], [GNP]**	**TAM-p[8] - [±cause], [GNP]**	**TAM-p (H) [±cause], [GNP]**	**TAM-p (O) [±cause], [GNP]**	**Inflection - (B)**	**Inflection link (H)**	**Inflection Link (O)**
ধো (dho)	ধু (dhu)	ধো (dho)	ধো (dho), ধু (dhu)	VERB	য়েছ, (yecha), য়েছো (yechho)	যা, ই, এ, এ, যে	0	0	0	0	0	0	0	ধুয়েছ, ধুয়েছো	ধোয়া, ধোই, ধোএ, ধোয়ে	
নে (ne)	–	লে (le) লু (lu)	নে (ne) (ni)	VERB	ব (ba), বো (bo)	ংগা ,ংগী, ংগে	ব (ba)	0	0	0	0	0	0	নেব, নেবো	লুংগা, লুংগী, লেংগে	নেব
যা (ja)	–	জা (ja)	যা (ja)	VERB	ছিলি (chhili), ছিলিস (chhilish)	রহা থা, রহী থী	ইচ্ছি (ichi)	0	0	0	0	0	0	যাছিলি, যাছিলিস	জা রহা থা, জা রহে থে, জা রহী থী	যাইচ্ছি
বল (bal	–	বোল (bol	বল	VERB	লাম	া,	ইলা(ilaa)	লুম (lu		0	লেম (le	0	0	বল	বোলা,	বলি

[6] TAM-g is suffix or prefix allomorphy
[7] TAM-c is colloquial form of suffix or prefix
[8] TAM-p is poetic variation of suffix or prefix

)		a)	(bol) କହ (ka ha)		(la a m a)	ି, ୋ		ma)			ma)			ଲା ମ, ବ ଲ ଲୁ ମ, ବ ଲ ଲେ ମ	ବୋଲେ , ବୋ ଲି, ବୋଲୁ	ଲା, କହ ଥିଲା
ଜାନ (ja an)		ଜାନ (jaa n), ପତା (pat aa)	ଯା (jaa) ଯି(ji)	VER B	ତା ମ (ta a m a)	ତା-ଥା, ଥା ; ତୀ-ଥୀ, ଥୀ ; ତେ-ଥେ, ଥେ ;	ଜାଣିଥିଲି	ତୁମ (tu m)	ୁ	0	ତେମ (te m)	0	0	ଜା ନ ତା ମ, ଜା ନ ତୁ ମ, ଜା ନ ତେ ମ	ଜାନ ତା ଥା, ଜାନ ତୀ ଥୀ, ଜାନ ତେ ଥେ	ଜାଣି ଥିଲି

3.3. Three Levels of Generation

According to the framework of BanglaGen.vbp derivative generation has been derived into three levels of matrix. In what follows, the affix and root/stem generation are (i) conditioned by two levels of generation for prefix and (ii) three levels of generation for suffix. Subsequent subsections will explain these 'layers of level' of the framework.

3.3.1 Level - 1 Derivative

Languages considered are Bangla, Hindi and Odia. POS and sense tagging has been introduced as discussed in previous sections. Links are established for affix mapping and derivation mapping till 4th increment. Consider few examples with feature expansion for (Root/Stem+Suffix1) in the following table:

For Example in Bangla:

চল – চলন্ত 'in motion', চলানো 'to run/execute'

বোকা – বাকামি, বোকামো 'idiot-ness'

In Hindi:

दोस्त 'friend'- दोस्ती 'friendship' - दोस्ताना 'friendship (ness)'

Table 10: Level 1 generation of BanglaGen.vbp

Example (B)	**Equivalent (H)**	**Equivalent (B)**	**POS**	**Sense**	**Affixation - Root / Stem + Suffix (1st+2nd+3rd+4th Increment level for derivative generation)**												**Derivative (B)**	**Derivative link (H)**	**Derivative link (O)**
					1st Increment			**2nd Increment**			**3rd Increment**			**4th Increment**					
					Derivational affix (B)	**Affix Link (H)**	**Affix Link (O)**	**derivational affix (B)**	**Affix Link (H)**	**Affix Link (O)**	**derivational affix (B)**	**Affix Link (H)**	**Affix Link (O)**	**Derivational affix (B)**	**Affix Link (H)**	**Affix Link (O)**			
					[±	**[±**	**[±**	**[±**	**[±**	**[±**	**[±**	**[±**	**[±**	**[±**	**[±**	**[±**			

					classifier]	classifier]	classifier]	classifier]	classifier]	classifier]	classifier]	classifier]	classifier]	classifier]	classifier]	classifier]			
					[± post-position]	[± post-position]	[± post-position]	[± post-position]	[± post-position]	[± post-position]	[± post-position]	[± post-position]	[± post-position]	[± post-position]	[± post-position]	[± post-position]			
চল (chalal)	চল (chala)	চ।ল্ (chal)	VERB	anger	ানো (aano)	না (naa), ানা	চালি (chaali), চালনা (chaaLanaa)	ন্ত (anta), ানো (aano)		থন্ত। (antaa)							চালনো	চলনা, চালানা	চালি, চালনা, চালথন্ত।
বোকা (bokaa)	মূর্খ (murkha)	ব৬।ক্া (bokaa)	ADJECTIVE	anger	মি (mi)	তা (taa)	ব৬।ক্া (mi)	মি (mi), পনা (panaa)									বোকামি	মূর্খতা	ব৬।কাব৬।ক্া
পাকা (pakaa)	পকা (pakaa)	x	ADJECTIVE	disgust	পনা (panaa)	পন (pana), ক্ত (U)		মি (mi), পনা (panaa)		পি (mi), পণ (paNa)							পাকাপনা	পকাপন, পকাক্ত	
ন্যাকা (nyakaa)	নখড়া ~ নখড়ে	নখর। (nakh	ADJECTIVE					নি (ni), পনা (panaa)	দার	পি (mi), পণ (paNa)							ন্যাকামি, ন্যাকাপনা	নখড়ে দার	নখরামি, নখরাপণ

		a r a a)																
ঘর (gh ar)	ঘর (gh ara)	সুখ (su kh a)	NOUN	0	ময় (m aya)	ভর	ময় (m ay a)										ঘর ময়	ঘর ভর	সুখ ময়
ফর ফর	চুল বুল (ad hira)	ফর ফর	ECHOWORD	Happiness							ানি (a ani), ানো (a an o), ে (e)	না ,0 ,0					ফর ফরা নি, ফর ফরা নো , ফর ফরে	চু লবু লা না	
অঙ্গ (an ga)	অংগ	অঙ্গ	NOUN											বান (ba an) , হা নি (ha ani), হীন (hii na) , দ (da)	হী ন	বা ন (b aa na), হা নি (h aa ni) , অঙ্গ (hi in a), দ (d a)	অ ঙ্গ বা ন, অ ঙ্গ হা নি, অ ঙ্গ হী ন, অ ঙ্গদ	অং গ হী ন	অঙ্গ বা ন, হানি হানি, অঙ্গ অঙ্গ, অঙ্গ গদ

বন্ধু (bandhu)	বন্ধু	ବନ୍ଧୁ (bandhu)	N	Happiness		তা (taa)		ত্ব (twa)		ତ୍ୱ (twa)								ବନ୍ଧୁତା	ବନ୍ଧୁତ୍ୱ
	দোস্ত					তানা, ী												দোস্তানা, দোস্তী	

3.3.2 Level - 2 Derivatives

In level 2 derivatives, languages considered are Bangla, Hindi and Odia. POS and sense tagging has been introduced according to the previous sub-sections. Links are established for affix mapping and derivation mapping till 4th increment level. Consider few examples with feature expansion for ((Pefix2+((Prefix1+Root/Stem))+Suffix1))/+(Suffix2) in the following table for examples in Bangla, such as:

লঙ্ঘন - উলঙ্ঘন 'crossover' – অনুলঙ্ঘন 'not to cross over' – অনুলঙ্ঘনীয় 'no crossing over'

কর্ম 'work'– কর্মশীল 'working' – কর্মশীলতা 'capability/eligibility'

Table 11: Level 2 generation of BanglaGen.vbp

Example (B)	**Equivalent map (H)**	**Equivalent map (O)**	**POS**	**Affixation - Prefix1+Prefix2+Root/Stem+Suffix1+Suffix2 (Antonymy, Synonymy, Polysemy)**												**Derivative (B)**	**Derivative Link (H)**	**Derivative Link (O)**	**Sense**
				Prefix1	**Prefix link (H)**	**Prefix link (O)**	**Prefix2**	**Prefix link (H)**	**Prefix link (O)**	**Suffix1**	**Suffix link (H)**	**Suffix link (O)**	**Suffix2**	**Suffix link (H)**	**Suffix link (O)**				
				[± intensifier]	[± intensifier]	[± intensifier]	[± intensifier]	[± intensifier]	[± intensifier]	[± Noun]	[± Noun]	[± Noun]	[± Noun]	[± Noun]	[± Noun]				
				[± Negation]	[± Negation]	[± Negation]	[± Negation]	[± Negation]	[± Negation]	[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]				

ল ঙ্ঘ ন (la ngha na)	লাং ঘ ন	ଲଙ୍ଘନ (la ngha na)	N O U N	ଉ (u)	০	০	অ ন (a na)	০	থ ন (a na)	নীয় (ni iy a)	না	নীয় (nii ya)	0	0	0	উলঙ্ঘ ন (ulla ngha n)	লাং ঘনা	ଉଲଙ୍ଘନ (ulan ghan a)	fea r/di sgu st
	লং ঘ ন			উ			অ ন									অনুল ঙ্ঘন (anul angh an)	উলং ঘন , অনু লংঘ ন	অনুল ঙ্ঘন(anula ngha na)	fea r/di sgu st
																লঙ্ঘ নীয় (lang hanii ya)		ଲଙ୍ଘ ନୀୟ (lang hanii ya)	fea r/di sgu st
																অনুল ঙ্ঘনী য় (anul angh aniiy a)		অଲଙ୍ ଘନୀୟ (alan ghani iya)	fea r/di sgu st
কর্ম (ka rm a)	ক র্ম,	କର୍ ମ(kar ma)	N O U N	0	0	0	0	0		শী ল (s hii la)	বী র	ଶୀ ଲ(s hee La)	তা (ta a)	তা	০	কর্মশী ল (kar mash iila)	কর্ম বীর , কর্ম বীর তা	କର୍ମ ଶୀଲ (kar mash iiLa)	ha ppi nes s
																কর্মশী লতা (kar mash iilata a)		କର୍ମ ଶୀଳତା (kar mash iiLata a)	ha ppi nes s

3.3.3 Level - 3 Derivatives

In level 3 derivatives, language considered are Bangla, Hindi and Odia. POS and Sense tagging has been introduced as discussed in previous sections. Links are established for affix mapping and derivation mapping till 4th increment level. Consider few examples with feature expansion for (((Root + Suffix 1) + Suffix 2) + Suffix 3) in the following table. Few Bangla, Hindi and Odia examples, such as:

Bangla:

অংশ ‘part’– অংশী ‘portion’ – অংশীদার ‘partner’ – অংশীদারী ‘partnership’

Hindi:

हिस्सा (base morph) ~ हिस्से (changed morph) 'part' – हिस्सेदार 'partner' - हिस्सेदारी 'partnership'

Odia:

ଅଂଶ 'part' – ଅଂଶୀ 'portion' – ଅଂଶୀଦାର ଅ) 'partner' – ଅଂଶୀଦାରୀ 'partnership'

Table 12: Level 3 generation of Banglaen.vbp

Example (B)	Equivalent (H)	Equivalent (O)	POS	Affixation - Root / Stem + 1st Level Suffix + 2nd Level Suffix + 3rd Level Suffix									Derivative (B)	Derivative Link (H)	Derivative Link (O)	Sense
				Suffix 1	S1 Link (H)	S1 Link (O)	Suffix 2	S2 Link (H)	S2 Link (O)	Suffix 3	S3 Link (H)	S3 Link (O)				
				[± Noun]	[± Noun]	[± Noun]	[± Noun]	[± Noun]	[± Noun]	[± Noun]	[± Noun]	[± Noun]				
				[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]	[± Adjective]				
অংশ (ansha)	হিস্সা - হিস্সে	অংশ (ansha)	NOUN	ী (ii)	0	1 (ii)	দার (dar)	দার	ଦାର (dar)	ী (ii)	ী	1(ii)	অংশ (ansha)	অংশ, ভাগ, হিস্সা	অংশ (ansha)	happiness
	ভাগ				ী			দার			ী		অংশী (anshi)	অংশী, ভাগী	অংশী (anshi)	happiness
	অংশ												অংশীদার (anshidara)	ভাগীদার, হিস্সেদার	অংশীদାର ଅ) (anshidaar)	happiness
													অংশীদারী (anshidarii)	ভাগীদারী, হিস্সেদারী	অংশীদାରୀ(anshidaarii)	happiness

In current section 3.0, all levels of derivative generation are discussed in a multilingual and contrastive perspective (for Bangla, Hindi and Odia). In the following sub-section 4.0, the concept of incremental lexicon of the extended framework is discussed at length.

IV. Incremental Lexicon: A Multilingual and Multi-layered Mapping

Incremental lexicon is a token-addition to the concept of BanglaGen.vbp's extended framework. The term incremental is synonymous to distribution of feature and tags to derivatives or inflections of vocabulary list of BanglaGen.vbp. The increment order of vocabulary is translation equivalence, TAM feature, classifier information,

Language family information with intricate coagulation of root-stem/affix information. Following is the process to arrive at concept of incremental lexicon in BanglaGen.vbp, for further description:

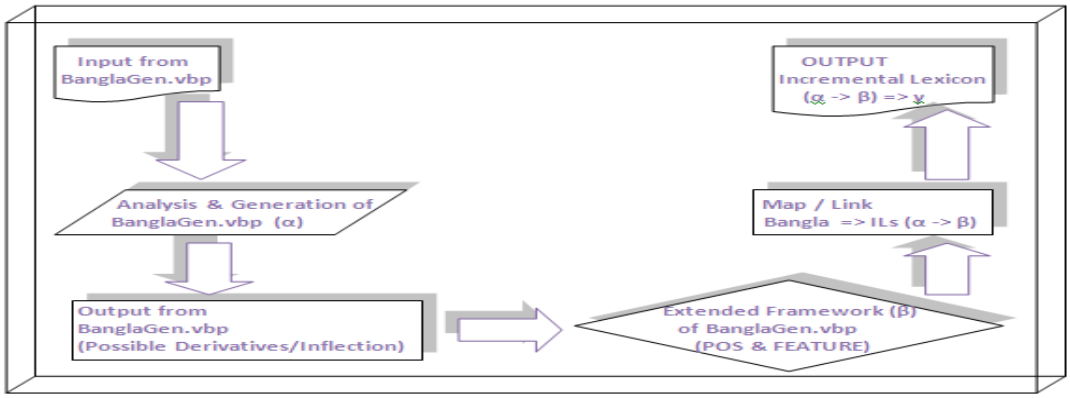

Figure 8: Process Flow for Incremental Lexicon

The following table is only a representation of above examples for various derivatives and inflection generations. The incremental lexicon is open ended where more features for classifier based languages can be included. For example, diglossic or poetic variation of derivatives in Bangla can be a layered information (which is

not included in the below table). Incremental lexicon can be further classified for root/stem and affix category in terms of inflectional and derivational morphology of Indian languages.

Following is the distribution of feature of Incremental Lexicon to the Translation equivalence of Head Word Derivative of BanglaGen.vbp:

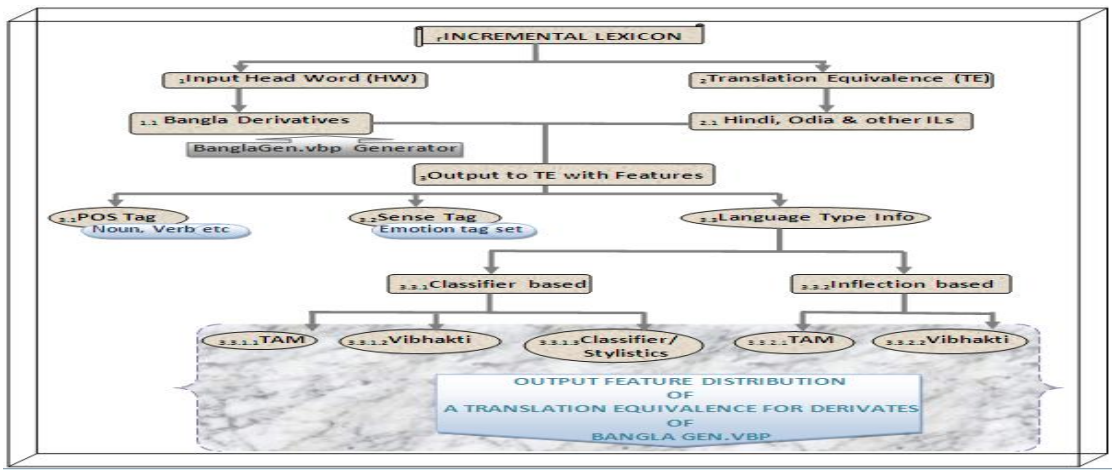

Figure 9: Process flow of incremental lexicon of BanglaGen.vbp

Consider few examples, to explain and elaborate on the concept of incremental lexicon introduced in extended framework of Banlagen.vbp. For a, input Head Word (HW) consider the following examples and cross referring to the above figure 9, i.e.

$_1$Input head word is ঘর [abode], নেব [to accept, to take] and বোকামি [idiot-ness]

$_2$Translation equivalent (TE) in Hindi and Odia correspondingly is घर (Hindi), ଘର (Odia).

Subsequently, the derivatives of ঘর from the generator frame will be imputed to list incremental lexicon, whose entries are:

$_{1.1}$Bangla Derivatives are (i) ঘরটা, (ii) ঘরটি, (iii) ঘরগুলো, (iv) ঘরগুলি, (v) ঘরের, (vi) ঘরে.

$_{1.1}$Bangla Derivatives with spelling variants: (vii) নেব, (viii) নেবো

$_{1.1}$Bangla Derivative with Sense tag feature: (ix) বোকামি

In connection to, the Translation Equivalent of above derivatives in Hindi and Odia are:

$_{2}$TE(i) is घर and ଘରଟା,

$_{2}$TE(ii) is घर and ଘରଟି,

$_{2}$TE(iii) is घरों, घरें and ଘରଗୁଡ଼ିକ,

$_{2}$TE(iv) is घरों, घरें and ଘରଗୁଡ଼ିକ,

$_{2}$TE(v) is घर का, घर के, घर की and ଘରର and,

$_{2}$TE(vi) is घर में and ଘରେ

$_{2}$TE (vii) is लुंगा, लुंगी, लेंगे and ନେବ

$_{2}$TE (viii) with spelling variants are not found in Hindi and Odia.

$_{2}$TE(ix) with sense tag feature in Hindi is मूर्खता and Odia is ବେ·ାକାବେ·ାକା

Accordingly, the Head Word Entry is annotated for POS tag and sense tag wherever applicable. For above HE, from (i) to (vi) whose POS tag is 'NOUN' but sense tag is 'null' and (vii) whose POS tag is verb and sense tag is 'null'; where else, in (ix) with POS tag is NOUN and a sense tag is 'disgust'.

In third category, language identifier identifies HE 'ঘর' the Hindi translation equivalent is inflection based and Odia is classifier based. Classifier based language further segmented in terms of TAM (or GNP), Vibhakti etc. and Inflection based language is further featured for TAM (or GNP) and Vibhakti (and GNP).

$_{3.3.1}$Classifier based (i to vi): Odia is a classifier based language, and the details of TE for the corresponding HEs are: [ଘରଟା, ଘରଟି, ଘରଗୁଡ଼ିକ, ଘରଗୁଡ଼ିକ, ଘରର, ଘରେ]

$_{3.3.1.1}$(TAM): [Singular, Singular, Plural, Plural, Singular, Singular]

$_{3.3.1.2}$(Vibhakti): [Nominative, Nominative, Nominative, Nominative, Genitive, Locative]

$_{3.3.1.3}$(Classifier / Stylistics): [Nominative, Nominative, Nominative, Nominative, Genitive, Locative]

$_{3.3.2}$Inflection based: Hindi is inflection based language and the details of TE for corresponding HEs are: [घर, घर, घरों~घरें, घर का~ घर के~ घर की, घर में]

$_{3.3.2.1}$(TAM): [Singular, Singular, Plural, Plural, Singular, Singular]

$_{3.3.2.2}$(Vibhakti): [Nominative, Nominative, Nominative, Nominative, Genitive, Locative] / Null

In third category, language identifier identifies HE 'নেব' in Bangla has a spelling variant that is treated as a form of derivative in BanglaGen.vbp. Nonetheless, in Hindi and Odia translation equivalent is लुंगा, लुंगी, लेंगे and ନେବ. This derivative is POS tagged as a 'verb'. Subsequently, residing the TAM and GNP information in Hindi as:

लुंगा: Future Tense, Masculine Gender, Singular number

लुंगी: Future Tense, Feminine Gender, Singular number

लेंगे: Future Tense, Masculine and Feminine Gender, Singular and Plural number

TAM and GNP information in Odia are as:

ଘରଟା: Future Tense, Gender (NULL), Singular and Plural number

ixth Bangla derivative is POS tagged as NOUN with sense tag as 'disgust' (as per Paul Ekman's six Emotion tag set).

Following table is a representation for database design of incremental lexicon for BanglaGen.vbp, with a few distinct features such as sense tag and language identifier (with language type information) largely supporting the Indian language computing for machine translation, information extraction and retrieval and with recent research area on sentiment analysis.

Table 13: Representation of Incremental lexicon from above extended framework

Derivatives			**Annotation**		**Language Identifier**				
Headword Word Entries & Translation Equivalents			**POS & Sense**		**Classifier Based**			**Inflectional Based**	
Bangla(B)	**Hindi(H)**	**Odia(O)**	**POS**	**Sense**	**TAM/ GNP**	**Classifier/ Stylistics**	**Vibhakti**	**TAM/ GNP**	**Vibhakti / Stylistics**
ঘরটা	घर	ଘରଟା	NOUN	0	Singular	(B) & (O)	Nominative	Singular	Nominative
ঘরটি	घर	ଘରଟି	NOUN	0	Singular	(B) & (O)	Nominative	Singular	Nominative
ঘরগুলো	घरों, घरें	ଘରଗୁଡ଼ିକ	NOUN	0	Plural	(B) & (O)	Nominative	Plural	Nominative
ঘরগুলি	घरों, घरें	ଘରଗୁଡ଼ିକ	NOUN	0	Plural	(B) & (O)	Nominative	Plural	Nominative
ঘরের	घर का, घर के, घर की	ଘରର	NOUN	0	Singular	-	Genitive	Singular	Genitive
ঘরে	घर में	ଘରେ	NOUN	0	Singular	-	Locative	Singular	Locative
নেব	लुंगा, लुंगी, लेंगे	ନେବ	VERB	0	Future, (Singular & Plural)	(B), (O) / (B)	-	Singular	Future, (Mas/Fem), (Singular /Plural)
নেবো	-	-	VERB	0	Future, (Singular & Plural)	(B)	-	-	-
বোকামি	मूर्खता	ବେ·ାକାବ େ·াକା	NOUN	Disgust	-	-	-	-	-
ন্যাকামি	नखड़ेदार	ନଖରାମି	NOUN	Anger	-	-	-	-	-
ন্যাকাপনা		ନଖରାପଣ	NOUN	Anger	-	-	-	-	-
ফরফরানি	चुलबुलाना	-	ADJECTIVE	Happy	-	-	-	-	-
ফরফরানো	-	-	GERUND	Happy	-	-	-	-	-
ফরফরে	-	-	ADVERB	Happy	-	-	-	-	-

This is very specific to mention that the concept of incremental lexicon is a *'Framework'* approach on derivative generation of BanglaGen.vbp. Nonetheless, equivalents relevant to feature/s with respect to language computing can be included within the existing periphery of the SLs and TLs (or those languages if included to execute within the purview of derivational or inflectional morphology) within bi or multilingual diaspora of Indian sub-continent.

V. Conclusion

In what explained above, are few intrinsic feature systems of Bangla, Hindi and Odia that are linked and mapped through a concentrated feature matrix as an extension to existing framework of BanglaGen.vbp, developed (the then only for Bangla) on windows on Visual Basic 5.0. The derivative generator has been extended to a morphological (and also a syntactico-semantic) analyzer and a concept of incremental lexicon with a multilingual mapping, as a '*gap spotting'* in Indian Language processing and computing. In what may follow, are more deep level feature with sense/emotion tagging outlined with methodologies for catering word-level analysis and formalism in Indian languages, currently prevailed computational framework (can be migrated to JAVA or more advanced version). Apparently, language specific facets and unseen or un-attempted assumptions of morphologically rich languages of Indian sub-continent that are meant for exploration, manifestation, and standardization.

VI. Transliteration

Table 14 (i): Transliteration of vowels

a	aa	i	ii	u	uu	e	ai	o	au	ã	a:	r
অ	আ	ই	ঈ	উ	ঊ	এ	ঐ	ও	ঔ	ং	ঁ	ঋ
অ	আ	ই	ঈ	উ	ঊ	এ	ঐ	ও	ঔ	অং	-	ঋ
অ	আ	ই	ঈ	উ	ঊ	এ	ঐ	ও	ঔ			ঌ

Table 14 (ii): Transliteration of consonants

ka	kha	ga	gha	Ng	ca	cha	ja	Jha	Ñ	Ta	Tha	Da	Dha	Na
ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ	ট	ঠ	ড	ঢ়	ণ
ক	খ	গ	ঘ		চ	ছ	জ	ঝ		ট	ঠ	ড	ঢ	ণ
‹	Œ		Ô	—			‘	’	“	”	•	“	“	~
ta	tha	da	dha	na	pa	pha	ba	bha	ma	ya	ra	la	va	
ত	থ	দ	ধ	ন	প	ফ	ব	ভ	ম	য	ড়	ল	য়	x
ত	থ	দ	ধ	ন	প	ফ	ব	ভ	ম	য	র	ল	ব	x
™	Š	›	œ	¬	Ç	Ÿ»	¡	¢	£	¤ý	¥	¬	§	x
sha	Sa	Sa	ha	ksha	tra	jna								
শ	ষ	স	হ	ক্ষ	ত্র	জ্ঞ	x	x	x	x	x	x	x	x
শ	ষ	স	হ	ক্ষ	ত্র	জ্ঞ	x	x	x	x	x	x	x	x
..		a	«				x	x	x	x	x	x	x	x

VII. Colophone

This proposed framework in is an expansion to my PhD thesis awarded in the year 2001 and was presented in Computational Linguistics session as a selected paper-abstract at 36th International Conference of Linguistic Society of India (ICOLSI 36) in Thiruvananthapuram, India during 1-4 December 2014. My sincere gratitude to my supervisor and mentor Prof. Udaya Narayana Singh in completing, submitting and receiving the PhD degree awarded in the year 2001. My sincere thanks to Dr. Ranjan Das for Odia examples and their mapping; and to Mr. Shashipal Singh for installing the BanglaGen.vbp project on VB 5.0 in current Operating System. My constant and significant inspiration and aspiration are due to my parent and all those invaluable and gifted beings for rendering my life in a gracious way.

Reference

[1] Dasgupta, Aparupa and Udaya Narayana Singh. (1997). *A Morphological Analyzer for Bangla*. Paper presented at the **2nd ICSALL**. Punjabi University, Patiala.
[2] Sproat, Richard. (1991). *Morphology and Computation*. MIT Press, Cambridge.
[3] Ekbal, A., Mandal, S., & Bandyopadhyay, S. POS *tagging using HMM and rule based chunking* . Workshop on Shallow Parsing for South Asian Languages. 2007.
[4] Dandapat, S. Part-*of-Speech Tagging and Chunking with Maximum Entropy Model*. Workshop on Shallow Parsing for South Asian Languages. 2007.
[5] Alvesson, Mats and JÖrgen Sandberg. 2011. *Generating Research Questions Through Problematization.* Academy of Management Review. University of Lund and University of Queensland. *Vol 36. No. 2. 247-271*.
[6] Dubey, Debasri and Dasgupta, Aparupa. (2010). *Morphological Analysers and Generators*. Knowledge Sharing Event 1. CIIL. Mysore. (In Press).
[7] Sinclair, J. 1991. *Corpus, concordance, collocation*. Tuscan Word Centre, Oxford: Oxford University Press.
[8] Krishnamurti, Bh., C.P. Masica and A. K. Sinha (eds). (1986). *South Asian Languages: Structure, Convergence and Diglossia*. Delhi: Motilal Benarsidass.
[9] Paul Ekman. 1993. *Facial expression and emotion*. American Psychologist, *48(4):384–392*.
[10] Baskaran S. et al. 2008. *Designing a Common POS-Tagset Framework for Indian Language*. The 6th Workshop on Asian Language Resources.
[11] Singh, Udaya Narayana. (1980). *Comments on 'On rule ordering in Bengali phonology'*. ***IL*** *40:2.91-101*.
[12] Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma and Lakshmi Bai. 2006. AnnCorra:*Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages*, Technical Report, Language Technologies Research Centre IIIT, Hyderabad.
[13] Kellog, S.H.(1875). *A Comparative Grammar of Hindi Language*. *3rd ed.* London: N. P. [Reprinted in London, Routledge and Kegan Paul, 1955].
[14] _____. 2004. *Developing Linguistic Corpora: A Guide to good practice*. Oxford: Oxford University Press.
[15] _____, & Sarkar, S. *Part-of-Speech Tagging for Bengali with Hidden Markov Model*. NLPAI ML workshop on Part of speech tagging and Chunking for Indian language. 2006.
[16] WEBSITE. ftp.cis.upenn.edu:/pub/ldc: http://clwww.essex.ac.uk/w3c/corpus_ling/content/introduction.html http://wombat.doc.ic.ac.uk/ (on-line dictionary of computing)