Vice Chancellor, Deccan College Post-Graduate and Research Institute (Deemed University), Pune

SUFFIX CONCATENATION IN BANGLA NOUN MORPHOLOGY

Author(s): Atiur Rahman Khan

Source: *Bulletin of the Deccan College Post-Graduate and Research Institute*, Vol. 77 (October 2017), pp. 336-343

Published by: Vice Chancellor, Deccan College Post-Graduate and Research Institute (Deemed University), Pune

Stable URL: https://www.jstor.org/stable/10.2307/26609194

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Vice Chancellor, Deccan College Post-Graduate and Research Institute (Deemed University), Pune is collaborating with JSTOR to digitize, preserve and extend access to *Bulletin of the Deccan College Post-Graduate and Research Institute*

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms

Title of the Research	**:**	**SUFFIX CONCATENATION IN BANGLA NOUN MORPHOLOGY: A FINITE STATE AUTOMATA (FSA) MODEL**
Research Scholar	**:**	**Atiur Rahman Khan**
Research Guide	**:**	**Prof. Sonal Kulkarni-Joshi**
Year of Award of Degree	**:**	**2017**

Part I

Introduction

The present thesis titled 'Suffix Concatenation in Bangla Noun Morphology: A Finite State Automata (FSA) Model' aims to discuss and analyze the process of suffix concatenation in inflectional morphology of Bangla with focus on the written form of Standard Colloquial Bangla (SQB). The suffix concatenation in the linear morphology is described and explained through the Unification-based Two-level Morphology (UTLM) model using its Finite State Automata (FSA) technique.

The thesis follows a rule-based approach in morphological analysis in the field of Natural Language Processing (NLP) by analyzing the morphotactics of Bangla nouns.

Bangla (English exonym: Bengali), is one of the widely spoken languages in the world. It is spoken in India (West Bengal, Tripura) and Bangladesh as enjoys the status of official language in the Indian state of West Bengal and Bangladesh. As per the 2011 Census Bangla is the second most spoken language in India.

To make the computer understand human language NLP is involved with two processes- Analysis and Generation. Natural language analysis and natural language generation is achievable only with a thorough study of the morphology of the language. Computational linguistics endevours to research, evolve and develop suitable formalisms to represent natural language behaviour. The present thesis attempts to choose one of these formalisms to represent the morphotactics of Bangla nouns for language processing tasks. Different levels of linguistics has significance for NLP tasks.

Importance of Morphological studies in NLP:

Morphology is placed at the conceptual centre of linguistics, hence the centre of all language processing tasks. This is not because it is a dominant sub-discipline, but because morphology is the study of word structure, and words are at the interface between phonology, syntax and semantics (Spencer, 2005). Pertaining to the morphological phenomena some basic activities such as stemming, lemmatising and more advanced ones, like parsing,

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms

morphological analysis and generation, began to be undertaken. The task of stemming, lemmatizing and morphological analysis in NLP has been dealt with various methods over the last few decades. As a result the computational approach to linguistics has its influence on the major sub-disciplines such as phonology, morphology, lexicography among others. Such research-based activities concerned with morphology becomes even more challenging in inflectionally rich languages.

Approaches in Natural Language Processing

In NLP the methodology adopted for morphological analysis[1] -generation[2] and related researches could be broadly categorized as supervised and unsupervised approaches. Following the broad classification in morphological analysis one can incorporate approaches such as rule-based ones as a supervised method, and statistical and corpus based approaches as unsupervised ones.

Supervised and Unsupervised approaches of Morphological Analysis

Any morphological study in NLP is done applying different methods and approaches. A broad classification can be done as supervised and unsupervised approaches. An unsupervised morphological analysis is the task of segmenting words into prefixes, suffixes and stems without the prior knowledge of the language-specific morphotactics and morpho-phonological rules.

Part II

An Overview of the Doctoral Thesis

The following paragraphs provide a brief overview of the hypothetical question, aims, objectives, scope, limitations and sources of the present thesis. It also provides a terse description of each chapters beginning with the Introduction and Review of Literature to chapters on Methodology, Bangla Noun Morphology and Computational Morphology of Bangla Nouns. At the end is discussed the Testing and Conclusion chapter along with the appendix and citations. This will help understand the present research work and make the reader aware of the general flow of the thesis.

Hypothesis

Considering a large written data of modern Bangla, as a sample, it seems that every noun in the language does not inflect with all the suffixes valid for inflectional morphology. There is a systematic choice of suffixes due to certain factors.

Research Question

What are those grammatical factors that influence the concatenation process in Bangla noun morphology and how do they condition the choice of suffixes? Can this process of suffix concatenation be described and explained with the help of a Natural Language Processing (NLP) formalism?

Aim

The aim of the present thesis is to describe the morphotactics of nouns in

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms

Bangla within the technical framework of Finite State technique of Antworth's Unification-based Two-level Morphology. This shall cater to Morphological Analyzer (MA) and Morphological Generator (MG) for Bangla.

Objective

The main objective of thisdoctoral thesis is to formulate a tangible framework for noun classification system in Bangla on the basis of semantic, phonological and orthographic parameters. Based on the classification system a schema for concatenation in Bangla nouns could be designed and subsequently attempt to represent the process through Unification-based Two-level Morphology (UTLM) and Finite State Automata (FSA).

Scope

The scope of the thesis is inflectional noun morphology in Bangla.

The morphological study in the thesis does not cover all inflectional markers. It includes only plural marker-**rā**, case markers -**ke**(acc./dat.), -**r**(genitive), -**te** and -**y**(locative/ nominative), the portmanteau morph-**der**(acc.pl./ dat. pl./gen. pl.), with their orthographic variants, the classifiers -**Tā**, -**Ti** (singular), -**gulo**, -**guli**(plural), and the emphatic and inclusive markers -**i** and -**o** respectively. The selection of these inflectional morphological markers is based on the fact that all of these can join with almost every Bangla nouns.

Limitations

The study takes into account only the standard form of written Bangla and does not include any morphological markers from any of its dialects.Following its aim to deal with the inflectional morphology, the present thesis includes suffixes only. Prefixes and derived morphemes are excluded in the study.Compound words with sandhi and compounds with the second unit (word) as a derived one are also not included in the present morphological study. These type of nouns are not taken into account for analysis or testing.

Review of Literature (Chapter 2)

The review covers various morphological analyzers for Bangla and the **extant methods, formalisms** and **approaches**, particularly the FSA technique adopted by various computational linguist and research institutes.The literature review ends with the approach and model adopted for the present research activity. It lays down the advantages choosing the **Unification-based Two-level Morphology** represented through the **Finite State Automata model**(FSA), one of the modern and productive models for analysis in computational morphology.

Methodology (Chapter 3)

Different methods have been applied for different purposes at different levels of the research work. In the collection of data, analysis and classification of nouns, in pilot survey it is **Deductive approach**. The approach used to demonstrate themorphotactics of Bangla nouns follows the **Unification based Two-level Morphology** (UTLM).TheFSA model is used in representing the suffix concatenation in

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms

Bangla inflectional morphology (discussed above in section 1).The testing method, therefore, follows the **Corpus-Based Approach** as far as its methodology is concerned.

Bangla Noun Morphology (Chapter 4):

This chapter discusses the broad three-fold noun categorization system working in Bangla based on the phonological, semantic and orthographic parameters. It explains the phenomenon by creating a generic taxonomy of Bangla nouns based on these factors. The semantic parameter makes distinction between animate and inanimate, count and mass noun, concrete and abstract noun and the inherent plural nouns. While the phonological influence distinguishes between a monosyllabic and a polysyllabic, orthographic parameter conditions the shape (form) of the suffix to be concatenated.

There is a clear and direct influence of animacy in deciding suffixes for human and other animate objects owing to the **semantic** factors.

Another factor that conditions choice of suffix of the word (noun) is the last character or the word-ending. This is called **orthographic** restriction which changes shape and calls for morphophonemic changes during the suffixation.

Computational Morphology of Bangla Nouns (Chapter 5)

It provides a detailed and exhaustive FSA representation of Bangla nouns through an FSA diagram, state description and transitions, and equations based on the suffix concatenation process in inflectional morphology. The FSA representation made shall help a programmer write flawless algorithm for Bangla inflectional nouns.

Testing and Conclusion (Chapter 6)

The testing is done on a large corpus of written Bangla procured from GIST Group, Centre for the Development of Advanced Computing (C-DAC), Pune. For further accuracy, testing has also been done on corpus collected from the internet (the google search engine, www.google.co.in) and the online AnandabazarPatrika (www.anandabzar.com).

Major outcome of the research

The major outcome of the research work in the thesis on morphology of nouns endevours to achieve the following:

1. A **taxonomy of Bangla nouns** based on phonological, semantic and orthographic parameters.
2. Identification of common inflectional **Suffix list** that uniformly joins with all stems (nouns)under a particular class.
3. Formation of **Suffix sequence** or suffix order for each noun class in accordance with Bangla morphotactics.
4. One of the outcomes of the research work would be an **exhaustive list of noun forms** with all types of concatenation with inflectional suffixes possible in Bangla Inflectional morphology. Each table lists the representative noun of that particular class along with the morphotactically relevant combinations of suffixes

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms

(denoting number, case, classifier and clitics). It is provided in the appendix.

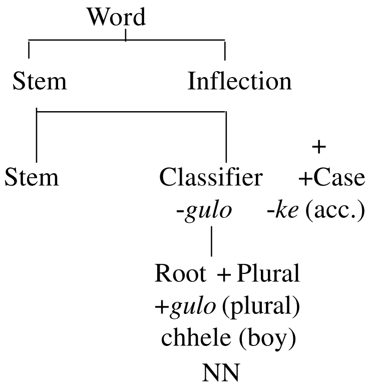
5. A **Finite State Automata rule** pertaining to morphotactics for Bangla Inflectional Nouns.
6. Such a rule-based tackling of nouns shall be helpful for various basic and high-end NLP tasks such as **morphological analyzer/morphological generation.**

Computational Grammar for Noun Morphology in Bangla

For example, the Bangla inflected word cheleguloke 'to the boys' is tokenized into this sequence of morpheme structures:

Form: chhele+ (boy)	gulo(plural classifier)	+ke (acc.)
Cat: Root	+PL	+CASE
Feat: [Lexcat: N	[fromcat: N	[fromcat: N
Afrom: !PoS	tocat: N	tocat: N

This analysis is then passed on to the word grammar which returns the tree and feature structure after the parsing shown below: (Antworth's model)

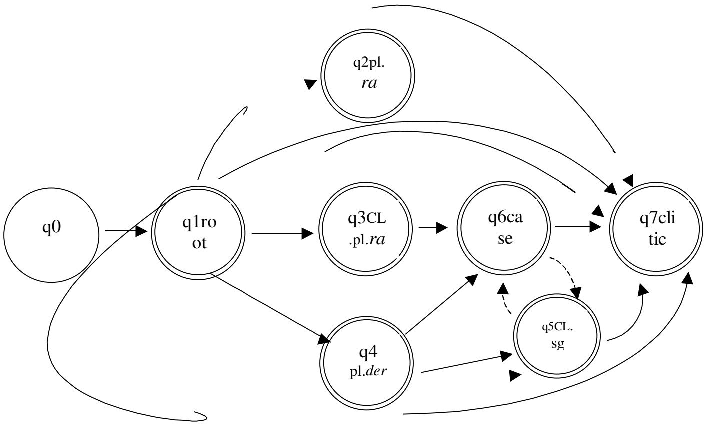


The final output (as provided by the unification-based Two-level morphology (feature structure) (Antworth (1993), Shanmugam 2010)

Word: chheleguloke (boys-pl.classifier-Acc.)
[lexcat: N
Number: PL
Case: Accusative

Finite State Automata for Bangla Noun Morphology

Here is how the Bangla inflectional noun morphology could be represented in Unification-based Two-level Morphology through FSA.



STATES
q0 = initial state,
q1 = State (root noun);
q2 = state (plural-1);
q3 = State (classifier plural);

Fig. 2. The FSA Diagram for Bangla Nouns

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms

Bibliogralphy

Adams, K. L. 1989. Systems of Numeral Classification in the *Mon-Khmer, Nicobarese and Aslian Subfamilies of Austroasiatic*. Australian National University.

Aikhenvald, Alexandra Y. *Noun Classes and Classifiers, semantics of...* .Research Centre for Linguistic Typology, La Trobe University, Melbourne.

Aikhenvald, A. Y. 2000. Classifiers: *A Typology of Noun Categorization Devices.* Oxford University Press, Oxford. Oxford studies in typology and linguistic theory.

Allan, K. 1977. 'Classifiers'. *Language*. Vol. 53: 284-310.

Andronov, M.S. 1964. *On the typological Similarity of new Indo-Aryan and Dravidian.* Indian Linguistics, Vol. 25.

Antworth, Evan L. 1994. *Morphological Parsing with a Unification-based Word.* Grammar University of Texas at Arlington.

Antworth, Evan L. 1990. PC KIMMO: *A Two-level processor for morphological analysis*. Summer Institute of Linguistics, Dallas, TX.

Bangla Script Grammar. 2011. Compiled by GIST Group, C-DAC, Pune.

Beams, John. 1891. *A Comparative Grammar of the Modern Aryan Languages of India: to wit, Hindi, Panjabi, Sindhi, Gujarati, Marathi, Oriya, and Bangali*. Vol. III. Trubner & Co., Ludgate Hill, London.

Bhattacharya, Pareshchandra. 1996. *Bhashabidya Parichay*. Jaydurga Library. Kolkata.

Bhattacharya, S, Choudhury, M, Sarkar, S. 2005. *'Inflectional Morphology Synthesis for Bengali Noun, Pronoun and Verb Systems.'* In Proceedings of the National Conference on Computer Processing of Bangla (NCCPB 05).pp. 34 - 43, Dhaka, Bangladesh.

Bolinger, D. 1975. *Aspects of language.* Harcourt Brace Jovanovich (83-90/ 99-123), New York.

Carey, William. 1818. *A Grammar of the Bengalee Langauge*. 4thed. The Serampore Mission Press, Serampore.

Chatterji, Suniti Kumar. 2003. *Bhasha Prakash Bangla Byakaran.* Rupa and Co. New Delhi.

Chatterji, Suniti Kumar. 2002. *The Origin and Development of the BengalILangauge*. Rupa and Co. New Delhi.

Das, Amitava & Sivaji Bandyopadhyay. 2009."*Morphological Stemming Cluster Identification for Bangla*". Jadavpur University, Kolkata.

Dasgupta, Sajib, and Vincent NG. 2009. *Unsupervised Part-of-Speech Acquisition for Resource-Scarce Languages*. University of Texas at Dallas.

Dasgupta, Sajib, et al. *Morphological Analysis of Inflecting Compound Words in Bangla*. BRAC University, Dhaka, Bangladesh.

Das, Suprabhat and PabitraMitra. 2011. *A Rule-based Approach of Stemming for Inflectional and Derivational words in Bengali*. Proceedings of the 2011 IEEE Student's Technology Symposium. 14-16 January, IIT Kharagpur.

Dasgupta, P. 1985. *On Bangla Nouns. Indian Linguistics*. Vol.46. no. 1-2-4. Deccan College, Pune

Dash, Niladri Shekhar. 2015. *A Descriptive Study of Bengali Words*. Cambridge University Press. Cambridge House, Delhi, India.

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms

Gangopadhyay, M. 1990. *The Noun Phrase in* Bengali: Assignment of Role and the Karaka Theory.MotilalBanarsidar Publishers Pvt. Ltd. New Delhi.

Halle, K & S J Keysar. 1993. *Distributed Morphology and the Pieces of Inflection*. The View from Building 20th, Chapter 2. Morris Halle and Alec Marantz. MIT Press.

Jurafsky, Daniel and Martin, James H. 2002. *Speech and Language Processing-An Introduction to NLP, Computational Linguistics and Speech Recognition*. Pearson Education.

Karttunen, L. Kaplan, R.M & Zaenen 1992. Two-Level Morphology with Composition. *Proceedings of COLING-92*. Nantes. Xerox Palo Alto Research Center, Centre for the Study of Language and Information, Stanford University.

Karttunen, L. 2003. *Computing with Realizational Morphology*. In: Gelbukh, A. (ed). CICLing 2003 Lecture Notes in Computer Science 2588. pp. 205-216. Springer-Verlag, Berlin, Germany.

Karttunen, L. 2007. *Word Play.* Computational Linguistics. 33 (4). pp. 443-467.

Koskenniemi, K. 1983. T*wo-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Publication 11, Department of General Linguistics, University of Helsinki, Helsinki, Finland.

Koskenniemi. 2007. *Notes on the Two-Level Morphology*. A Man of Measure. Frestschrift in Hounour of Fred Karlsson, pp.422-431.

Mudassar M. Majgaonker et al. 2010. *Discovering suffixes: A Case Study for Marathi Language*. (IJCSE) International Journal on Computer Science and Engineering. Vol. 02, No. 08, 2010, pp. 2716-2720.

Minnen, Guido. Carroll, John. Pearce, Darren. Robust. *Applied Morphological Generation. Cognitive and Computing Sciences.* University of Sussex, Brighton, UK. (A paper presented at the North Texas Natural Language Processing Workshop).

Mohanty S, P K Santi, K.P. Das Adhikary. 2004. *Analysis and Design of Oriya Morphological Analyser*: Some Tests with OriNet. Proceeding of symposium on Indian Morphology, Phonology and Language Engineering. IIT Kharagpur.

Nicholl, G.F. 1885. *A Bengali Grammar also an Assamese Grammar*. W.H. Allen & Co. London. pg. 13-14.

Porter, M.F. *An Algorithm for suffix stripping*. Cambridge.

Ray, P R. Basu, A, & Sarkar, S 2004. 'Bengali Derivational Morphological Analysis using Finite State Methods.'*Lecture Compendium*. Symposium on Indian Morphology, Phonology & Language Engineering. IIT Kharagpur, India.

Racova, A. 2007. '*Classifiers in Bengali*'. Asian and African Studies, 16. Vol. 2, pp. 125-137.

Reed, N. Harvey, Mark (eds). 1997. *Nominal classification in Aboriginal Australia*. John Benjamins Publishing Company, Amsterdam.

S.Viswanathan, S. Ramesh Kumar, B. Kumara Shanmugam, S. Arulmozi, & K.Vijay Shanker. 2003. *A Tamil Morphological Analyser.* ICON-2003, pp. 31-39.

Sarkar, Pabitra. 2006. *Pocket Bangla Byakaran.* Aajkaal Publishers Ltd. Kolkata.

Sarkar, Pabitra. 2006. *Bangla Byakaran Prasanga*. Dey's Publishing. Kolkata.

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms

Sen, Sukumar. 2007. *Bhashar Itibrittya.* Ananda Publishers.

Shelton. J., Caramazza A. 2001. *The organization of Semantic Meaning.*

Simons, Gary F. 1991. *Computing in Linguistics: a Two-level Processor for Morphological Analysis.* Notes on Linguistics 53. From Ethnologue.

Sproat, Richard. 1992. *Morphology and Computation.* The MIT Press. Cambridge.

Senapati, Apurbalal & Utpal Garain. (2012). Bangla Morphological Analyzer using Finite Automata: ISI@ FIRE MET 2012.

Shanmugam, R. 2010. *Issues in Morphological Parsing for Modern Tamil.* Ph.D. Thesis submitted to the Department of Tamil Language, University of Madras, Chennai.

Tagore, Rabindranath. 1995. *Bangla Shabdatatva.* Vishwa Bharati University.

Vitrant, Alice. 2002. *Classifier Systems and Noun Categorization Devices in Burmese.* Proceedings of the Twenty-Eighth Annual Meeting of the Berkeley Linguistics Society: Special Session on Tibeto-Burman and Southeast Asian Linguistics (pp. 129-148).

Walsh, Michael. 1997. *Noun Classes, Nominal Classification and Generics in Murrihnpatha.* In Harvey, M. and Reid, Nicholas's 'Nominal Classification in Aboriginal Autralia'. John Benjamins Publishing Company.

Yates, William. 1849. *A Bengali Grammar.* Ed. John Wenger. Baptist Mission Press. Calcutta.

Yona, S. and Wintner, S. 2007. *A finite-state morphological grammar of Hebrew.* Natural Language Engineering. 14. pp. 173-190.

This content downloaded from
123.108.244.227 on Tue, 06 Oct 2020 06:03:39 UTC
All use subject to https://about.jstor.org/terms