Bangla pronouns - a corpus based study

Article *in* Literary and Linguistic Computing · December 2000

DOI: 10.1093/llc/15.4.433

Niladri Sekhar Dash
Indian Statistical Institute

Bangla Pronouns: a Corpus Based Study

Niladri Sekhar Dash
Indian Statistical Institute, Calcutta, India

Abstract

[Bangla is the second most widely-spoken language in the Indian subcontinent, yet has not been the focus of much research activity in either corpus linguistics or language engineering to date. This paper describes the automatic processing of pronouns in three and a half million words of Bangla corpus data. A corpus based analysis of Bangla pronouns is developed, and a new approach to the analysis of Bangla pronouns is taken as a consequence. On the basis of this analysis a system is then developed to identify and analyse Bangla pronouns in corpus data]

1. Introduction

Words[1] in Bangla and other human languages are subjected to complex morphological processes. These processes result in the component parts of a word not being susceptible to analysis by a simple, general, algorithm (Spencer 1991). Such morphological processing is, however, an important part of natural language processing (NLP), as it can yield important information to a linguistic analysis. To take Bangla as an example, a morphological analysis can aid in the identification of the morphosyntactic category of a word, the identification of the functional role or roles the word plays in a sentence in which it appears and the recovery of semantic information. In order to achieve such an analysis, two obvious strategies present themselves - working at either the word level, or above the word level. At the word level, a context free analysis is possible which can automatically extract information such as part-of-speech, tense, aspect, case, person, number, gender, particle, honorification etc. However, a context free approach may lead to ambiguous analyses. Above the word level, a context bound approach is possible, using, for example, co-occurrence information (Garside, Leech and Sampson, 1987) or grammatical rules working at the sentence level (Karlsson *et al,* 1995) to achieve disambiguation. Needless to say, word level and context bound analyses need not be independent, and information gathered from word level processing can be used for context bound analysis, with context being applied to achieve disambiguation. The results of such a disambiguated analysis can then, in turn be used to generate further levels of analysis such as word sense disambiguation or information extraction.

Given that morphological analyses are such a useful first step towards NLP, the work presented here on a morphological processor for Bangla, working from the context free to the context bound level, is justifiable, especially as this will provide a basis for future work on the

Dash, Niladri Sekhar (2000) “Bangla pronouns: a corpus-based study”. *Literary and Linguistic Computing*. Vol. 15. No. 4. Pp. 433-443.

automatic processing of Bangla. Reliable morphological processing is a good building block for the development of further language processing systems. This paper presents the first steps taken towards an automated morphological analysis of Bangla.

Bangla has a rich inflectional morphology, and as such it was decided to minimise the scope of the initial morphological processing to Bangla pronouns. Pronouns form a useful focus for this work, as there are few individual pronoun word forms in Bangla, yet any given word form is used frequently, as is typical of closed class words. The morphological process of suffixation used on Bangla pronouns and the system of case marking for inflected pronouns is similar to that used for the much more populous category of nouns. The only noticeable difference is that prefixation cannot be used with pronouns. In addition to being a useful means of examining the morphology of nouns by default, a context free analysis of pronouns yields a less ambiguous analysis than a corresponding study of nouns would. Typically, a context free analysis of a pronoun should lead to a non-ambiguous morphosyntactic analysis of the word, though there are some forms such as *ye*, *ei*, *ki*, etc. which yield ambiguous analyses in some contexts.

For the analysis of pronouns in Bangla, a written corpus of three million words was used. The corpus was obtained from the Ministry of Information Technology in India. In addition, a corpus of 500,000 words (Dash and Chaudhuri, 2000) was employed to test the automated pronoun analysis system developed on the basis of the larger corpus. Both corpora represent a range of written genres of Bangla. There are some problems with these corpora, however. Neither represent newspaper data, a genre common to many corpora in use. Also, the larger of the two corpora employs a random sampling method that deeply perturbs text structure in the corpus, but as the focus of this paper is not upon textual analysis this can safely be ignored. However, for other purposes, this limitation may well render the 3 million word corpus of Bangla unusable.

Using the Bangla corpus data we were able to explore the morphological structure of Bangla pronouns. However, before outlining that analysis, it is necessary to explain why existing studies of Bangla pronouns were not considered to be a useful basis for automatic processing.

2. Studies of Bangla pronouns to date

Traditionally, a pronoun is used in language to replace one or more nouns with which it holds a syntactico-grammatical relationship. However, the identification of a pronoun in a running text is generally carried out by identifying the role it plays in the sentence as well as its meaning. There exist few observations of the formation and function of pronouns in Bangla. Among them Chatterji (1926, 1993) gives a list of pronoun roots with their origin and classification and provides a comparative study with pronouns available in other Aryan languages. Sen (1993) provides an etymological history of Bangla pronouns and Sarkar and Basu (1994) classify them and explore their uses in texts. Chaki (1996) discusses the classes of pronouns, nouns used as pronouns and some pseudo pronouns found in Bangla. Common to all of the studies is a focus on defining Bangla pronouns with reference to their meaning, rather than with reference to an analysis of the component morphemes of the pronoun. The following classification by Chaki (1996) is a good example of this:

(1) personal pronoun	: aami (I), tumi (you), se (he) etc.
(2) demonstrative pronoun	: ei (this), ai (that), sei (that) etc.
(3) indefinite pronoun	: keu (someone), kichu (some) etc.
(4) interrogative pronoun	: ke (who), ki (what), kaake (to whom) etc.
(5) reflexive pronoun	: aap(a)ni (yourself), nija (oneself), khod (oneself) etc.
(6) relative and correlative pronoun	: ye (that), yini (who), yaa (what) etc.
(7) reciprocal pronoun	: paraspar (each other), aapnaa-aapni (among us)
(8) inclusive pronoun	: sab (all), sakal (all), sarba (all) etc.
(9) denoting pronoun	: anya (other), apara (other), para (other) etc.

This kind of classification, while providing a general linguistic base for identifying and understanding the role of pronouns in the language is not helpful for automatic language processing. Importantly, such an analysis is too general and over arching. For the purposes of NLP we wished not simply to categorise individual pronouns, but to analyse the morphology of each instance of a pronoun to indicate the role that the pronoun was playing in any given sentence. The majority of Bangla pronouns are inflected, either with case marking and/or other suffixes. For example, *tomraaderke* (to you, plural), would simply be classified as a 'personal pronoun' in Chaki's, yet this word is clearly somewhat more complex than this analysis would suggest. Pronominal roots undergo a morphophonemic change whenever they use case markers or suffixes, as happens with *tomraaderke*. A study of the roots *tum-* and *tu-* (you) in the corpus show that they change into *tom-* and *to-* respectively whenever the plural suffix *-raa* is used with them, as occurs in *tom* (you)+ *raa* (plural) + *der* (genitive) + *ke* (nominative). So, in addition to being a case marked personal pronoun, it is also a pronoun liable to morphophonemic change. To give some further examples of this kind of morphophonemic change:

(i) *tumi* (you, singular) becomes *tomraa* (you, plural)
(ii) *tui* (you, singular) becomes *toraa* (you, plural)

Consequently, rather than rely on an analysis of Bangla pronouns which simply categorises base forms of the pronoun into general categories, the aim of this study was to examine pronouns used in a corpus of Bangla. Such an analysis must take into account the complex morphemic and morphophonemic processes operating upon Bangla pronouns as a prelude to developing a system to automatically analyse and identify Bangla pronouns.

3. Corpus-based observations of the structure of Bangla pronouns

Structurally, Bangla pronouns are of two types: (i) pronouns with case, number and other markers, and (ii) non-inflected pronouns. In order to explore Bangla pronouns, each pronoun in the 3,000,000 corpus of Bangla used in this study corpus was identified manually. This provided some raw data about pronoun use in Bangla. In the corpus, 3% of the word forms are pronouns. Of these pronouns, 75% are inflected with case markers or other suffixes, while 25% of the pronouns were root forms. The number of individual pronoun word forms in the corpus is around 800. However, the number of pronoun roots is 65. There are 90 suffixes and case marking morphemes used with the pronominal roots. Among the inflected forms personal and demonstrative pronouns are most frequently used in the text. One final observation about Bangla pronouns derived from the corpus was the identification of a new pronominal form, kei which is not mentioned in any of the existing Bangla grammars and dictionaries. While more detailed explorations will be required to discover the genesis of this pronoun, it is probably derived from *keui* "whoever" by a process of morphophonemic change whose discussion is beyond the scope of this paper.

Having identified the pronouns in the corpus, the next stage was to examine those pronouns. For this purpose, a system of analysis was developed which identified the major processes operating on the pronominal root form of each word. This yielded three broad categories of pronoun - simple, inflected and adjectival. Simple pronouns like *se* (he), *e* (this), *ei* (this), *o* (that), *ke* (who), *ki* (what) are used in the text without suffix or case markers. Inflected pronouns like *tomaake* (to you), *aamaar* (mine), *seguli* (those), *aapnaader* (to you) bear inflectional suffixes and case markers. Adjectival pronouns like *tvadiya* (yours), *svakiya* (of one), *kata* (how much) are mostly used as adjectives in the text without any inflectional suffixes or case markers. This three way division of pronouns is useful, as it corresponds to the major roles that the pronoun can play in Bangla: a simple pronoun, a pronoun bearing a great deal of syntactic information or an adjective. Having broadly classified the pronouns into these three categories, the focus of the analysis switched to a consideration of the morphological processes operating on the pronouns.

The corpus data shows that Bangla pronouns are generated following a relatively fixed pattern of morpheme arrangement. However, in some cases, the sequence of arrangement of the suffix elements can be changed in the surface form. For instance, the sequence number/case can be reversed to case/number. In table (1) below the structure of Bangla pronouns derived from the corpus data is shown. The table shows variable sequences such as case/number and number/case by exhaustively listing the observed possibilities. Following the table, subsections outline specific observations regarding the morphology of Bangla pronouns made on the basis of the corpus analysis.

Prononminal form	**Surface form**	**Structure**	**Gloss**
Base Form (BF) only	*se*	se + 0	he
BF with Person Marker	*tumi*	tum + I	you
BF with Particle Marker	*sei*	se + I	he himself
BF with Person and Particle	*tumii*	tum + i + I	you yourself
BF with Number Marker	*tomraa*	tom + raa	you (pl)
BF with Case Marker	*tomaake*	tomaa + ke	to you
BF with Number and Case	*tomaaderke*	tomaa + der + ke	to you (pl)
BF with Case and Number	*tomaarT*^{2}*aa*	tomaa + (aa)r + Taa	of yours
BF with Case, Number and Case	*tomaarTaate*	tom + aar + Taa + te	in yours

Table 1. The structure of Bangla pronouns.

3.1 Suffixation in Bangla

As is apparent from the above table, Bangla pronoun suffixes denote number, person, honorification markers and particles. The concept of grammatical gender is absent in Bangla. The total number of word final suffixes used with pronouns in the corpus is 20. Of these, 15 are number markers, 2 are person markers, and 3 are particles which are generally used at the end of the pronouns. For honorification, there are some specific pronominal roots such as *aap(a)n-, tin-* which force verb forms to employ certain specific suffixes like *-en, -n* whenever they are used in a sentence.

3.2 Markers of singularity and plurality

The personal pronominal roots use *-Ø* (null) and *-i* for denoting singularity and *-raa* for plurality. Among these, *-i* is used with all personal pronominal roots irrespective of first, second or third person. Moreover, it does not cause any morphophonemic change in the surface form. The plural marker -*raa* can also be used with all personal roots irrespective of number or gender. However, it has a tendency to alter the morphophonemic structure of the root, as noted in section two with reference to the pronoun *tomraaderke*.

The singular and plural markers used with personal pronominal roots are also used for indicating their nominative case because there are no specific nominative case markers for pronouns in Bangla. However, a number of specific enclitics exist (mostly derived from words denoting numerical or quantitative loads inherent with the forms), which are generally used with the pronouns for this purpose. The enclitic forms such as *-kaTaa, -kaTi, -khaanaa,* and *-khaani* are typically post modifiers marking number: e.g. *yekaTaa* (those), *sekhaani* (that). The notable aspect of these enclitics is that in the corpus none of them are suffixed to personal (human) pronominal roots with the exception of se which denotes either a non-human object or a person well known and regarded by the speaker such as *seTi* (that), *seTa* (that chap). Among plural markers *-kaTaa* and *-kaTi* are probably derived from *kayekTaa* (some) and *kayekTi* (some) respectively. Moreover, the plural markers are suffixed to both personal and impersonal pronominal roots. Among them *-kaTaa, -kaTi, -guli, -gulo, -gulaa* are used with impersonal forms while *-raa*, *-diga*, *-der* are attached to personal forms. The number markers for pronouns in Bangla are give in table 2 below.

	Singular	Plural
	-Taa	-eraa
	-Te	-raa
	-khaanaa	-kaTaa
	-Ti	-kaTi
	-Tuku	-guli
	-khaani	-gulaa
		-gulo
		-diga
		-der

Table 2. Number markers used with pronouns.

Bangla pronouns have a peculiarity with reference to their use of number markers. A singular stem can use a plural suffix e.g. *aamaargulo* (of mine) while a plural stem can use a singular suffix e.g. *taaderTaa* (of them). Therefore, the final determination of the number of a pronominal form depends on the number of the marker. That is why forms like *aamaargulo* are considered plural forms whereas forms like *taaderTaa* are considered singular. This kind suffix information is clearly of importance in the process of developing an automatic morphosyntactic analysis of Bangla.

3.3 Emphatic markers

Two emphatic particles are found in the corpus: *-i* and *-o*, which are normally the terminal morpheme of a word form, as in *tinii / tinio* (he himself, honorific) or *sei / seo* (he himself, non-honorific). However, there is another emphatic particle *-to* (indeed), which may appear at the end of pronouns with or without emphatic particles. The addition of this particle forms pronouns with double emphatic markers such as *tinito / tiniito / tinioto* (he himself indeed, honorific) or *seto / seito / seoto* (he himself indeed, non-honorific).

3.4 Case

The suffixation of case markers to stems is an important aspect of Bangla word formation. For pronouns, the use of case markers is almost identical to that of nouns. In the corpus, markers for nominative, accusative, genitive and locative cases are used with pronouns. There are no separate markers for the dative and oblique cases, rather, the accusative case marker is used to denote these cases also. This means that, with respect to the identification of the accusative, dative and oblique case, there is a notable ambiguity that a context free analysis will be unable to resolve in Bangla.

The cases marked in Bangla use 20 separate case markers. Among them, the marker *-der* is used in the genitive case e.g. *taader* (of them) and denotes plurality. Table 3 below gives a list of case markers used with Bengali pronouns found in the corpus data.

Nom.	Acc.	Gen.	Loc.
-Ø	-(e)re	-(aa)r	-(e)te
–e	-ke	-er	-y
–y	-y	-r	-te
–ke	-e	-der	-e
-re	-ere		-ite
	-re		

Table 3. Case markers used with pronouns.

The analysis of Bangla pronouns presented in this section allowed the division of the pronouns examined into three separate types. These pronouns could then be subclassified based upon such features as number and case. Additionally, the analysis of the process of suffixation allowed a prediction to be made of the sequencing of suffixes appended to a pronoun root in Bangla. On the basis of this analysis and information a system was developed to allow for the automatic identification and analysis of pronouns in Bangla.

4. Automatic processing of Bangla pronouns

A number of computational models of morphological processing have been developed for European languages to date. For the processing of English words a two level morphological analysis scheme was proposed by Kartunen and Wittenburg (1983) while Koskenniemi (1983, 1984) and Koskenniemi and Church (1988) used a finite state automata based approach to morphological processing for analysing Finnish words. Similarly, Lun (1983) worked on a two-level morphology of French words. Work on Bangla is far rarer, however, and the only attempt at morphological processing of any note to date was the directed acrylic graph based approach proposed by Sengupta and Chaudhuri (1993). However, this approach was proven to be rather complex and far from robust. Consequently there have been further proposals, firstly from Chaudhuri *et al.* (1997) and then from Sengupta (1999) aimed at improving morphological processing for Bangla. Chaudhuri *et al.'s* work suggested that a Trie-structure based approach, exploiting phonemic transcriptions, would be advantageous because it promised robustness, simplicity and computability. However, Sengupta (1997, 1999) argued that a grapheme-based approach (GSMorph Theory) was advantageous for the processing of Bangla words. On the basis of the analysis of Bangla pronouns undertaken using the corpus data available, it was decided to combine the suggestions of Chaudhuri and Sengupta and take an approach to Bangla morphological analysis that used both the grapheme-based and the Trie-structure approach.

The implementation of Chaudhuri *et al.'s* proposals required some adaptations to be made to their proposals. Specifically, changes were made because the Trie-based approach was intended to work on phonemic transcriptions, yet the pronunciation of Bangla words can be reasonably estimated from conventional spelling because of the Bangla writing system. The necessity for the generation of time consuming phonemic transcriptions therefore seems unnecessary, and consequently the system outlined here works on text corpora. Given that this observation about the Bangla writing system applies equally to a number of other South Asian writing systems derived from Sanskrit, the method outlined here could potentially be applied to other inflectional South Asian languages such as Hindi, Gujarati and Panjabi.

Based upon the analysis of the Bangla pronouns undertaken on the corpus data, two resource files were generated to aid in the process of the automated analysis of Bangla pronouns. Firstly, a root lexicon, which stores all of the pronoun roots identified in the corpus, and secondly a suffix lexicon, which stores all of the suffixes observed in our training data. Both lexicons carry main entries (roots or suffixes) and each main entry is associated with corresponding attributes (e.g. singular/plural, case marking etc). Additionally, the lexicon entries carry markers to indicate which suffixes may be attached to which roots, as well as information showing the order in which suffixes should appear. Using these lexicons a process of analysis was developed which is largely that of Sengupta. However, to ensure fast access, the root lexicon and the suffix lexicon are kept in two different Trie structures as proposed by Chaudhuri. The rest of this section outlines the processing system in some detail.

The system developed exploits the root and suffix lexicons in order to search for and analyse pronouns in unconstrained Bangla text. Consequently, the first task of the processing system is to identify whether any given word is a pronoun or not. The system adopts a left to right analysis of a word-form, attempting to match the string in question with the characters of a known pronoun root. Consider, for instance, that ***W*** represents a string of characters. For non-inflected pronouns the algorithm works as follows: the system tries to match the string ***W*** with a pronoun string in the root lexicon. If a complete match is found then ***W*** is accepted as a pronoun and there is no need for further processing. If a complete match is not found, the word-form is then subject to further analysis in order to discover whether it is an inflected form.

The process of determining whether a word is an inflected pronoun is as follows. If a sub-string of ***W*** matches with a string of the pronoun root lexicon, then its pronoun class is noted and the corresponding suffix lexicon file is examined in order to determine whether the rest of the sub-string of W matches with a suffix string or strings of the suffix lexicon. If a complete match is found, the analysis is considered successful. ***W*** is accepted as a valid inflected pronoun and its associated semantic and other information is returned from the root lexicon, as well as its possible suffixes from the suffix lexicon. If no match is found in the suffix lexicon then ***W*** is rejected and a new word is selected for processing. The information accumulated can be stored for further levels of automatic analysis and processing.

This algorithm is heavily dependent upon the coverage provided by the root and suffix lexicon - if the lexicons contain all, or the great majority, of Bangla pronoun roots and suffixes, such a system could be quite successful. If, however, the corpus used yielded only a limited subset of Bangla pronoun roots and suffixes, then the resulting system would not be very successful at all. Consequently, the system was tested on the 500,000 word test corpus in order to explore the robustness of the system developed. The algorithm performed well, identifying and processing 98% of the pronouns used in the test corpus successfully. Table 4 below gives an example output from the system.

Word form : *aamaaderTaakei*		
Root part:	*aamaa-*	Suffix part: *-derTaakei*
Person:	First	Number: *-Taa* (Sing)
Hon.:	Null	Particle: *-i* (emphatic)
Case:	*-ke* (Acc)	Meaning: to our

Table 4. An example output from the system.

Of the two percent of errors, some of these were caused by ambiguity irresolvable at the context free level. This may occur where a word form is polysemous, such as the word form ye which may be a pronoun though may also belong to the lexical category indeclinable. Similarly, a context free analysis may misrepresent an adjectival pronoun as a non adjectival pronoun. An example of this is found in the demonstrative pronouns such as *ei* "this", *sei* "that". These are generally used as demonstrative adjectives in the text, but where they are used immediately before nouns as in *sei din* (that day) or *ei kathaa* (this word) they change their lexical categories to adjectival pronoun, though this is not marked in the word form at all. Consequently, the need to scale up to a context bound analysis, rather than lexicon coverage, is the overwhelming cause of the 2% of errors that the system displays.

5. Conclusion

A corpus based approach proved very useful for the purposes of building a pronoun identifier and analyser for Bangla. As well as identifying a new pronoun in Bangla, the corpus based analysis revealed some basic data about the distribution of inflected and non-inflected pronoun forms, and provided a sound basis for a comprehensive description of the morphology of Bangla pronouns. This description in turn was the basis for the development of a robust system for pronoun identification and analysis for Bangla. As such, the work outlined in this paper shows not only the promise of corpus linguistics in the generation of improved descriptions of languages, but also shows the usefulness of the generation of such descriptions for language engineering.

Acknowledgements

The paper was first presented at the *International Conference on South Asian Languages (ICOSAL-II)* at the Punjabi University, Patiala, India in January, 1999. The modified version of the paper was produced after the comments of unknown reviewers were taken into account. These reviewers deserve thanks for their contributions. I would also like to thank the Ministry of Information and Technology (MIT), Govt. of India, for supplying the Bangla text corpus. The comments of M.K. Nath of Calcutta University and of Tony McEnery of Lancaster University, UK are also acknowledged with thanks.

References

Aronoff, M. (1981). *Word Formation in Generative Grammar.* Cambridge, Mass.: MIT Press.

Bloomfield, L. (1933). *Language.* New York: Holt, Rinehart & Winston.

Chaki, J. B. (1996). *Bangla Bhasar Byakaran* (The Grammar of the Bangla Language). Calcutta: Ananda.

Chatterji, S. K. (1926). *The Origin and Development of the Bengali Language.* Calcutta: Calcutta University Press.

Chatterji, S. K. (1993). *Bhasa Prakash Bangla Byakaran* (The Grammar of the Bangla Language). Calcutta: Rupa.

Chaudhuri, B. B., N. S. Dash and P. Kundu. (1997). "Computer Parsing of the Bangla Verbs" *Linguistics Today* 1(1): 64-86.

Dash, N. S. and B. B. Chaudhuri (2000) "The Process of Designing A Multidisciplinary Monolingual Sample Corpus". *International Journal of Corpus Linguistics*, forthcoming.

Garside, R., Leech, G. & Sampson, G. *The Computational Analysis of English*, Longman: London, 1987.

Joshi, A. and B. Srinivas. (1994). Disambiguation of Super Parts of Speech: Almost Parsing *in Proceedings of COLING-94*. Kyoto, Japan.

Karlsson, F., A. Voutilainen, J. Heikkilä and Anttila, A. (1995). *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin.

Koskenniemi, K. (1983) "Two Level Model for Morphological Analysis" in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAL-83),* Karlsruhe, West Germany, pp. 683-685.

Koskenniemi, K. (1984). "A general computational model for word-form recognition and production" in *Proceedings of Conference on Computational Linguistics (COLING-84).* Stanford, CA, 1984. pp. 178-181.

Koskenniemi, K. and K. W. Church. (1988) "Complexity, two-level morphology and Finnish" in *Proceedings of Conference on*

Computational Linguistics (COLING-88), Budapest, 1988 pp. 335-340.

Kartunen, L. and K. Wittenburg (1983) "A two-level morphological description of English" in *Proceedings of Advanced Computational Linguistics (ACL-83), 23rd Annual Meeting*. 1983, pp. 217-228, also in *Texus Linguistic Forum* 22: 227-238. TX 78712.

Lun, S. (1983). "A Two-level Morphological Analysis of French". *Texus Linguistic Forum.* 22: 271-278.

Sarkar, P. and G. Basu. (1994). *Bhasa Jiggnasa* (Queries of Language). Calcutta: Vidyasagar Pustak Mandir.

Sen, S. (1993). *Bhasar Itivritta* (The History of Language). Calcutta: Ananda.

Sengupta, G. (1997) "Three models of morphological processing", *South Asian Language Review*, 7(1): 1-26.

Sengupta, G. (1999). "GSMorph: A Grapheme Oriented Structuralist Morphological Processor". *Presented in ICOSAL-II*, Punjabi University, Patiala. (MS).

Sengupta, P. and B. B. Chaudhuri. (1993). "A Morpho-Syntactic Analysis Based Lexical Sub-system". *Int. Journal of Pattern Recognition and Artificial Intelligence.* 7(3): 595-619.

Sengupta, P. and B. B. Chaudhuri. (1993). "Natural Language Processing in an Indian Language (Bangali)-I: Verb Phrase Analysis". *IETE Technical Review.* 10(1): 27-41.

Spencer, A. (1991). *Morphological Theory*. Oxford: Basil Blackwell.

Notes:

1. For the purposes of this paper, the term word refers to a string of graphemes as it appears delimited by white space in written language.
2. In this transcription of Bangla, the capitalised T represents the retroflex form of the phoneme /t/.