Some Techniques Used for Processing Bengali Corpus to Meet New Demands of Linguistics and Language Technology

Niladri Sekhar Dash

The utility of a language corpus is drastically enhanced when it is properly processed in various ways for retrieving relevant linguistic information to be used in language description and analysis as well as in various applications related to applied linguistics and language technology. Unfortunately, the text corpora developed for the Indian languages are not yet processed properly for making them useful for the tasks related to both mainstream linguistics and natural language processing. Keeping this in mind I present here in brief a few techniques of Bengali text corpus processing, which we use for various linguistic activities. These techniques, however, become far more complicated due to orthographic, morphological, and lexicological complexities involved in the language.

Keywords: frequency counts, concordance, collocation, key-word-in-context, corpus, local-word-grouping, lemmatization, parsing, language technology, applied linguistics, Bengali

1. Introduction

The Bengali text corpus (Dash 2005) — after it is generated in electronic form — is used in various ways for several text processing works. There are many text techniques, which are often run on present-day electronic corpora, but were seldom used on corpora manually compiled in the earlier days. These corpus processing techniques are the outcomes of close interface developed between language databases and computer. These are used to analyse text corpora from newer perspectives, give new directions in the study of a natural language and to shed lights on the language properties found in the corpora.

In reality, application of corpus processing techniques on corpora results in finding out new evidences, which are furnished to describe a language and its properties from new perspectives. There are other advantages also in utilization of these techniques both in mainstream linguistics and language technology. By applying these, we gather examples to provide explanation that may fit into observations, rather than adjusting evidences to fit our pre-supposed explanation. Our experience shows that processing of Bengali corpus produces results, which directly contradict our intuition-based observation about the language and its properties.

The application potential of linguistic information in developing tools and systems for language technology motivates us to develop corpus processing techniques. These include systems like statistical frequency counting, lexical collocation, concordance, key-word-in context, local word grouping, parts-of-speech tagging, morphological processing, lemmatisation, annotation, parsing, etc. These methods can run automatically on corpus for obtaining data and information required for designing systems for language technology in Bengali. Thus, these techniques become instrumental to open up new avenues to language research and applications unknown before.[1]

2. Frequency Counts

The study on the frequency of use of various linguistic elements in a corpus is of great value in understanding of a language as well as for developing systems for language technology. Also, information of this kind becomes relevant in language teaching because at the earlier stage of language teaching learners need to be presented with sets of common linguistic items along with the patterns of their use for enhancing their linguistic competence and performance. For instance, in 'usage-based model' of language learning it has been argued that information about the frequency of use of various linguistic items has direct effect on language education (Johns 1991). This had led some scholars to argue that:

> ...while frequency data is presumably of minor importance in a parameter-setting model of language learning in which the data has only the 'triggering' function in grammar formation, frequency is very important in an alternative conception of grammar formation based on a model of grammar which is 'use-based'. Such models assume that grammar formation is inductive to a large extent, and that the frequency of the linguistic usage events has a direct effect on the form of the grammar (Barlow 1996: 5).

The information available from frequency lists of words is usually rendered either alphabetically or numerically. The lists are arranged either in ascending or in descending order based on the needs of the target users. In essence, a frequency list provides necessary clues to know how words actually occur in a language with regard to their recurrence patterns. By examining the lists, we get rudimentary ideas about the basic structure of a language to plan our future course of investigation and analysis accordingly.

Frequency count also becomes useful to project into the patterns of distribution of words and other lexical units within a piece of text. It is known to us that patterns of distribution of words not only shade light on the overall discourse structure of a text but also focus on its basic ingredients related to content, target readership, type, function, addresser, addressee, etc. Scholars have observed that "...if in a technical text, there is very little technical vocabulary for some time and then a rush of it, that may be a clue to a high-level structural boundary in the text, perhaps the end of a general, layman's introduction to a technical subject" (Sinclair 1991: 30).

We have observed that two synonymous words (say, W_1 and W_2) in the Bengali corpus have same frequency of occurrence, but while W_1 occurs at the first section of a text, W_2 occurs at the last section. This gives us vital clues to furnish suitable interpretations to their patterns of occurrence in the text. It also helps us to select and design texts for Bengali teaching in a more efficient way than traditional manners where selection of texts is randomly made by intuition. Quite often, a randomly selected piece of text appears to be an introductory matter, but further analysis of distribution of words occurring in subsequent sections turns it into a typical technical text rich with complex ideas and analysis. In general, frequency count is of two types:

(a) Alphabetic frequency count, and
(b) Numerical frequency count

In true sense, however, these are not two types. Rather, these are two different modes of display of the same results. In case of alphabetical frequency listed items (e.g. characters,

words, idioms, phrases, etc.) are arranged in alphabetical order while in numerical frequency same items are arranged according to their degree of occurrence in the in the text (from high to low or from low to high). For instance, let us consider a list of some Bengali words arranged in alphabetical order along with their frequency of occurrence in the corpus (Table 1).

Word	Percentage	Word	Percentage
ār	0.513 %	tār	0.497 %
ei	0.939 %	theke	0.549 %
ek	0.367 %	nā	1.153 %
ebaŋ	0.653 %	bā	0.419 %
o	0.910 %	yā	0.404 %
karā	0.408 %	ye	0.653 %
kare	0.989 %	sa'nge	0.371 %
kintu	0.423 %	sei	0.321 %
kon	0.358 %	hay	0.764 %
janya	0.322 %	haye	0.394 %

Table 1 *Frequent Bengali words arranged in alphabetical order*

Numerical frequency list, on the other hand, is generated to identify which linguistic items are most frequent as well as which items are least frequent in use in a language. The information is sorted in such a way that the list begins with the most frequent item to end with the least frequent one. Generally, in a corpus of hundred thousand words a frequency list is too small to be interesting and faithful. But in a corpus of billions of words frequency list in numerical order is useful for provide interesting insights into the language, because the list of items in particular order becomes comparable to a large population of samples for statistical measurement and analysis. For instance, let us consider the list of some of the most frequent Bengali words arranged in their alphabetical order (Table 2).

Word	Percentage	Word	Percentage
nā	1.153 %	kintu	0.423 %
kare	0.989 %	bā	0.419 %
ei	0.939 %	karā	0.408 %
o	0.910 %	yā	0.404 %
hay	0.764 %	haye	0.394 %
ebaŋ	0.653 %	sa'nge	0.371 %
ye	0.653 %	ek	0.367 %
theke	0.549 %	kon	0.358 %
ār	0.513 %	janya	0.322 %
tār	0.497 %	sei	0.321 %

Table 2 *Bengali words arranged in numerical frequency order*

It has been observed that normally, the most frequent linguistic items tend to keep suitable distance in their distribution. As a consequence, many marked changes in the order of their distribution become quite significant both in linguistic analysis and generalization.

For instance, the 'asymmetrical frequency' shows that some highly common words make up a high percentage in all text types while large number of less frequent words make up the rest (Zipf 1949: 173). This signifies that while highly frequent words are attested even in a small corpus, less frequent words will not occur in a corpus unless it is made with large amount of samples obtained from various text types of language and represented in the corpus in a balance manner.

The usability of the two types of frequency list is also different. In case of alphabetical frequency list, items are normally displayed in a tabular form for general reference in descriptive linguistic studies. It plays secondary role in language analysis, since it is referred to only when there is a need to check frequency of particular items in a language. However, it becomes helpful in formulating hypotheses to be tested as well as for checking prior assumptions (Kjellmer 1984). Similar frequency counts are also used to evaluate the patterns of use of words, compounds, idiomatic expressions, phrases, collocations, and other multiword units in a language.

Similar to the phenomenon of usage of words, we use frequency list to measure patterns of meaning of the lexical items in a language. We find that individual meanings of words have a frequency curve, where the most common meaning of a word occurs more often than the least common meaning. Similarly, certain words and phrases may occur more frequently in particular text while others occur in another kind of text. Such observations work also in case of multiword units, sentences and other grammatical properties used in a language. The basic point of our argument is that frequency information, run on a well-designed and properly balanced corpus, reveals accurately many new patterns of use of linguistic items to understand the language in a new perspective.

Information about the frequency of use of various linguistic items carries utmost importance in language teaching. In fact, it makes sense to observe frequency of occurrence before selecting examples of linguistic items for reference and use in designing grammars and course books for language teaching (Wills 1990: 142). Thus, corpus-based frequency counts register their functional as well as referential advantages over intuitive ways of language description and analysis, which often tends to furnish evidences collected through intuition or assumption.

3. Concordance of Words

Technically, concordance is a process of indexing words used in a piece of text. It enables us to display the total list of occurrence of a word — each occurrence in its own contextual environment. Since words in concordance are indexed with close reference to the place of their occurrence in a piece of text, it helps us to know possible range of usage varieties of words in a corpus. Thus, concordance becomes indispensable in corpus analysis because it gives us better scopes to access possible patterns of word use in a language.

Although the scholars of earlier ages felt a genuine need for such a technique for understanding distributional and semantic patterns of words in a language, they had no opportunity to acquire it as they had no computational device under their disposal that could automatically arrange words collected from a corpus database in a desired manner for subsequent analysis and observation. Introduction of computer technology and electronic corpora has made concordance an easy process to compile and arrange words from large corpora within a short span of time. Due to flexibility in the technique, determination of

contextual frame of words may vary depending on various criteria, such as fixed number words on either side of the target word, finding sentence boundaries of the target words, etc.[2]

The application of a concordance technique on a corpus, either small or big, allows us to understand the variations of linguistic features, since it gives us scope to study lexical, semantic, syntactic patterns of words as well as genre and type of a text (Barlow 1996). In case of language education, this technique helps the learners to access and understand language properties in their syntagmatic and paradigmatic frames. With maximum sorting option for both left and right sorting of a word it becomes useful tool to investigate if words are polysemous with a range of multiple senses embedded within a single surface form. In the area of 'data-driven learning', various lessons on grammar and vocabulary become highly helpful if these are made with concordance-based materials complied from a corpus. Johns (1991) argues that in 'data-driven learning', attempt should be made in order to

> …cut out the middleman as far as possible and give direct access to the data so that the learner can take part in building up his or her own profiles of meanings and uses. The assumption that underlies this approach and that effective language learning in itself is a form of linguistic research. The concordance printout offers a unique resource for the stimulation of inductive learning strategies — in particular the strategies of perceiving similarities and differences and of hypothesis formation and testing (Johns 1991: 30).

Although the technique of concordance was not available in ages, some scholars diligently did the work manually on small corpora to come out with new evidences and observations. In late 18th century, the corpus of the Bible was processed to generate word lists, lists of lexical collocation, and lists of word concordance to prove factual consistencies within various parts of its text. In 1769, Alexander Cruden, a British publisher, produced concordance list of words on an authorized version of the Bible, which was considered as one of the monumental pieces of laborious scholarship in English language. It included list of concordance of major content and function words, as well as lists of lexical collocations (Kennedy 1998: 13). Similar attempts were also made on the works of Shakespeare (Gibson 1962, Elliott and Valenza 1996), Milton, and others.[3] In the table below (Table 3) we present a sample list of concordance of a Bengali word *mānuṣ* "man" to show how the word used and how the word varies in sense due to its occurrence in different contexts.

In general, application of a concordance program on a corpus yields varieties of information, which are not available via intuition. Due to this excellent quality, this technique is frequently used in the work of dictionary compilation to search out words, compounds, idioms, and multiword units from corpus along with contexts of their occurrence. Sometimes, the technique is complemented with a range of statistical tools that provide information about relative frequency of the items, their distributions across text types, and the list of lexical items with which they are most likely to occur in a piece of text. That means with the help of a concordance program, it is no more a difficult task for us to examine all types of occurrence of all linguistic items in a corpus to describe a language or language variety with new insights and examples.

araṇyacārī	mānuṣ	kṣudhār samay phalmūl kheto
jaler janya	mānuṣ	nadītīre basabās karta
ŷaubane pā dile	mānuṣ	nānā rūp dhāraṇ kare
sei daler	mānuṣ	hala debdās o romio
kī kare pṛithibīte	mānuṣ	elo tā aneke bhebechen
kathā hala banmānuṣ ār	mānuṣ	duṭo biśeṣ dharaner prāṇī
ākṛṣṭa haye grāmer	mānuṣ	bhiṛ kare elo
sei maner	mānuṣ	āchen mānuṣer mājhe
uttar paścimer anek	mānuṣ	māch dharār jībikā niyeche
choṭabelā theke se	mānuṣ	hayechila māsīr kāche
ek caṛei se anya	mānuṣ	haye geche
bhāluk marā	mānuṣ	khāy nā balei jāni
tāderke ṭhik	mānuṣ	balā yāy kinā bhebe dekhben
śudhu śarīr diye to ār	mānuṣ	hay nā, maner diko āche
sei samykār guhābāsī	mānuṣ	saṇkhyā hisābe cinta kichu dāg
jānā geche erā ādhā	mānuṣ	ādhā garilā jātīya jīb
pān biṛi sigāreṭ	mānuṣ	abhyāsbaśata bhog kare

Table 3 *Concordance list of mānuṣ 'man' from the Bengali corpus*

4. Lexical Collocation

Lexical collocation has been defined as the "occurrence of two or more words within a short space of each other in a text" (Sinclair 1991: 170). The technique for identifying lexical collocation in a piece of text is considered an important method for evaluating the value of consecutive occurrence of any two words in a piece of text. In return, it projects into the functional nature of the lexical items used in a language as well as on the "interlocking patterns of the lexis" in a text (Williams 1998). In analysis of lexical collocation, we are normally interested to know to what extent the actual patterns of lexical occurrence differ from the patterns that have been expected (Barnbrook 1998: 87). This measurement is also used to evaluate the argument that claims that our mental lexicon is made up not only with single words but also with larger multiword units — both fixed and variable.

The technique of lexical collocation, when it runs on a large corpus database, produces various kinds of information about the nature of collocation words in a language. In fact, systematic analysis of the process of collocation helps us to understand the position and function of words that often participate in collocation in a language. Usually, the list of examples about contextual use of words obtained through the program of concordance on a corpus often contains some preliminary information about the patterns of lexical association of words needed for analysing the nature of lexical collocation. The list of lexical collocation also includes information about the frequency of words used in collocation as well as specific statistical counts used to calculate the figures needed for comparison and authorization of the examples of collocation.

Lexical collocation is a well-known linguistic phenomenon of a natural language. It is discussed in full length with evidence carefully selected from a language. For instance, in the Bengali corpus, the adjective *kā̃cā* 'raw' is found to be associated with more than thirty

different words to denote equal number of collocation and sense variation. Without reference to their frequency of use, it is however, impossible to understand all the finer aspects involved in the distribution of these words. In the examples given below all possible sense variations of the word are taken into consideration to make distinction among the senses implied by the word used in different contexts of lexical collocation in Bengali.

(1)
Word : kc̃ācā
Default word class : Adjective
Primary meaning : "raw"
No. of sense variations: 30+
Examples: *kc̃ācā phal* "unripe fruit", *kc̃ācā māch* "raw fish"', *kc̃ācā māŋsa* "raw meat", *kc̃ācā iṭ* "un-burnt brick", *kc̃ācā rāstā* "earthen road", *kc̃ācā ghar* "mud house", *kc̃ācā kathā* "initial talk", *kc̃ācā bhāşā* "obscene word", *kc̃ācā khisti* "slang", *kc̃ācā sabji* "green vegetable", *kc̃ācā māthā* "young brain", *kc̃ācā lok* "novice", *kc̃ācā hāt* "new hand", *kc̃ācā rasid* "primary draft", *kc̃ācā kāj* "useless work", *kc̃ācā ra'n* "washable color", *kc̃ācā sonā* "pure gold", *kc̃ācā cul* "black hair", *kc̃ācā kāṭh* "wet log", *kc̃ācā prem* "calf love", *kc̃ācā ojan* "less weight", *kc̃ācā paysā* "easy money", *kc̃ācā ghum* "incomplete sleep", *kc̃ācā bayas* "immature age", *kc̃ācā māl* "raw material", *kc̃ācā mukh* "filthy mouth", *kc̃ācā lekhā* "bad writing", *kc̃ācā kalā* "green banana", *kc̃ācā ŷauban* "early adulthood", *kc̃ācā jal* "non-boiled water", *kc̃ācā hiśāb* "initial estimate", *kc̃ācā rod* "rays of early sun", etc.

The examples of collocation cited above signify that with reference to the contexts of use of words in a piece of text we can empirically determine which pairs of words maintain substantial collocational relation between them. The most common formula we use here is a method of 'mutual information' that helps us to compare probability of any two words (W_1 and W_2) occurring together as an event with their probability of occurrence as a result of chance. For each pair of words, we derive a statistical score from the corpus to calculate that where there is higher score, the greater is the possibility of lexical collocation. Thus, the method of reference to 'mutual information' becomes useful in evaluation of lexical collocation in Bengali for the following reasons:

(a) Empirical information of lexical collocation enables us to extract multiword units from the corpus to compile dictionary of lexical collocation, develop databases of translational equivalents, and to design text materials for language education.

(b) It helps us to group all collocations of a word together to identify the range of its sense variation as well as to know how it is able to generate new senses by collocating with new words. This, in return, directs us towards the phenomenon of semantic gradience of words (Leech, Francis, and Xu 1994). For instance, in Bengali, the word *mukh* 'mouth' collocates with *bandha* 'closed' to refer to 'introduction', with *pātra* 'person' to refer to "spokesperson", with *patra* 'leaf' to mean 'manifesto', with *jhāmṭā* 'rage' to mean 'rebuke', with *rocak* 'taste' to mean 'tasteful'.[4] Thus, in each case, the core sense of W_1 (i.e. *mukh*) is changed due to its collocation with a different word (i.e. W_2). It signifies that proper understanding of the actual meaning of W_1 we require to refer to the meaning of W_2 also.

(c) Understanding lexical collocation of a language also supports to understand and identify the differences of use of the synonymous words (different in form but similar in meaning) in a language. For instance, although words *strong* and *powerful* are similar in sense, mutual information score obtained form their association with other words reveals some interesting differences in English. While *strong* often collocates with *motherly, showings, believer, currents, supporter*, *odour*, etc. *powerful* usually collocates with *tool, minority, neighbour, symbol, figure, weapon, post*, etc. to denote sense variation in their distribution (Church et al. 1991).

(d) Finally, lexical collocation helps us to understand the nature and pattern of grammatical association of two synonymous words in a language. For example, grammatical association of *little* and *small* in the *British National Corpus* exhibits that *little* tends to co-occur with concrete, animate nouns such as *things, boy(s), girl(s),* etc. while *small* co-occurs with nouns that tend to indicate *quantity, amount*, *number*, *proportion*, etc. (Biber, Conrad, and Reppen 1998: 94).

While exploring the nature of association of two nearly synonymous words (i.e. *din* and *divas* 'day') in Bengali, we observe some really interesting patterns of lexical association. While *din* mostly co-occurs with regularly used words having an informal sense such as *janma din* 'birth day', *kājer din* 'work day', *chuṭir din* 'holiday', *barṣār din* 'rainy day', etc., *divas* normally co-occurs with words having a flavour of formal dignity such as *śramik dibas* 'May Day', *śiśu dibas* 'children's day', *svādhīnatā dibas* 'independence day', *śahīd dibas* 'Martyr's Day', *mṛtyu dibas* 'death day', *mātṛ dibas* 'mother's day', etc. Thus, empirical analysis of patterns of lexical association with examples obtained from the corpus shows that some synonyms may have important differences in their grammatical and lexical association resulting from differences in distribution across discourse types. In essence, analysis of examples of lexical collocation shows that nearly synonymous words are rarely equivalent in function when considered in terms of their distribution in the text.

Detailed information about the delicate differences of collocation of any two synonymous words is an important input for students in their way of learning a language in advanced stages. Moreover, this information becomes important input for the people working in the area of dictionary and thesaurus compilation, machine translation, speech and language processing, and similar works, although for the non-native speakers it is not easy to determine which collocation is a significant phenomenon of a language.[5]

5. Key-Word-In-Context

Key-word-in-context (KWIC) is another technique of text display, which is also widely used in corpus processing. Technically, it is another format of word concordance, which saves us from looking up each occurrence of particular words in the corpus. However, it differs from concordance in the perspective that while in concordance the target word under investigation is the central point of attention in case of KWIC it is the environment that arrests our attention.

In case of KWIC, generally, the target word appears at the centre of each line with extra space on either side of the word where length of the sentence is previously specified by investigators. The system thus displays an environment of two, three, or four words on either

side of the target word located at centre. Since the pattern of word presentation may vary according to our need, we can ask computer to provide relevant citation of a word according to the specifications we have determined previously. In general, a KWIC technique performs the following things quite usefully for us:

- It helps to identify all the occurrence variations of the key word we have selected in a corpus, and
- It presents results of a search for the contextual environments in a way that may help to define usage patterns of the key word with regard to its contexts.

The success of the process in finding out occurrences of key words in a corpus is based on processing efficiency of the system. The display method, however, relies on the standard display options used in most of the concordance packages.

The KWIC method registers some advantages over concordance in mode of display because it allows us to select which factors are to be analysed to detect changes found in context. The central block of display, occupied by the key word, captivates our eyes for scanning the lines to identify the context patterns (Barnbrook 1998: 69). Due to this advantage, most of the KWIC techniques use various display options in text representation. For example, while some use sentential context in KWIC format, other may use paragraph or the whole text. This liberty in selection for variable length format allows adjustment of the size of the search of text, within which the key word is entered, in proportion to size and display facilities provided in computer. In case of context determined by sentence and paragraph, it places the keywords in sentences and paragraphs in which key words occur.[6] Thus, the facility to browse the contexts of a whole text allows us to move backwards and forwards from the point of occurrence of the key word and permits us to access, as much we require, for checking details about the usage patterns of the key words.

The access of information from a corpus through KWIC helps us to formulate various objectives in description of the words as well as in devising procedures for pursuing these objectives. For instance, the execution of a KWIC the *Bank of English* reveals that in English, as the following examples show, the most frequently used verb with a reflexive form is *find* followed by *see*, *show*, *present*, *manifest*, and *consider*, etc. all of which refer 'viewing' as a part of representation or proposition (Barlow 1996).

(2) a. I always *find myself* in trouble.
    b. Better *see yourself*.
    c. *Show yourself* the path.
    d. *Present yourself*.
    e. It was *manifested* in *itself*.
    f. I *consider myself* fortunate.

Thus KWIC helps to understand the importance of context in analysis of words as well as to estimate the role of associative words in sense variation. It also helps us to explore the actual behaviour of words in context-bound situations, decipher actual environment of occurrence of various language properties, and to evaluate contextual restrictions exercised in use of various language properties in speech and writing (Sardinha 1996).

KWIC technique is found to be convenient and useful in analysis of idioms, phrases, clauses, and proverbial expressions, which require additional texts and contextual information

for their understanding. Such facilities lead us to think that a KWIC output is a text in itself, which we can use separately to examine the frequency and pattern of the words that occur within the environment of the key word. It is not that we use total information extracted from contexts every time but we can use it in description of key words as and when required.

6. Local Word Grouping

The process of local word grouping (LWG) is another important way of corpus processing and text analysis, which unlike concordance and KWIC, aims at throwing lights on patterns of use of words, idioms, phrases and other language properties from different perspective. We find that LWG technique becomes useful in those places where word order is an important aspect for determining semantic information of a sentence, and where semantic information of individual constituent affects or is affected due to presence of another constituent in the sentence.

The LWG technique provides valuable information to deal with functional behaviour of the constituents at the time of parsing — both at phrase and sentence level. For instance, information obtained from LWG run on the *British National Corpus* shows that verb *manifest* is mostly associated with third person neuter reflexives, whereas *enjoy* occurs with all reflexive forms except neuter gender (Barlow 1996). Using same method, the distribution of verbs like *amuse, please, lend, remind*, and others, which are not very common in use but have a special kind of affinity for reflexive forms, have been elaborately studied with illustrative examples for the corpus. The most striking advantage of this technique is that linguistic knowledgebase acquired about the patterns of use of rare lexical items in a language become highly fruitful for moving the language learners from intermediate to more advanced levels of their linguistic proficiency.

In the Bengali corpus we have noted that all non-finite verbs are most often followed by finite verbs while nouns are mostly followed by postpositions. These so called verb and noun groups, which we can easily retrieve by LWG run on the corpus, are best analysed by using local information, which in return, supplies contextual clues for understanding their functions as idioms, phrases and set expressions. Also, information extracted from the corpus through LWG becomes useful for dissolving lexical ambiguities that arise from association of various lexical items within local contexts (Miller and Leacock 2000: 156). It implies that finer shades of meaning are often linked with the association of specific lexical units within a word group.

It also suggests that finer shades of meaning become explicit by the internal relations underlying between the members of the group along the line of their occurrence in many contexts. For instance, in case of compounds, idioms and set expressions, meanings denoted by a particular association of words are not obtainable from meanings of individual words put together in a random manner. Therefore, for understanding and translating these multiword units, meanings of the related members need to be grouped together. And this task is best possible by way of suing LWG technique on a corpus.

7. Lemmatisation of Words

The term 'lemma' refers to the basic form of words disregarding their grammatical changes such as tense and plurality (Biber, Conrad and Reppen 1998: 29). Lemmatisation involves identification of part-of-speech of words used in a piece of text and reducing them to their respective lexemes — the headword that we look for in a dictionary (Dash 2006). For various works related to corpus processing (e.g. statistical counts, concordance, numerical sorting, lexical collocation, etc.), this process is indispensable to group together different types of inflected and affixed forms of words, so they are collectively displayed under one head (Barnbrook 1998: 50). In the list given below we show an example of lemma along with a list of inflected forms of the word collected from the Bengali corpus.[7]

(3)
Lemma : kathā 'word'
Inflected forms: *kathā, kathāi, kathāo, kathāte, kathātei, kathāteo, kathāke, kathākei, kathākeo, kathār, kathāri, kathāro, kathāṭi, kathāṭii, kathāṭio, kathāṭir, kathāṭiri, kathāṭiro, kathāṭite, kathāṭitei, kathāṭiteo, kathāṭike, kathāṭikei, kathāṭikeo, kathāṭā, kathāṭāi, kathāṭāo, kathāṭār, kathāṭāri, kathāṭāro, kathāṭāte, kathāṭātei, kathāṭāteo, kathāṭāke, kathāṭākei, kathāṭākeo, kathāguli, kathāgulii, kathāgulio, kathāgulir, kathāguliri, kathāguliro, kathāgulite, kathāgulitei, kathāguliteo, kathāgulike, kathāgulikei, kathāgulikeo, kathāgulo, kathāguloi, kathāguloo, kathāgulor, kathāgulori, kathāguloro, kathāgulote, kathāgulotei, kathāguloteo, kathāguloke, kathāgulokei, kathāgulokeo, kathāgulā, kathāgulāi, kathāgulāo, kathāgulār, kathāgulāri, kathāgulāro, kathāgulāte, kathāgulātei, kathāgulāteo, kathāgulāke, kathāgulākei, kathāgulākeo, kathāy, kathāyi, kathāyo*, etc.

By way of lemmatisation we can also assemble all derived forms of a verb (e.g., *calan, calā, calti, calita, calanta, calamān, calamānatā, calişņu, calanśīl, calansai, caleble,* etc.) under a single group to link up with lemma, *cal* "to move". Thus, it enables us to accumulate all variants of a headword without searching through the whole corpus, which otherwise, is a tedious, time consuming and error-prone task. It also helps us to cluster morphologically irregular forms (e.g. *ŷete* 'to go', *gela* 'went', *ŷāba* 'will go', etc.) under one 'head so that all variants belong to a single lemma (e.g., *ŷā* 'to go') for linguistic analysis and interpretation.

Because of so many advantages, we consider lemmatization as an important process in corpus research and application. In the area of vocabulary study and lexicography it allows us to produce frequency and distribution information for lemmas (Sánchez and Cantos 1997). The process of lemmatization is successfully used on several corpora of English for last few years (Beale 1987). For instance, *SUSANNE corpus* includes information where lemmatized forms are displayed parallel to the actual words in a vertical format along with part-of-speech and syntactic information. Also, some parts of the *Brown Corpus* contain lemmatized forms of words along with other lexical and grammatical information. In some recent attempts, some texts of the *CRATER Corpus* of English, French and Spanish (McEnery and Wilson 1996: 43) and the *Frankenstein Text* (Barnbrook 1998: 51) are passed through lemmatization. The process is hardly applied on any of the Indian text corpora developed so far. In the table below we furnish a sample set of lemmatization from the Bengali corpus where lemmatized

forms, their original surface forms as well as their part-of-speech are arrayed in three separate columns (Table 4).

Input text: svādhīnatā lābher par theke gata calliś bachare kendrīya sarkār katakguli bhrāntimūlak nīti anusaraṇ kare esechen

	Lemma	Surface form	Part-of-speech
L	svādhīnatā	svādhīnatā	Noun
E	lābh	lābher	Noun
M	par	par	Noun
M	theke	theke	Postposition
A	gata	gata	Adjective
T	calliś	calliś	Adjective
I	bachar	bachare	Noun
S	kendrīya	kendrīya	Adjective
A	sarkār	sarkār	Noun
T	katak	katakguli	Adjective
I	bhrāntimūlak	bhrāntimūlak	Adjective
O	nīti	nīti	Noun
N	anusaraṇ	anusaraṇ	Noun
	kar	kare	Non-finite Verb
	es (< ās)	esechen	Verb

Table 4 *Lemmatization of words from the Bengali corpus*

8. Parsing Sentences

Parsing is a kind of annotation that is operated on sentences collected in a corpus after the corpus passes through the stages of grammatical annotation. With the help of this technique we carry out some kinds of syntactic analysis of sentences collected in a corpus in accordance with the grammar of a language (Leech and Eyes 1993). The analysis of sentences is done completely automatically or with partial manual assistance or by combining both techniques. The result is an output of an annotated version of a text in which each individual lexical item is tagged with salient (and relevant) grammatical information to exhibit their syntactic functions and relations with other constituents in the sentence.

The notable difference between a tagged and a parsed corpus is that in a tagged corpus lexical items are annotated at lexical and/or semantic level to provide adequate intralinguistic and extralinguistic information about each item compiled in the corpus. On the other hand, in a parsed corpus, information is provided for identifying the structural relationships of words, word-groups, phrases, and clauses etc. used within a sentence (Barnbrook 1998: 127).

Technically, parsing refers to the practice of assigning syntactic structure to the sentences used in a text. It is usually performed in a corpus after the identification of basic morphosyntactic categories of a language. Since identification of basic morphosyntactic categories of a language is not an easy task, it involves automatic context-bound as well as

context-free analysis of the sentences by using necessary linguistic information acquired from processing of words.

Based on different models of grammar analysis, in parsing, we try to elevate all morphosyntactic categories to a higher level of their syntactic relationship they develop with each other.[8] Whatever may be the model, approach, technique or methodology, the basic goals of a parsing technique are the followings:

- Proper identification of words used in a sentence,
- Assignment of appropriate syntactic description to the words,
- Identification of boundary of phrases and clauses,
- Allocation of groups to clause components,
- Grouping phrases and clauses to identify syntactic constituents of a sentence, and
- Naming of the constituents accordingly.

Most of the parsing techniques developed so far for English and other languages aim at using some of the existing linguistic formalism such as principles of Government and Binding theory, context-free Phrase Structure Grammar, Tree Adjoining Grammar, and others to exhibit the inherent syntactic relations underlying the constituents used in formation of a sentence. This has been the normal practice since it is believed that implementation of parsing technique based on certain grammar formalism is far more convenient for encoding syntactic relations of words and useful for achieving higher level of success in sentence processing. However, in case of Bengali, we have noted that application of grammatical formalism does not yield commendable amount of success, which eventually leads us to design a method, which is exclusively dependent on a set of linguistic rules defined manually. A large set of these linguistic rules needs to encode human knowledgebase to develop a creditable parsing technique.

A comprehensive parsing scheme usually assigns phrase marker or labelled bracketing to each sentence of a corpus in the manner of a phrase structure. The resulting 'parsed corpus' is identified as 'Tree Bank' that aims at depicting a map similar to the tree diagram used in phrase structure grammar. Since representation of a tree structure is rare in corpus parsing, identical information is represented using sets of labelled brackets. Thus an English sentence like *Pearl sat on a chair* will appear in a tree bank in the following manner (Fig. 1):

```
[S[NP Pearl_NP1 NP]
    [VP sat_VVD
            [PP on_II
                    [NP a_AT1 chair_NN1 NP]
            PP]
    VP]
S]
```

Figure 1 *A sample tree bank with labeled brackets from English*

The morphosyntactic information is attached here with the words by underscore characters while constituents are indicated by the opening and closing square brackets annotated at the beginning and at the end with the phrase type e.g. [S ... S] (McEnery and Wilson 1996: 44).

In case of Bengali sentences, this system however, does not yield good outputs, since normal Bengali sentences do no adhere to the model of grammar used for English and other languages. Let us, for instance, consider the following normal Bengali sentences obtained from the Bengali corpus, which do not fit into the frame designed for English.

(4) a. byāpārṭā bujhte etadin samay lege gela
'So many days are spent to understand the matter'

b. lekhak se kathā ullekh karenni
,The did not mention that event'

c. eto gela anyer kathā
'This is related to other's story'

d. eṭāke kṛtrimatā bale choṭa karā uchit nay
'One should not ignore it claiming to be artificial'

e. Svādhīnatār par Gāndhījīr path theke āmrā sare esechi
'We have deviated from the path of Gandhiji after independence'

Tree banks are considered useful resources for providing annotations of natural languages at various levels of structure: word level, phrase level, sentence level and sometimes at the level of function-argument structure. They become crucially important for designing data-driven approach to natural language processing, grammar development and language education. There are a number of on-going projects for compiling representative tree banks for English, Spanish, Bulgarian, Portuguese, Dutch, and others. Also a number of project are going on for compiling tree banks for specific purposes in English, German, Russian and others.[9] The *Tübingen Treebank of Written German* is a manually parsed newspaper corpus of data taken from daily issues of *Die Tageszeitung*. The parsing scheme distinguishes four major levels of syntactic constituents: lexical, phrasal, topological and clausal. In addition to the constituent structures, parsed trees contain node labels to encode grammatical function. This tree bank currently comprises approximately 15,000 sentences. The *Tübingen Partially Parsed Corpus of Written German*, that contains collection of articles from newspapers, is annotated at clause structures, topological fields, and chunks, in addition to more low-level annotation including parts-of-speech and morphological ambiguities. All the texts are parsed at paragraph, sentence and tokens that include information about some regular types of named entities, dates, telephone numbers and unit/ number combinations.[10]

Since all parsing techniques are not identical in form, method, and representation, they usually differ in the following parameters:

(5) a. The number of constituent types, which a system employs, and

b. The ways in which constituent types are allowed to combine with each other in a sentence.

Despite such differences, most of the parsing schemes developed for English and similar other languages depend on a form of context-free phrase structure grammar. Within

this system, a ‘full parsing’ scheme aims at providing detailed analysis of a sentence structure (Fig. 2), while a ‘skeleton parsing’ aims at using less finely distinguished set of syntactic constituent types and ignores the internal structure of the certain constituent types used in a sentence (Fig. 3).

Sample Text:

Another new style feature is the wine-glass or flared heal, which was shown teamed up with pointed, squared, and chisel toes.

Full Parsing:

> [S[Ncs another_DT new_JJ style_NN feature_NN Ncs] [Vzb is_BEZ Vzb] [Ns the_AT1 [NN/JJ& wine-glass_NN [JJ+ or_CC flared_JJ JJ+]NN/JJ&] heal_NN ,_, [Fr [Nq which_WDT Nq] [Vzp was_BEDZ shown_VBN Vzp] [Tn [Vn teamed_VBN Vn] [R up_R] [P with_INN [NP [JJ/JJ/NN& pointed_JJ ,_, [JJ-squared_JJ JJ-] ,_, [NN+ and_CC chisel_NN NN+] JJ/JJ/NN&] toes_NNS Np]P]Tn]Fr]Ns] ._. S]

Figure 2 *Full parsing of LLC* (McEnery and Wilson 1996: 45)

Sample Text:

Nemo, the killer whale, who’d grown too big for his pool on Clacton Pier, has arrived safely at his new home in Windsor safari park.

Skeleton Parsing:

> [S[N Nemo_NP1 ,_, [N the_AT killer_NN1 whale_NN1 N] ,_, [Fr[N who_PNQS N][V 'd_VHD grown_VVN [J too_RG big_IJ [P for_IF [N his_APP$ pool_NN1 [P on_ll [N Clacton_NP1 Pier_NNL1 N]P]N]P]J]V]Fr]N] ,_, [V has_VHZ arrived_VVN safely_RP[P at_ll [N his_APP$ new_lJ home_NN1 [P in_ll [N Windsor_NP1 [N safari_NN1 park_NNLl]N]P]N]P]V] ._.S]

Figure 3: *Skeleton parsing of a spoken corpus* (Leech 1993)

The outputs of parsing are often post-edited by human analysts since parsing is full of differences depending on the robustness of a system, and as a result, has lower success rate than part-of-speech tagging. The disadvantage of full parsing lies in inconsistency on behalf of the analysts engaged in parsing or editing corpus. Detailed guidelines are needed to overcome these limitations. Even then, ambiguities may occur when multiple interpretations are possible for a single phrase or clause. Implementation of parsing systems for the Bengali corpus still remains a dream for us and it requires more research into the intricate form and structure of normal Bengali written sentences.[11]

Although many attempts are being made with several new ways and techniques for parsing normal written sentences, the present state of success is not very impressive. Some highly acclaimed parsers, which are often referred to as models for other languages, also cut a

sorry figure with around 50% accuracy. After nearly four decades form the first attempt for parsing, it appears far more depressing in the context when we know that, at initial stage, nearly 77% accuracy was achieved in part-of-speech tagging (Greene and Rubin 1971). The marginal success achieved in parsing leads scholars to argue that "such is the difficulty of this goal that if you are reading this book twenty years from its publication date, the authors would not be in the least surprised if no robust parser for general English has yet been created" (McEnery and Wilson 1996: 130). We will not be surprised at all if, except in some limited cases, generation of a robust parser still remains a fantasy even half a century from today.

9. Conclusion

The majority of corpus processing techniques available are designed for English, Portuguese French, German, Spanish, Swedish, Dutch, Finnish, and other language corpora. Although, these techniques are theoretically applicable to any natural language, in reality, these need to be modified to a large extent before they become useful for the Bengali corpus. Strategic modification is required due to differences existing between the Bengali language in one hand and other western languages in the other. They become useful for Bengali only when necessary modifications are made in their operational system. Even then, application of these techniques on the Bengali corpus may not yield expected results due to certain technical problems related to Bengali orthography and text samples. Therefore, the best solution is to design our own corpus processing tools, which may differ both in approach and methodology used in other languages.

The advantage of these is that these are designed keeping in view the techniques used for other languages as well as the basic nature of the Bengali language. This leads us to develop much better systems, since it infuses sophistication of western techniques with peculiarities noted in the Bengali texts. Careful consideration of relevant features of both domains produces techniques suitable for Bengali.

At present there are some widely known corpus processing tools. Among these some are language-independent while others are mostly language-specific and object-oriented with less applicability beyond the scope of the languages for which these are designed[12]. Therefore, it is realized that unless these are converted to an acceptable standard, blind application of these techniques on the Bengali corpus will yield wrong results to tarnish the actual image of the language. That means serious consideration about the potentiality of those techniques is our primary prerequisite before they are implemented on the Bengali text corpus. However, for the Bengali text corpus, there exist some tools and techniques that include frequency counting of words, morphological generation, word concordance, and word tagging (Vikas *et al*. 2003). Although most of these techniques are applied on English and other languages, these are used on Bengali as well as on other Indian language corpora for the first time with necessary modifications. Keeping this information at background, in the present paper we have focused on these tools to show how they operate on the Bengali text corpus.

Notes

[1] There are a large number of papers, which deal with corpus processing techniques of various types (Garside, Leech and Sampson 1987, Souter and Atwell 1993, Thomas and Short 1996, Garside, Leech and McEnery 1997, Oakes 1998, Biber, Conrad and Reppen 1998, Tognini-Bonelli 2001). None of these tries to use corpus of Indian languages to produce results out of them.

[2] Some concordance software available are available in the market. For example, *MonoConc* and *Conc* are used for sorting and frequency information of words, *ParaConc* is used for the purpose of parallel texts processing, *FreeText* is used for processing and sorting words. Details are available at www.ruf.rice.edu/~barlow/corpus.html

[3] Almost similar works are also reported in Bengali on some writings of *Rabindranath Tagore*, the Nobel Laureate (Mallik *et al.* 1994, Mallik *et al.* 1996, and Mallik *et al.* 2000).

[4] This is noted that among the *karmadhāraya* (descriptive), *abyayībhāba* (adverbial) and *bahubrīhi* (reciprocal) compounds in Bengali. Probably, this also true to other Indian languages which have genealogical relation with Bengali language.

[5] Illustrative empirical investigations about the phenomenon of various types of lexical collocation both in English and German text corpora are reported in some details in Barnbrook (1998).

[6] The boundary of a sentence or a paragraph is normally determined by careful examination of text, which appears to be more of a linguistic arrangement than anything else based simply on the length of context. Moreover, since conceptual interactions between words often operate beyond the boundaries of sentence and paragraph, use of these divisions may prejudge important questions of the nature of textual organisation.

[7] The list presented in table contains examples where lemmas are tagged with particles, enclitics, plural suffixes and case markers. Compound words as well as derived words, which also contribute to the process of lemma extraction, are not presented here with tags of inflection, which otherwise will make a huge list. This is a unique property of Bengali language but a rare feature for English.

[8] Some relevant information regarding different grammars used in the work of parsing is available in Souter and Atwell (1993).

[9] The practice of building syntactically parsed corpora proves that aiming at more detailed description of data becomes more and more theory-dependent. For instance, one can refer to *Prague Dependency Treebank*, *Italian Treebank*, *Turkish Treebank*, *Polish Treebank*, *Bulgarian Treebank*, etc. Therefore, development of tree banks as well as formal linguistic theories needs to be more tightly connected in order to ensure necessary information flow between them.

[10] License for accessing database is granted only for scientific purposes. For more information regarding the user license one can easily refer to http://www.sfs.uni-tuebingen.de/en_tuepp.html.

[11] A sentence boundary is determined by careful examination of a text, which appears to be more of a linguistic arrangement based on length of sentence. Since conceptual interaction between words often operates beyond the boundary of a sentence, use of traditional sentence boundary identification method appears to be quite useful in automatic sentence boundary identification and textual organization.

[12] For instance, there are downloadable corpus processing systems such as *Xtract*, which is used for lexical collocation in English; *Perl* which is used for frequency counting and sorting processing of words; *LEXA* that is used for tagging, lemmatisation and frequency count; *TextAnalyst* that is used for producing semantic network on the basis of text input; *Paai's Text Utilities* that is used in frequency count, lexical cohesion, etc. These have partial implication for Indian languages.

References

BARLOW, Michael. 1996. Corpora for theory and practice. *International Journal of Corpus Linguistics*, 1996, 1(1), pp. 1-38.

BARNBROOK, Geoff. 1998. *Language and Computers.* Edinburgh: Edinburgh University Press, 1998.

BEALE, Adam. D. 1987. Towards a distributional lexicon. In GARSIDE, R., LEECH, G., and SAMPSON, G. (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman, 1987, pp. 149-162.

BIBER, Douglas, CONRAD, Susan, and REPPEN, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

CHURCH, Kenneth, W., GALE, William A., HANKS, Patrick and HINDLE, Donald. 1991. Using statistics in lexical analysis. In ZERNIK, U. (ed.), *Lexical Acquisition.* Englewood Cliff, NJ: Erlbaum, 1991, pp. 115-164.

DASH, Niladri Sekhar. 2005. *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. New Delhi: Mittal Publications, 2005.

DASH, Niladri Sekhar. 2006. The process of lemmatisation of inflected and affixed words in Bengali text corpus. In *Proceedings of the 28th All India Conference of Linguists,* Dept. of Linguistics, Banaras Hindu University, Varanasi, 2nd– 5th Nov. 2006, pp. 127-128.

ELLIOTT, Ward and VALENZA, Robert. 1996. And then there were none: winnowing the Shakespeare claimants. *Computers and the Humanities*, 1996, 30(3), pp. 1-56.

GARSIDE, Roger, LEECH Geoffrey, and SAMPSON, Geoffrey (eds.). 1987. *The Computational Analysis of English: A Corpus Based Approach*. London: Longman, 1987.

GARSIDE, Roger, LEECH, Geoffrey and McENERY, Tony (eds.). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman, 1997.

GIBSON, Harry Norman. 1962. *The Shakespeare Claimants: A Critical Survey of the Four Principle Theories Concerning the Authorship of the Shakespearean Plays.* London: Methuen and Co., 1962.

GREENE, Barbara, B. and RUBIN, Gerald M. 1971. *Automatic Grammatical Tagging of English*. Technical Report. Department of Linguistics. Brown University, RI, USA, 1971.

JOHNS, Tim. 1991. Should you be persuaded: two samples of data-driven learning materials. In JOHNS, T. and KIND, P. (eds.), *Classroom Concordancing*. *ELR Journal 4*. University of Birmingham, 1991, pp. 1-16.

KENNEDY, Graeme. 1998. *An Introduction to Corpus Linguistics.* London: Addison-Wesley Longman, 1998.

KJELLMER, Göran. 1984. Why 'great': 'greatly', but not 'big': 'bigly'? *Studia Linguistica*, 1984, 38, pp. 1-19.

LEECH, Geoffrey. 1993. Corpus annotation schemes. *Literary and Linguistic Computing*, 1993, 8(4), pp. 275-281.

LEECH, Geoffrey and EYES, Elizabeth. 1993. Syntactic annotation: linguistic aspects of grammatical tagging and skeleton parsing. In Black, E., R. Garside, and G. Leech (Eds.) *Statistically-driven Computer Grammars of English: the IBM/Lancaster Approach* Amsterdam: Rodopi, 1993, pp. 36-61.

MALLIK, Bhaktiprasad and NARA, Tsusuki (eds.). 1994. *Gitanjali: Linguistic Statistical Analysis.* Kolkata: Indian Statistical Institute, 1994.

MALLIK, Bhaktiprasad and NARA, Tsusuki (eds.). 1996. *Sabhyatar Sankat: Linguistic Statistical Analysis.* Kolkata: Rabindra Bharati University Press, 1996.

MALLIK, Bhaktiprasad *et al.* (ed.). 2000. *Shes Lekha: Linguistic Statistical Analysis.* Kolkata: Bangla Akademi, 2000..

McENERY, Tony and WILSON, Andrew. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press, 1996.

MILLER, George A. and LEACOCK, Claudia. 2000. Lexical representations for sentence processing. In RAVIN, Y. and LEACOCK, C. (eds.), *Ploysemy: Theoretical and Computational Approaches.* New York: Oxford University Press Inc., 2000, pp. 151-160.

OAKES, Michael P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1998.

SÁNCHEZ, Jose A. and GOMEZ, Pascual Cantos. 1997. Predictability of word forms (types) and lemmas in linguistic corpora, a case study based analysis of the CUMBRE Corpus: an 8-million-word corpus of contemporary Spanish. *International Journal of Corpus Linguistic*, 1997, 2(2), pp. 259-280.

SARDINHA, Andrew P. B. 1996. Applications of WordSmith keywords. *Liverpool Working Papers in Applied Linguistics*, 1996, 2(1), pp. 81-90.

SINCLAIR, John M. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press, 1991.

SOUTER, Clive and ATWELL, Eric (eds.). 1993. *Corpus Based Computational Linguistics.* Amsterdam: Rodopi, 1993.

THOMAS, Jenny and SHORT, Mick (Eds.) 1996. *Using corpora for language research: Studies in the Honour of Geoffrey Leech.* Longman: London, 1996.

TOGNINI-BONELLI, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamin, 2001.

VIKAS, Om; CHATURVEDI, Prasant Kumar; CHOPRA, Pradip; SHARMA, Vinay Kumar; JAIN, Mangesh; and CHANDRA, Suresh (eds.). 2003. *Vishwabharat* (Indian Technology Newsletter 10). July 2003.

WILLIAMS, George C. 1998. Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3(1), pp. 151-172.

WILLS, John D. 1990. *The Lexical Syllabus.* London: Collins.

ZIPF, George Kingsley. 1949. *Human Behaviour and the Principle of Least Effort: An Introduction of Human Ecology.* Cambridge, Mass.: Addison-Wesley.

Dr. Niladri Sekhar Dash
Linguistic Research Unit
Indian Statistical Institute
203, Barrackpore Trunk Road
Kolkata - 700108, West Bengal, India
Phone (O): + 91-033-25753281 (11am - 6pm: IST)
Email: niladri@isical.ac.in
Email: ns_dash@yahoo.com
Homepage: http://www.isical.ac.in/~niladri

In *SKASE Journal of Theoretical Linguistics* [online]. 2007, vol. 4, no. 2 [cit. 2007-06-14]. Available on web page <http://www.skase.sk/Volumes/JTL09/pdf_doc/2.pdf>. ISSN 1339-782X.