**CHUNKING IN BENGALI**

**Abstract:** Chunking or the identification or extraction of chunk in linguistics is a problem to identify the group of words belonging to a particular phrase. The work of chunk identification is generally taken up in particular phrases, like only nominal or verbal but here the identification has been done as the statistics of the language suggests. If required, we are taking all of them, be it nominal or verbal. The paper is slightly limited in its scope as it does not extracts the chunks rather it defines chunk and draws boundary between sentence and chunk. The other scope it does is to show the characteristics of chunk in Bengali language. Only on chunking this is perhaps the first work of its kind in any Indian language.

**Introduction:**

Language is built with words and sentences. Words or its unit morphemes are considered to be the minimum meaningful unit of the sentence. Words convey one meaning while one sentence conveys one total expression. We see sometimes some words convey one full expression but often on parsing it is seen that the word is actually part of an elided structure on sentence. Thus from my own experience of parsing and language structure, I may say as much as I have reviewed it sentence conveys the total meaning of a human expression, where the necessity, and purpose both are conveyed. Sentence contains not only words, but also group of words. As one sentence is constructed, it is constructed for conveying some meaning, this meaning has been generated by some and is conveyed with some purpose. The word or words which convey this governing attitude of meaning in a sentence is called the subject and constitutes the subject phrase. The word or group of words which are governed by subject phrase is called object phrase. Thus we can say sentence contains phrases, which are governed by meaning and those phrases take their choice of words, to give the meaning some form of some physical entities. On the other hand words constitutes only of alphabets and phrases constitute only of words. Parsing of words and phrases are one level and of sentence is bi-level.

Thus, if we want to understand the behavior of a sentence without going into the detail grammatical formulation or intricacies of full volume parsing we can go for shallow parsing to have the chunks.

Chunking constitutes thus mainly word or words which belong to either of the subject or verb phrases, and involves one simple singular level of parsing.

#It is very useful in task of n-gram or collocational pattern analysis.

But it is not that the group of meaningful words always mean the phrases it may also mean multiword expressions or group verbs or clauses.

The scope of the paper is thus to define phrases, the entities for chunking its scopes, and find the linguistic evidence for it. Other than this our work tries to identify the difference among the phrases and multiword expressions and clauses.

This paper is divided in the following segments, the first section contains the introduction, the second section contains previous works, the third section contains the problems, fourth section contains the analysis of chunking and its linguistic evidence, followed by a suitable reference.

**PREVIOUS WORKS:**

There are various works in Bengali, in Indian languages and in European languages.

The problem of chunking was tackled by Ramshaw and Marcus (1995) for NP-Chunking, by Freitag and McCallum, (2000), Kupiec ( 1992) or multilevel Markov models (Bikel et al. 1999), (McCallum et al., 2000; Ratnaparkhi (1996) and a variety of other linear classifiers, including Punyakanok and Roth, 2001, AdaBoost (Abney et al., 1999), and Support-Vector Machines (Kudo and Matsumoto, 2001), Bandyopadyay,Ekbal and Haldar(2007), Dandapat(2008), Chatterjee, Dutta, Dhar, Barik, Sarkar and Basu(2012), Singh, Bendre, Sangal(2011), Bhaskaran(2011) , Pammi, Prahlad(2011), Ghosh, Das, Bandyopadhya(2010) made nice efforts for finding out the chunking in Bengali and Hindi through Hidden Markov Model, Maximum Entropy Markov Model and Rule. Other than these approaches, chunking is working out as a problem of Subject Verb Object (SVO) also, where Subject Verb and Object means the word or phrase acting as Subject, Verb means the word or phrase acting as Verb and Object means word or phrase acting as Object. Cole, Royal, Valtorta, Huhns, Bowles\* worked on extraction of relationship on text from SVO angle, Rusu, Dali, Fortuna, Grobelink, Mladenic\* worked on Triplet Extraction, Li, Kim, Lee\* worked on Chinese and Korean SVO.

**PROBLEM:**

The problem we are trying to discuss are not the construction of chunker, but the aim, scope and basis of chunking. The first purpose is thus the definition and divisions. There are few interrelated topics, like Multiword expression, Clause. We are trying to define against them and trying to form the basis and scope of the study. The next and conclusive study is the applicatory area of research in chunking, like it is used in sentence detection, machine translation, author identification, etc.

**ANALYSIS:**

The important definitions of chunk are**,**

1. A chunk is a unit of memory organization, formed by bringing together a set of already formed chunks in memory and welding them together into a larger unit. (Nick C. Ellis: Constructions, chunking, and connectionism: The emergence of second language structure, Revised chapter for Doughty & Long (Eds.) Handbook of Second Language Acquisition. Oxford: Blackwell. [11 April, 2001].
2. A chunk is a sequence with the property that the elements within it predict one another, but do not predict elements outside the sequence.

[Word Segmentation as General Chunking,Daniel Hewlett and Paul Cohen]

"I begin with an intuition: when I read a sentence, I read it a chunk at a time" (Abney,1994). Abney perhaps gave a very good introduction to chunking. He stressed, chunks represent one particular pattern in speech, it takes one content word and few function words around it. Abney’s point was near to the point given by Gee and Grosjean(1983) as performance structure, where group of syntactically related words around a head forms the chunk. The other property of chunk is that these words within a particular group are often non repeating in nature.

Thus, we can come to the conclusion the syntactically related words which form around a head with a particular speech pattern is called chunk and their association lies within the particular group of chunk.

Chunks can be divided into two linguistic divisions, namely phrases and clauses. Phrases are the group of words that may or may not contain verbs. Phrases thus can be noun phrases, adjective phrases, verb phrases etc and the clause contains only verb phrases. Thus the common notion of the singular level of analysis, do not attract here rather get revised. Phrase analysis or chunking is also like parsing a multi-level task and a full fledged parsing attracts now more levels of analysis.

Our next task is to differentiate between the chunks and multi word expressions. Chunk here means both phrases and clauses. Multiword Expression here means idioms as well as proverbs. Idioms are also syntactically related group of words but they do not convey the meaning the head word supposed to convey. Proverbs also have the same definition of idioms but they are formed from the speech of important persons.

# Noam Chomsky(2002) in his book “On Language”, tried to see how phrase marker represented derivational history of rewriting rules. He tried to discuss interesting examples on direct object and indirect object with sentences like, “*I promised to John that I would leave”,* and “*John was promised that I would leave”*. In unificational grammars, phrases are projections of word categories, Noun phrases are headed by nouns,Verb phrases are headed by verbs, Adjectival phrases are headed by adjectives.(<http://www.cs.haifa.ac.il/~shuly/malta-slides.pdf>).

Syntactically, we can say phrases are formed by endocentric merger operation, that is merger operation of head.(Ref: 88-89 Radford, Andrew(1997)).

There is a minor revision we need to do is between phrase and compound words, phrases are said to be descriptive function, compounds are with name-giving function, compounds display syntactical integrity, compounds seem to be more restrictive in terms of interpretability.(“ <http://www.uni-kassel.de/fb02/fileadmin/datas/fb02/Institut_f%C3%BCr_Anglistik_Amerikanistik/Dateien/Linguistik/Presentations/B%C3%B6er_Kotowski_H%C3%A4rtl_talk_IMM15.pdf>”).