Automatic Extraction of Multiword Expressions in Bengali: 
An Approach for Miserly Resource Scenarios

Aswhini Agarwal, Biswajit Ray
Department of Electrical and Electronics Engineering
National Institute of Technology Trichi
Kerala, INDIA 620015
email: {ee10208, ee10211}@niit.edu

Monojit Choudhury, Anupam Basu & Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
West Bengal, INDIA 721302
email: {monojit, anupam, sudeshna}@cse.iitkgp.ernet.in

Abstract

Works on automatic extraction of multiword expressions (MWE) have banked on the use of deep or shallow syntactic pre-processing, rich lexical resources and large corpora. However, unavailability of such tools and resources for many languages makes the problem much more difficult. In this paper, we describe an algorithm for automatic extraction of MWE in Bengali from a medium-size untagged corpus using only morphological analyzer and a root lexicon. Each word in the corpus is morphologically analyzed and provided with possible parts-of-speech tags. Depending on the language (in this case Bengali) certain MWE categories are chosen for extraction. With the help of the possible POS tags, plausible candidates are extracted from the corpus with the corresponding frequencies. Finally, the candidates are assigned a significance value based on a new statistical approach. Initial results for Bengali are promising and the approach is being tested over other languages.

Keywords: multiword expressions, collocation, morphological analysis, morpho-syntactic variation, statistical filtering, co-occurrence, Bengali
Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenarios

Abstract

Works on automatic extraction of multiword expressions (MWE) have banked on the use of deep or shallow syntactic pre-processing, rich lexical resources and large corpora. However, unavailability of such tools and resources for many languages makes the problem much more difficult. In this paper, we describe an algorithm for automatic extraction of MWE in Bengali from a medium-size untagged corpus using only morphological analyzer and a root lexicon. Each word in the corpus is morphologically analyzed and provided with possible parts-of-speech tags. Depending on the language (in this case Bengali) certain MWE categories are chosen for extraction. With the help of the possible POS tags, plausible candidates are extracted from the corpus with the corresponding frequencies. Finally, the candidates are assigned a significance value based on a new statistical approach. Initial results for Bengali are promising and the approach is being tested over other languages.

1. Introduction

One of the key issues in both natural language understanding and generation is appropriate processing of multiword expressions (MWEs). MWEs can be roughly defined as “idiosyncratic interpretation that cross word boundaries (or spaces)” [1]. A wide range of linguistic expressions used in both spoken and written form of a language are MWEs. The list includes, but is not limited to, compound nouns (e.g. traffic lights), phrasal verbs (e.g. go through), idioms (e.g. beat around the bushes), non-syntactic units (e.g. at all, of course), institutionalized phrases (e.g. multiword expression), etc. Jackendoff [2, p. 156] notes that the number of MWEs in a speaker’s lexicon is of the same order of magnitude as the number of single words. Citing the fact that 41% of the entries in the WordNet 1.7 [3] are lexicalized MWEs, Sag et al [1] claim that correct treatment of MWE in any language is even more challenging than was once thought to be.

Non-compositionality of MWEs demands exhaustive enumeration of such units or a case by case rule/grammar generation (as in the case of kin terms or numbers for English [4]). Moreover, for applications such as machine translation, a database of MWEs for a particular language is not sufficient; the database should also include the corresponding single or multiword units for the target language. Manual creation of such databases is difficult due to several reasons. First, existing resources like thesauri and dictionaries, which are used for creation of the database, may not have an adequate coverage. Second, it is tedious, time-taking and prone to errors. Third, it is not always easy to decide whether a given expression qualifies as an MWE. Measures like collocation and anti-collocation frequencies are often the deciding factors and should be computed automatically from a representative corpus.

Automatic extraction of MWEs from text corpora using linguistic and statistical tools is an alternative to manual creation of such databases. The existing techniques for automatic extraction
of MWE rely upon accurate parts-of-speech (POS) taggers, shallow parsers and rich lexical resources such as WordNets. However, most of the Indian languages including Bengali cannot boast of even a good size representative corpus, let alone such sophisticated NLP tools. The paper describes an approach for automatic extraction of certain categories of MWEs for Bengali in such a miserly resource scenario. The technique uses a morphological analyzer and a moderate size untagged text corpus. The results for Bengali are encouraging and the generic nature of the approach makes us believe that similar results can be expected for other languages as well.

The paper is organized as follows. Section 2 provides a brief overview of the existing techniques for automatic extraction of MWEs and highlights the need of sophisticated tools and extensive resources for these approaches. Section 3 proposes a taxonomy for Bengali MWE based on morpho-syntactic variation and identifies some of the classes that can be extracted in a miserly resource scenario. Section 4 describes the architecture and the algorithms used for automatic extraction of MWEs. The results are described in the following section. The concluding section summarizes the work and discusses the scope of further improvement.

In this paper, Bengali words have been written using italicized roman-script following the ITRANS convention [5].

2. Related Work

There are three approaches to automatic extraction of MWEs from a corpus – statistical, rule-based and hybrid. Statistical methods are language independent and use collocation measures like \textit{pointwise mutual score} or \textit{Student’s t test} [6, 7], but cannot guarantee good coverage because, even for very large corpora (100M words), more than two-third of the MWEs occur too infrequently to yield any statistically significant collocation [8]. Moreover, statistically significant collocations are not necessarily MWEs and might result from semantic connections (e.g. \textit{water} and \textit{drink}). To circumvent this problem the concept of un-substitutability of MWEs is often used [9]. This method, however, requires thesauri or WordNets.

Rule-based approaches [10] target some specific types of MWEs and extract them from the corpora taking clues from morpho-syntactic properties. Such methods depend heavily on POS taggers,chunkers, shallow or deep parsers and sometimes even semantic taggers. These methods are language dependent and fail for narrow domains [11]. Therefore, Dias [11] suggested a hybrid approach for MWE extraction, where patterns are identified at word level and POS-tag level using statistical measures and combined appropriately to find out valid MWEs. Sag \textit{et al} [1] also proposed a hybrid approach that uses parsed structures and \textit{dependency subgraphs}. Hybrid extractors too depend on chunkers, taggers and rich lexical resources. Although many methods for MWE extraction have good precision, the recall values are in general very poor.

In the context of Indian languages, there has been some work on MWE categorization or representation [12], but we could not find any work on automatic extraction.
3. Multiword Expressions in Bengali

Characterization and classification of MWEs is a challenging task involving a lot of debatable issues. To illustrate this point, let us take the example of the Bengali verb kATA, which normally means “to cut”. However, as shown in table 1, the meaning of kATA may vary with the context. The table shows only some of the different senses of kATA when it occurs with different nouns. One might consider kATA as a polysemous verb and assign different senses to it. Some of the senses in table 1 (e.g. to spend) are indeed due to polysemy, however it will be a gross overgeneralization if we consider constructs like nAka kATA (humiliation) as a polysemous use of the verb, because with no other noun, but mAthA (head) and nAka (nose), kATA is used in that sense. Fillmore [4] therefore concludes that there is a continuum from purely lexicalized to purely compositional expressions; thus, one might extrapolate from this that for a particular NLP system the boundary between MWEs and compositional phrases is largely a design decision.

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Meaning</th>
<th>Example Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>chula (hair)</td>
<td>to dress hair</td>
<td>nApit lokera chula kATe Barber hair dresses the people</td>
</tr>
<tr>
<td>suto (thread)</td>
<td>to spin the yarn</td>
<td>se suto keTe rojagAra kare He earns by spinning the yarn.</td>
</tr>
<tr>
<td>samaYa (time)</td>
<td>to spend/pass time</td>
<td>Ami bandhudera sAthe samaYa kATAi I spend time with friends</td>
</tr>
<tr>
<td>nambara (number) kATA</td>
<td>to deduct marks</td>
<td>apariChanna lekhAra janya nambara kATA habe Numbers will be deducted for untidy handwriting</td>
</tr>
<tr>
<td>sA.NtAra (swim) kATA</td>
<td>to swim</td>
<td>Ami sA.NtAra kATAte pAri nA I cannot swim</td>
</tr>
<tr>
<td>TikiTa (ticket) kATA</td>
<td>to buy ticket(s)</td>
<td>TikiTa nA keTe Trene uThabena nA Do not board a train without buying tickets</td>
</tr>
<tr>
<td>nAka (nose) kATA [yAoYA]</td>
<td>to be humiliated</td>
<td>Chelera kukarme bAbAra nAka kATA gela Son’s misdeed led to father’s humiliation</td>
</tr>
</tbody>
</table>

Table 1. Different senses of the Bengali verb kATA in different contexts.

Moon [12] provides a classification of MWEs for English based on lexical, semantic, syntactic, pragmatic and statistical markedness. For a scanty resource scenario, where semantic or pragmatic information are largely unavailable, Moon’s classification might not be very helpful from the point of automatic extraction. Therefore, in this paper, a different taxonomy for MWEs in Bengali is proposed based on syntactic flexibility and POS categories.

3.1 Words with Spaces

This class consists of MWEs which are syntactically rigid. No word can be inserted in between the expression; neither the words can be inflected except for the last one in some cases. Such
expressions can be considered as a single lexical entity with spaces in between. We further classify them in terms of their morpho-syntactic category as follows.

1.a. Cranberry: No inflection allowed even to the last word and some individual words may not be a part of the standard Bengali vocabulary. E.g. yena tena prakArena (by any means), yAra para nAi (ultimate), sonAYe sohAgA (an excellent combination) etc.

1.b. Named Entities: This class include names of people (kaviguru rabindranAtha ThAkura), places (pashchima ba~Nga), institutions (saMyuktA rAShTra sa~Ngha), recipes (tela kai) etc., where inflections can be added to the last word only (e.g. paschima ba~Ngera). The MWE is a noun and individual words can be either nouns, titles or adjectives.

1.c. Idiomatic Compound Nouns: These are noun-noun MWEs that are idiomatic or unproductive in nature and inflection can be added only to the last word. The formation of such compounds may be due to hidden conjunctions (mA bAbA meaning mA and bAbA, similarly bhAta jala, sakAla bikela, dharA ChoYA etc.) or hidden inflections (lalATa lekhana meaning lalATera lekhana i.e. fate, similarly ashba Dimba, AkAsha kusuma etc.).

1.d. Idiomatic Noun Groups with Inflections: These are also noun-noun compounds with idiosyncratic meaning, but the first noun is in inflected (generally possessive) form. E.g. yakShera dhana, ghughura bAsA, tAsera ghara, khu.Dora kala, kalura balada, gho.DArA Dima etc.

3.2 Semi-productive with Minor Syntactic Variations

This category includes MWEs that are either (semi-)productive in nature with its own grammar (like the numbers) or allow slight syntactic variations like inflections or a limited number of word insertion. They can be further classified as follows.

2.a. Numbers: Numerical expressions highly productive and can be expressed by a small grammar. However, no word can be inserted in between and inflections can be added only to the last word. E.g. eka hAjAra chaurAsI, tina samasta pA.Ncha bhAgera dui, sA.De chuYAttara etc.

2.b. Kin Terms: Bengali kin terms are normally two word MWEs such as mAStuto bhAi, khu.Di shAshuri etc.

2.c. Productive Compound Nouns: same as 1.c, except for the fact that the meaning is not idiomatic. These are also called institutionalized phrases. E.g. mACha bhAjA, beguna bhAjA, durgA pujo, janma tithi, bibAha bArShikI etc.

---

1 These MWEs may sometimes occur as hyphenated words; however, the use of hyphen is not strictly followed.
2.d. **Noun – Noun collocations with inflections**: same as 1.d, except for the fact that these are semi-productive. E.g. mATira mAnuSha, sonAra Chele / meYe, hIrera / chA.NderA Tuka.Do, naYanera / Chokhera maNi / tARa etc.

2.e. **Conjunct Verbs**: These are a pair of similar verbs used together to denote some other action. When used in inflected form, the same inflection (normally e or te) is added to both the verbs. E.g. khAoYA dAoYA, hAsA khelA, pa.DA shonA etc.

### 3.3 High Syntactic Variation but Fixed POS Categories

This class includes noun-verb, adverb-verb and adjective-verb collocations, where the syntactic structure is quite flexible. For example, the ordering and the inflections of the words can vary, and the words can be separated by arbitrarily large number of words. However, the POS category of the words involved in such collocations is restricted. This class has both unproductive and semi-productive sub-categories as described below.

3.a. **Do/Is Support Verbs**: This is a productive class where verbs are formed by addition of “do” (karA) or “is” (haoYA) to a noun. E.g. pujo karA, snAna karA, sparSha karA, sammukhIna haoYA, AlApa haoYA, bRRiShTi haoYA etc.

3.b. **Light Verb Constructions**: Some verbs like deoYA (to give) or kATA can have different senses in different context. They are often referred to as light verbs [8]. Some of these variations can be considered to be polyseme, however if the context is not sufficiently productive, we can consider them as MWEs. E.g. manoyoga / mana deoYA, preme / phA.Nde / jAle pa.DA, kila / cha.Da / lAthi khAoYA etc.

3.c. **Idiomatic Noun-Verb Collocation**: This class includes unproductive and idiosyncratic noun-verb collocations such as chokha oThA, si.Di bhA~NgA, pakeTa mARa, ghara bAdhA, mAthA dhARa etc.

3.d. **Adjective-Verb and Adverb-Verb Collocations**: Might be idiomatic or compositional, but statistically marked. E.g. (lajjAYe) lAla haoYA, musaladhAre / TipaTipa bRRiShTi haoYA etc.

### 3.4 Completely Flexible MWEs

This category includes idioms and proverbs for which neither the word ordering, nor the POS category of the expression is fixed. High degree of syntactic variation and even synonym substitution is allowed. E.g. Thaga bAChate gA.N ujAra, ulu bane mukto Cha.DAno, gharera kheyA banera moSha tA.DAno etc.

MWEs of category 1 (§3.1) can be manually enumerated, except for the class 1.b. MWEs of category 4 (§3.4) are difficult to extract automatically and can also be enumerated manually. This work focuses on extraction of syntactically semi-flexible MWEs that belong to category 2 and 3. Since, numbers and kin-terms can be captured by a mini-grammar, such MWEs are not
considered during extraction. Thus, the problem boils down to identification of noun-noun, adjective-noun, noun-verb, verb-verb, adjective-verb and adverb-verb collocations.

4. Automatic Extraction of MWEs

Figure 1 gives an overview of the extraction procedure. First some basic preprocessing is done over the corpus and every word is assigned a set of possible POS tags. For extraction of noun-verb, adjective-verb and adverb-verb collocations, each sentence is further analyzed for identification of the main verb(s). Possible MWE candidates are extracted from the sentences and assigned a significance value based on statistical parameters like co-occurrence and individual frequencies. The lists of different classes of MWEs are finally sorted in the descending order of significance value.

---

**Figure 1.** Overview of the Extraction Procedure
4.1 Resources Used

The monolingual Bengali corpus developed by Central Institute of Indian Languages, Mysore was used for extraction of MWEs. The untagged corpus consists of around 1200 documents from different domains. There are a little more than 3 million words of which around 0.2 million words are distinct. Therefore, this is only a moderate size corpus and cannot be expected to have a good coverage of MWEs.

The morphological analyzer (MA) used in this work has been developed in-house and uses a root lexicon. It can analyze only inflected forms of verbs, nouns and pronouns, but not derivational morphology. If a word has multiple morphological parses, the MA gives all of them as output with corresponding POS tags, roots and inflections. For example, for the word *kare*, the MA gives the following results.

1. [FINITE VERB] root: *kar*, inflection: *e*, attributes: 3\textsuperscript{rd} person, simple present (does)
2. [NON-FINITE VERB] root: *kar*, inflection: *e* (having done)
4. [POST POSITION] root: *kare* (by/on)

4.2 Basic Preprocessing

The aim of the preprocessing is to identify sentence boundaries and possible POS tags for each word so that we can select possible candidates for further analysis based on POS categories. The sentence boundaries are determined on the basis of punctuation marks by using simple heuristics as given below.

- There is always a sentence boundary at ‘.’ (period) unless the period is a part of an acronym or number.
- Acronyms are identified by the presence of periods after non-sense (out of vocabulary) words. Periods within numbers are immediately followed by numerals.
- The marks of exclamation ‘!’; interrogation ‘?’; semi-colon ‘;’ always mark a sentence boundary.
- A new sentence begins from double quotes (“”) and continues till the end of the quote (“”), if there are no sentence boundaries in between the quotes as per the above three rules.
- Words within parentheses, braces and brackets are not considered as a part of the main sentence and treated as separate sentences.

The heuristics stated above may not mark the true sentence boundaries, but our basic assumption is that MWEs do not span over such punctuation marks and therefore, the search for possible co-occurrences can be restricted to the sentences or fragments of sentences identified using the aforementioned heuristics. Due to typographic errors and informal use of punctuation marks in Bengali, sometimes sentence boundaries were not properly identified and very large (more than 20 words) sentences were encountered. Such cases were reported as errors by the system and later corrected manually.
After identification of the sentence boundaries, each word is sent to the MA and all possible POS tags with corresponding roots are stored along with the word. The words that are not recognized by the MA, which may be due to typographic errors in the corpus or absence of the root in the lexicon used by the MA, are marked as unknown. They can be valid proper nouns as well. After the pre-processing phase, we have a set of sentences (or fragments of sentences), where each word of a sentence is followed by the possible POS tags and the corresponding roots.

### 4.3 Extraction of Possible Noun-Noun, Adjective-Noun and Verb-Verb Candidates

The noun-noun and verb-verb collocations, which fall in category 2 (§3.2), are semi-productive in nature and do not allow much syntactic variation. They are adjacent to each other and the second word can be inflected. For noun-noun category the first noun can have a possessive marker, *er* (2.d) and for verb-verb category, same inflection is added to both the verbs (2.e). Similar conditions hold for adjective-noun collocations. Therefore, each of the sentences in the corpus are scanned word by word and checked for the following conditions.

- If a word $w$ has NOUN as one of the possible POS tags and the word immediately preceding $w$ (say $v$) also has NOUN as a possible tag and has either the possessive inflection (*er*) or is uninflected, then the pair $(v, w)$ is a possible noun-noun MWE.
- If $w$ has NOUN as one of the possible POS tags and $v$ has ADJECTIVE as one of the possible tags, then the pair $(v, w)$ is a possible adjective-noun MWE.
- If $w$ has VERB as one of the possible POS tags and $v$ also has VERB as one of the possible tags and if both of them have the same inflection, but different roots, then the pair $(v, w)$ is a possible verb-verb MWE.
- If both $v$ and $w$ have VERBAL NOUN as a possible tag and either both has the same inflection or $v$ has null inflection, then $(v, w)$ is a possible verb-verb MWE.

### 4.4 Extraction of possible Noun-Verb, Adjective-Verb and Adverb-Verb Candidates

Noun-verb, adjective-verb and adverb-verb MWEs belong to category 3 (§3.3). They are syntactically flexible and need not be adjacent to each other. In noun-verb MWEs, both the noun and the verb can be inflected. Moreover, only the collocation of the main verb with the nouns, adverbs or the adjectives is meaningful, but not the auxiliary, modal or vector verbs. For example, in a sentence like “*chA nA khele mAthATA dhare yAYa*” (I get a headache, if I do not consume tea), there are three verbs *khele* (root: *khAoYA*), *dhare* (root: *dharA*) and *yAYa* (root: *yAoYA*). *mAthA dharA* is a valid MWE and *chA khAoYA* is a valid semantic collocation, but *mAthA yAoYA*, *mAthA khAoYA*, *chA yAoYA* and *chA dharA* are not\(^2\). The facts that *yAYa* is acting as a vector verb here and *khele* marks a clausal boundary in the sentence are important for proper computation of the collocation frequencies.

\(^2\) Although *mAthA yAoYA* (to become crazy), *mAthA khAoYA* (to irritate/vex) and *chA dharA* (get addicted to tea) are valid MWEs or collocations in Bengali, the example sentences cited here is not an evidence for any of these.
In order to identify the main verb(s) in a sentence the following heuristic was used. Let the word \( w \) be a possible VERB. \( w \) is considered to be one of the main verbs in the sentence if (1.1) \( w \) is a non-finite verb, (1.2) it is eventually followed by a finite verb \( u \) which is also a possible auxiliary, vector or modal verb and (1.3) all the words between \( w \) and \( u \) are possible vector, modal or auxiliary verbs in non-finite forms; or (2.1) \( w \) is a finite verb and (2.2) there is no non-finite verb \( v \) that precedes \( w \). The conditions do not guarantee identification of all the main verbs, but it works for majority of the cases.

The sentences are analyzed for main verbs and put in different sets according to the main verbs. A sentence can belong to more than one set if it has more than one main verb. For each verb, the possible list of nouns, adverbs and verbs are extracted by considering the set of possible POS tags of the words from the list of sentences. Although due to syntactic flexibility of this class of MWEs, occurrence of a noun-verb in the same sentence is a sufficient condition for considering it as a valid co-occurrence, it was found that normally for valid MWEs the nouns are adjacent or very close to the verb. The same goes true for adjectives and adverbs. Therefore, in order to reduce the effect of spurious co-occurrences, the co-occurrence frequency for this class of MWEs were defined as

\[
co(u, v) = \sum_{s \in S(u, v)} e^{-d(s, u, v)} \tag{1}
\]

where,
- \( co(u, v) \) = co-occurrence frequency between the words (after stemming) \( u \) and \( v \)
- \( S(u, v) \) = set of all sentences where both \( u \) and \( v \) occurs
- \( d(s, u, v) \) = Number of words between \( u \) and \( v \) in the sentence \( s \)

Therefore, for every adjacent occurrence of \( u \) and \( v \), \( co(u, v) \) increases by 1, but if in a sentence \( u \) and \( v \) are largely separated, \( co(u, v) \) increases only marginally.

4.5 Filtering and Sorting

The following parameters are extracted for every word pair \((u, v)\).
- \( co(u, v) \) = collocation frequency between \( u \) and \( v \) calculated by counting the number of occurrence for adjacent MWE (§4.4) and calculated using equation 1 for other MWEs.
- \( f(u, v) \) = number of sentences in which both \( u \) and \( v \) are present and the specific conditions stated in §4.3 or §4.4 hold.
- \( f(u) \) = frequency of the word \( u \) in the corpus
- \( f_{v}(u) \) = number of verbs with which the word \( u \) has occurred
- \( f_{n}(u) \) = number of nouns with which the word \( u \) has occurred

Some general parameters like \( T_{v} (= 583) \) and \( T_{n} \), the total number of verbs and nouns, are also extracted from the corpus. For MWEs of class 2 (§4.3), \( co(u, v) = f(u, v) \).
All the \((u, v)\) pairs with \(f(u, v) < 3\) were filtered out, because it is not possible to make any statistically significant claim about them. For the remaining candidates, a significance value for each pair is computed from the statistical parameters extracted from the corpus. The idea is to determine the probability that a given pair \((u, v)\) is a valid MWE. Previous works have focused on different language models and often taken the help of anti-collocation frequency to determine this probability [6,7,8]. Since it is not possible to obtain the anti-collocation frequencies without a WordNet, we take a different approach here to assign significance values to the word pairs.

On the basis of empirical observations, we propose the following hypotheses regarding the significance value (as an MWE) of a noun verb pair \((n,v)\). The first two hypotheses are valid for other MWE categories as well.

H1. The expected co-occurrence frequency of \(n\) with any verb with which it occurs at least once, is \(f(n)/\hat{f}_v(n)\). If \(co(n,v)\) greater than \(f(n)/\hat{f}_v(n)\), the probability that \((n,v)\) is a valid MWE increases sharply, whereas the probability falls sharply as \(co(n,v)\) decreases from the mean value.

H2. Nouns which do not collocate with a large number of verbs have better chance of forming valid MWEs. For example, the noun ghara (house/home) can be the location of any verb and hence co-occurs with a large number of verbs, but most of these are not valid collocations. Therefore, if \(f_v(n)\) is greater than a value \(\lambda\), the probability that \((n,v)\) is a valid MWE reduces drastically. \(\lambda\) can be considered to be the average of \(f_v(n)\) over all the nouns in the corpus.

H3. Unlike nouns, the verbs which co-occur with a large number of nouns (high \(f_n(v)\)) and are very frequent (high \(f(v)\)) have greater chance of forming MWEs.

Therefore, we arrived at a definition of significance of a noun–verb collocation as follows. The significance values for other categories can be defined similarly.

\[
\text{sig}_v(n) = \sigma[\kappa_1 \cdot (1 - co(n,v) \cdot \hat{f}_v(n)/f(n))] \cdot \sigma[\kappa_2 \cdot (\hat{f}_v(n)/\lambda - 1)]
\]

\[
\text{sig}(v,n) = \text{sig}_v(n) \cdot \exp\left[\frac{f_n(v)}{\max(f_n(v))} - 1\right]
\]

where,

\(\text{sig}_v(n)\) – significance of noun \(n\) with respect to the verb \(v\).

\(\text{sig}(n, v)\) – general significance of \((n,v)\).

\(\sigma(x) = e^x/(1+e^x)\), is the sigmoid function.

\(\kappa_1, \kappa_2\) – constants which determine the stiffness of the sigmoid curve

\(\lambda = \Sigma(f_i(n))/T_n\) is the average number of verbs a noun co-occurs with.

The values of \(\text{sig}_v(n)\) and \(\text{sig}(n)\) lie between 0 and 1. \(\lambda\) was found to be around 190 and based on empirical observations, \(\kappa_1\) and \(\kappa_2\) were assigned a value of 5.0.
5. Observations and Discussions

The extracted candidates were sorted on the basis of their significance and each list was evaluated manually. The candidates in the list were classified into four categories – (1) valid MWEs, (2) valid semantic collocations, (3) invalid MWE due to incorrect POS assumption and (4) invalid MWE due to other reasons. Table 2 illustrates these categories by reproducing some of the results (in order) for the verb kATA. Considering category (1) and (2) as correct and (3) and (4) as incorrect outputs, the precision was calculated for the first 500 entries.

Figure 2 shows how precision calculated using three different criteria changes with the rank. The monotonicity of the curves and very high precision for the high ranked candidates show that the significance function described in this paper (equations (2) and (3)) are quite accurate and have very good discriminative power. Also, out of 44730 noun-verb candidates, there is not a single valid MWE in the last 3500. Table 3 lists some valid collocations with ranks after 40000. There were only 7 such cases.

Although, the approach described here has a good precision, the recall is low. It is difficult to quantify the recall as we do not know the total number of MWEs in the corpus, however, initial investigation shows that many frequently used noun-verb MWEs are absent in the list of 44730 possible candidates, because they occurred less than thrice in the corpus.

<table>
<thead>
<tr>
<th>Noun</th>
<th>$f(n)$</th>
<th>$co(n, kATA)$</th>
<th>$f_s(n)$</th>
<th>$sig(n, kATA)$</th>
<th>Rank (overall/verb specific)</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>sA.NtAra</td>
<td>92</td>
<td>7.34</td>
<td>50</td>
<td>0.793</td>
<td>443/2</td>
<td>Idiomatic (1)</td>
</tr>
<tr>
<td>kATA</td>
<td>100</td>
<td>6.14</td>
<td>44</td>
<td>0.789</td>
<td>456/3</td>
<td>POS error (3)</td>
</tr>
<tr>
<td>suto</td>
<td>123</td>
<td>3.92</td>
<td>65</td>
<td>0.743</td>
<td>632/4</td>
<td>Idiomatic (1)</td>
</tr>
<tr>
<td>kATA</td>
<td>459</td>
<td>41.81</td>
<td>126</td>
<td>0.686</td>
<td>907/5</td>
<td>POS error (3)</td>
</tr>
<tr>
<td>dAga</td>
<td>265</td>
<td>5.467</td>
<td>85</td>
<td>0.670</td>
<td>978/7</td>
<td>Idiomatic (1)</td>
</tr>
<tr>
<td>Churi</td>
<td>76</td>
<td>2.135</td>
<td>50</td>
<td>0.566</td>
<td>1528/8</td>
<td>Semantic (2)</td>
</tr>
<tr>
<td>pAtA</td>
<td>871</td>
<td>3.050</td>
<td>166</td>
<td>0.092</td>
<td>9922/23</td>
<td>Invalid (4)</td>
</tr>
</tbody>
</table>

Table 2. Different Nouns retrieved for the verb kATA. $f(kATA) = 739$, $f_s(kATA) = 1261$

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>$co(n,v)$</th>
<th>$sig(n, v)$</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>mana</td>
<td>TAnA</td>
<td>1.560</td>
<td>0.000865</td>
<td>40138</td>
</tr>
<tr>
<td>mana</td>
<td>yog.Ano</td>
<td>0.405</td>
<td>0.000693</td>
<td>40576</td>
</tr>
<tr>
<td>kathA</td>
<td>kaYA</td>
<td>1.100</td>
<td>0.000532</td>
<td>41157</td>
</tr>
<tr>
<td>samaYa</td>
<td>mApA</td>
<td>0.922</td>
<td>0.000420</td>
<td>41566</td>
</tr>
</tbody>
</table>

Table 3. Few valid collocations ranking above 40000
6. Conclusion

In this paper, we have described an approach for extraction of MWEs of specific kinds from a moderate size untagged corpus using morphological analysis and statistical methods. The precision achieved for Bengali MWEs are comparable to other existing techniques that rely upon much more sophisticated linguistic tools and resources. Apart from being the first work of its kind for any Indian language, the contributions of this work are

- A hybrid approach for extraction of MWEs in a miserly resource scenario.
- A new ranking method using *sigmoid*, which is flexible and can be tuned for different languages/contexts by adjusting the sharpness parameters and means.
- A method to handle sparse linguistic data
- A morpho-syntactic classification of Bengali MWEs
- A list of possible Bengali MWEs, which can be corrected manually and used for building good lexical resources
Further improvements of automatic extraction techniques under scanty resource scenario can focus on local disambiguation of POS tags from simple heuristics or use of accurate POS-taggers, supervised or unsupervised learning of the sharpness parameters ($\kappa_1$, $\kappa_2$) for better discrimination, methods to handle extremely sparse MWEs etc. Work is underway on extraction of idioms and more flexible MWEs (category 4) under similar conditions and application of the technique on other Indian languages.

References