A New Approach for Hindi to English Idiom Translation

Himani Mishra (corresponding author)
PG Scholar, Department of CSE,
Shri Vaishnav Institute of Technology and Science,
Indore, India
himanimishra.hm21@gmail.com,

Rajesh Kumar Chakrawarti,
Reader, Department of CSE,
Shri Vaishnav Institute of Technology and Science,
Indore, India
rajesh_kr_chakra@yahoo.com,

Dr. Pratosh Bansal
Professor, Department of IT, Institute of Engineering and Technology,
Devi Ahilya Vishwavidyalaya,
Indore, India
pratosh@hotmail.com,

***Abstract*—Idioms are the significant part of any language and are widely used during communication [1] as they enhance the elegance of that conversation and language as well. Idioms carry a figurative meaning [2] which makes the translation of idioms rather difficult than any text translation. They are the real challenge for machine translation from the preliminary stage of machine translation development. A lot of research has been done for extraction and translation of text in many languages, but no significant research has been captured in Hindi to English idiom translation. In this paper, we have designed a new hybrid machine translation architecture using sub-categories - transfer-based and interlingual-based machine translation of rule-based approach to automate Hindi to English Idiom Translation. This Hindi to English idiom translation system can be expanded for other language pairs and can be embedded with other machine translation system to improve their translation by encapsulating correct idiom translation with their ordinary translation.**

***Keywords*—Idioms, translation, machine translation, Hindi, English, language.**

I. INTRODUCTION

The collection of words which offers a different meaning when taken together as compared to when considered individually or when only its literal meaning is considered is defined as idioms [1][2][3]. 'Idiom', taken from Greek-Latin word 'Idioma' which means special feature or property or special phrasing [3]. In most of the idioms, it is useless to find literal meaning, since they contain cranberry words [2] which lack literal meaning. They are valuable and meaningful only within that idiom. For instance- consider the idiom 'Spill the beans' whose literal meaning is -to spill out the beans; but in the true sense, it means to leak out secret information. This twist in extracting the meaning of idiom and then translating that into idiom of another language, makes idiom machine translation difficult as machine translation methods generally do not apply algorithms for translating idioms with their figurative meaning [2], rather it simply maps source text word to target text words using lexical dictionary-direct approach or by applying some analogy which translates using some pre-translated source-target example sentence pairs-example based approach, by using some predefined rules for translation-rule-based approach and many more approaches of this type.

In this paper, first, we have given a brief note on the research work done in the field of Idiom translation. Following that, we have discussed a new methodology for Hindi to English Idiom Translation to ease the machine translation of Idioms using some classic methods in a completely newer approach or newer architecture for better result and accuracy than other approaches.

II. LITERATURE SURVEY

Present day is an era of digitalization and zillions of data is available, either on our hard drives or in our pen drives or in our mobile phones or on clouds, in different languages. There was a need for a better approach for

translation as compared to manual and its evaluation as the translation by humans was time-consuming, expensive, subjective and highly prejudiced task [4].

A. Approaches Available for Idiom Translation

Development in the field of machine translation solved many of these issues using various approaches. Machine translation can be bilingual or multilingual.The approaches to machine translation are:-

- *Corpus-Based Machine Translation*: -Corpus-based approach relies on the study of bilingual text corpora [5] [6]. Corpus-Based machine translation is further categorized as Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT). In SMT, the translation is generated on the basis of Statistical model whose parameters are taken from text corpora [7]; whereas EBMT approach is translation by analogy in which we provide EBMT system with some input sentences (source language) together with their output translation sentences (in target language), then system utilizes these instance pairs to translate other similar source language sentences to target language sentences [8]. After a lot of discussion from 1986 to 2006, now idiom translation can be related to EBMT approach [2].
- *Direct Machine Translation*: - A simple translation approach where individual words of input source sentence are translated into corresponding words of target output sentence followed by some syntactical rearrangements in order to keep the structure of sentence of that language alive, is called direct machine translation [5][7][9]. It is probably the simplest machine translation approach present and is used till date for automatic translation of any text or speech.
- *Rule-Based Machine Translation*: -As the name itself suggests, in rule-based approach, the source language sentence is parsed by the system and an intermediate representation (IR) is produced. It can be some sort of parse tree or abstract representation of source sentence. Using this intermediate representation, target sentence is generated with taking into account semantics, morphological and syntactic information [5] [9]. It is subcategorized into Transfer-Based Machine Translation (TBMT) and Interlingual-Based Machine Translation. The basic steps of both the methods are same as that of its parent approach. Transfer-based MT consists of three modules namely- Analysis module, Transfer module and Generation module [9] [10], whereas Interlingual-Based machine translation comprises of two modules- Analysis module and Synthesis module [9].
- *Hybrid Machine Translation*: - The combination of two or more machine translation methods is defined as Hybrid machine translation. The idea is to utilize the strength of both the approaches while dropping out the weakness. Some organizations have claimed to have implemented such hybrid systems. Omniscien Technologies, LinguaSys , Systran are among such companies which uses a hybrid method which is the combination of rule-based approach and statistic approach.

III. NOTEWORTHY CONTRIBUTIONS

In 1952, Bar-Hillel in his presentation on “The treatment of ‘Idioms’ by a Translating Machine” on Mechanical Translation at MIT said- “The only way for a machine to treat idioms is -not to have idioms!” [2]. This sentence of Bar Hillel shed light on the extent of difficulties during Idiom Translation by machines together with the lack of solutions except getting rid of them. Hutchins in 1995 wrote an article about critiques about MT in which he quoted examples to show problems of ambiguity and lexical issues which mostly included idioms.

During 1990, Santos discussed a specific treatment of lexical gaps and idioms in MT system “PORTUGA”. In 1998, Wehrli described fixed word expressions in "ITS-2", a French-English MTS; FROMTo K/E which contains Korean dependency parser with an idiomatic expression recognizer was developed by Ryu et al. in 1999.

Many researchers worked for the machine translation of multiword expressions, phrases, metaphors and idioms applying the various combination of approaches available together with some grammar rule or sometimes with an improvised parser or sometimes with finite state automata or by separately storing these expressions. Following are some of the Noteworthy contributions of our researchers in machine translation of idioms from 1986 to 2007.

TABLE 1. RESEARCH IN THE FIELD OF IDIOM TRANSLATION

S.No.	Year	Work
1.	1986	Schenk came up with idiom translation using "isomorphic grammar" approach that idiom translation system was called as -Rosetta.
2.	1990	Santos defined a parser that studies and produces the MWE due to lexical transfer - PORTUGA.
3.	1998	Wehrli introduced this system in which the idiom is initially parsed, retrieved if all lexical constraints related to that idiom is satisfied- ITS-2.
4.	1999	Ryu et al. Described a dependency parser which contains an idiomatic expression recognizer- FromTo K/E.
5.	2000	Tools for collocations, identification and representations were introduced by Krenn, Franz et al. Implemented a system called HARMONY. Here they explained a design using example-based machine translation approach.
6.	2001	Bi-directional Finite State Automata (BFSA) for parsing the idioms was the idea of Poibeau
7.	2002	Fellbaum provided details like status and representation related to the type of VP idioms.
8.	2004	Lexical resource focusing on German verb phrase idioms using corpus-based approach was given by Neumann et al.
9.	2005	A collection of German idioms were implemented using Sailer's approach in TRALE by Bela Usabaev, Automatic Idiom Extraction by applying graph analysis and asymmetric lexico-syntactic patterns was given by Widdows and Dorow.
10.	2006	The technique for Auto lexicon construction of idioms was given by Fazly and Stevenson. After this, idiom translation was somehow related to EMBT approach. Sohn designed an HPSG analysis based on Sailer . This was for German verb phase Idioms.
11.	2007	An idea for storing idioms separately in a file was given by Gangadharairah and Balakrishnan. The format for storing was given as *SL lemma-> TL lemma.*

IV. PROPOSED SYSTEM ARCHITECTURE

To understand the architecture of the idiom translation system, we must understand the elementary concepts on which the system is designed. Below is the basis of Idiom categorization according to various researchers, using which the idiom translation system will be implemented. The second part of this section covers the design modules and component description of those modules together with the reason for their design and working.

A. Idioms in Hindi-English Language Pair

Many terms like idioms, collocations, (dead) metaphors, periphrastic phrases, proverbs etc describe expression with figurative meaning [2] i.e. the meaning of such Multiword Expression (MWE) are different from the one that is coming out from that MWE. Basically, there are two problems with idiom machine translation. Those are:-

1. How to understand what actually the idiom is trying to convey. For instance- "Too many cooks spoil the broth [12]" which actually sounds that the MWE is trying to express that a soup (broth) is getting spoiled if many cooks are involved. But actually, it means- "if too many people are participating in a task, it will be screwed up".
2. After understanding that, second task or issue is how to translate that idiom into the same meaning idiom in another language i.e. how to convert the Hindi idiom into equivalent English idiom. By equivalent here, we mean the one having the same meaning in both literatures. Again taking the example of the first problem "Too many cooks spoil the broth" [11][12], in Hindi the equivalent idiom is "बहूत से जोगी मठ उजाड़ । "[11]. It is quite difficult to translate these two idioms without using any intelligent trick in between or without exploiting connecting edge in these two different language's idioms.

Nunberg et al. in 1994, considered idioms can be applied to a fuzzy category defined on the one hand by ostention of prototypical examples (…) [2]. Idioms are classified into many sub-categories by a wide range of scholars. Some of those classifications are listed below:-

- Encoding vs Decoding idiom by Makkai (1972), Fillmore et al. (1988).
- Grammatical vs Exragrammatical,
- Lexical Filled vs Lexically open Idiom by Fillmore(1988),
- Idiom with vs without Pragmatic point [2].

We can also categorize idioms into another sub-divisions which will be utilitarian for implementing the Idiom Translation System. On the basis of 'ways idioms can be translated'; we can classify them into three sub-categories [3]. The sub-classification are as under:

- CASE I- Idioms which have similar meaning and similar form in both languages [3][11][12]:
  For e.g.
  All's well that ends well---अंत भला , तो सब भला
  Contentment is happiness--- संतोषी परम सुखम ।
  Unity is strength--- एकता ही बल है ।
- CASE II- Idioms which have similar meaning and dissimilar form [3][11][12]:
  For e.g.
  To add fuel to fire---आग में घी डालना ।
  To sleep like log---घोड़े बेच कर सोना ।
  No pain, no gain--- बिना सेवा मेवा नहीं मिलता ।
  One nail drives out another--- कांटे से काँटा निकलता है ।
- CASE III- Idioms which have completely different meaning and different form [3][11][12]:
  For e.g.
  A nine day's wonder---चार दिन की चाँदनी , फिर अंधेरी रात ।
  An empty vessel sounds much--- थोथा चना बाजे घना ।
  A drop in the ocean---ऊठ के मुँह में जीरा ।

B. System Architecture

We are implementing Hindi to English Idiom Translation System using a Rule-based approach in which we will apply both Transfer-based method as well as Interlingual-based method. The System architecture is divided into two phases, in accordance with the type of input idiom (from the last classification). Those phases are Phase I- Comparison phase and Phase II- Translation phase.

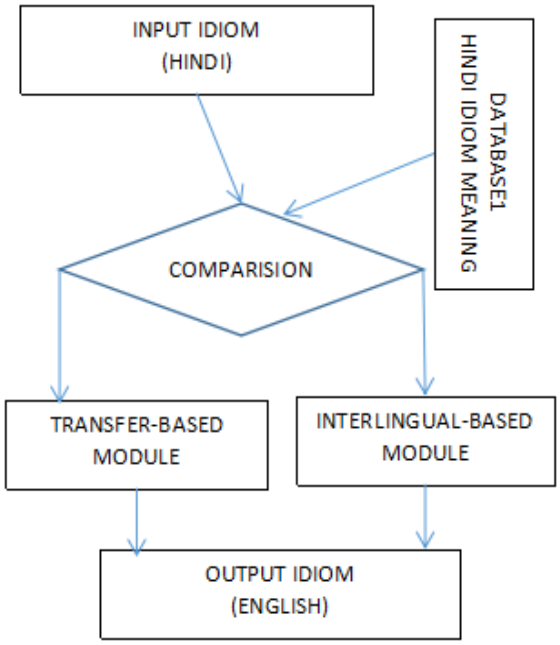

Fig 1. Block Diagram of System Architecture [1] [5] [10]

1. *Phase I- Comparison Phase:* - When the user provides input(Hindi idiom) to the idiom translation system, the Phase I gets activated. Comparison between the user given (Hindi idiom) and available list of idioms from Database1(containing Hindi Idioms together with its meaning) is done, based on which the comparison algorithm decides to which translation module the idiom should be forwarded for Phase II. If the user given idiom is present in database1 that means it must be sent to Interlingual-Based module(i.e. that idiom is of dissimilar meaning and different form, third case) and if the user given idiom is not available to the database1(that means it is either of similar meaning and similar form, first case or similar meaning and different form, second case), it should be forwarded to Transfer-Based module for machine translation.
2. *Phase II- Translation Phase:* - As soon as it is decided that to which translation module the given Hindi idiom is to be forwarded for translation, the second Phase comes into the frame. In translation phase, the translation module (either transfer-based or interlingual-based) initiates the translation according to the algorithm steps presented in that module. Finally, the translation of the given input Hindi Idiom arrives as equivalent English Idiom.

1. Transfer- Based Module:

In a transfer-based module, Hindi to English Idiom translation is performed using transfer rules.This module is basically for idioms with similar meaning and either similar or dissimilar form in both language pairs i.e. this translation module is for case- I and case- II. As we are more concerned about the improvised or advanced approach for idiom machine translation which is a combination of interlingual-based approach and improvised transfer-based approach, we will discuss transfer-based approach in brief. It consists of following components:

- *Input:* - The input Hindi Idiom is presented to the translation system from comparison module after deciding to which module the user idiom input should be forwarded.
- *Tokenizer*: - Tokenizer [4] or lexical analyzer or word segmenter [10] or splitter [5]; splits or segments the input idiom into units known as tokens. This token set generated here is passed onto next component is known as the parser.
- *Parser and POS tagger*: - Parser studies the syntax and semantic structure of Hindi Idiom. A parser contains POS tag-set which can be used for Part Of Speech (POS) Tagging. The output of this part is processed in Declension tagger [4] [5] [10].
- *Declension Tagger*: - In declension tagger phase, the tagged output of the parser is again re-tagged on the basis of some declension rules [5].
- *Reordering*: -Hindi has a structure of Subject Object Verb (SOV) but English sentence follows Subject Verb Object (SVO) structure. To translate idioms from Hindi to English, we must reorder idioms according to the structure of English language [5] i.e. Subject Verb Object.

- *Translator*: - Translator, as the name itself suggests, translates Hindi words to correct English words from any Lexical dictionary based on translation rule [5] [10].
- *Output:* - Output is the English idiom equivalent to the given input Hindi idiom.

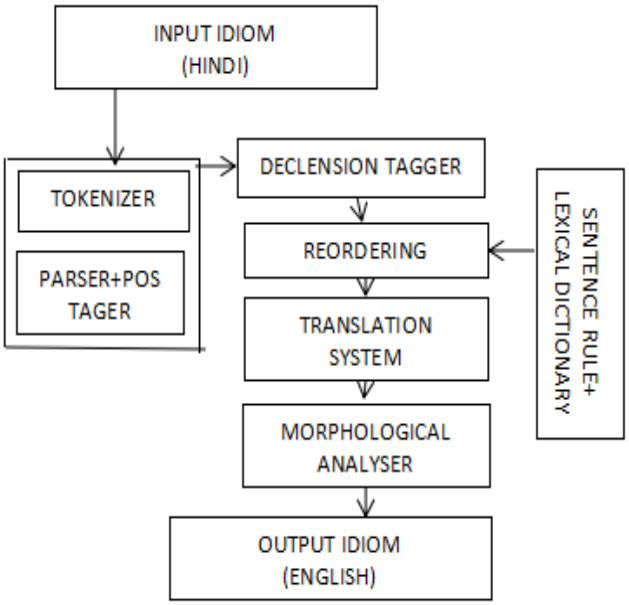

Fig 2. Block Diagram of Transfer-Based Module[4] [5] [10]

The problem encountered with this diagram of the transfer-based module is that it will work only for idioms with similar meaning and similar form (case- I) like--- " हवाई किले बनाना । "[11] which is equivalent to the English idiom " to make castles in air "[12]. Here the idiom translation is any ordinary translation. We do not require any sort of extra effort to translate as its Hindi Idiom = English Idiom. The above idiom's translation can be done using the classic transfer-based module. But this transfer-based module cannot work correctly with idioms with similar meaning and dissimilar form (case- II) i.e. " जले पर नमक छिड़कना । "[11] which is equivalent to " To rub salt in wounds "[3][12]. Here, the translation is same except that if "जले" which is "burn" in English can be replaced with "wound". Now, if the dictionary is little bit improvised and if it replaces certain words with their idiomatic meaning also, then the above transfer-based module can be used for case- II also. So below we present an improvised Transfer-based module design as a block diagram which contains an improvised idiomatic dictionary which will be utilitarian for the second case. So this advanced module design will be applicable for both the cases i.e. idioms with similar meaning and similar form case I and idioms with similar meaning and dissimilar form case- II.

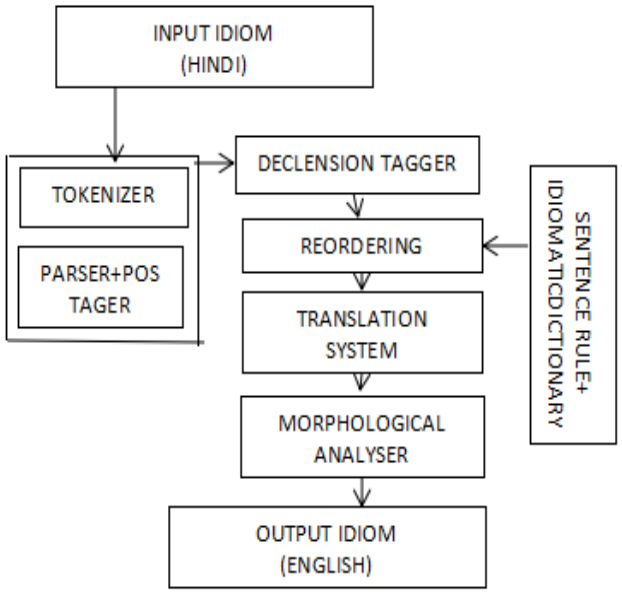

Fig 3. Block Diagram of Improvised Transfer-Based Module[4] [5] [10]

2. Interlingual-Based Module

If the idiom matches one of the idioms present in the Hindi Idiom Database which is Database1, which indicates that this idiom belongs to the case- III (dissimilar meaning and dissimilar form). This denotes that it needs to be handled separately from the case- I and case- II, by using our second module namely, the Interlingual module. The comparison algorithm decides this forwarding in phase I. The working of different components of an Interlingual module can be explained as below:-

- *Input:* - The input Hindi Idiom by the user for translation arrives as an input to the Interlingual module after comparison phase. For instance, the user input is- "सांच को आच नहीं । " [11].
- *Database1:* - Database1 contains Hindi Idioms together with their meaning in Hindi language only. This database1 is also used by the comparison phase for taking Idiom forwarding decision in phase- I, we are using this database twice in the system.
- *Mapper1:* - The mapper1 is a simple mapping algorithm which takes the input idiom and maps it correctly to its meaning in the same language by accessing Database1 which contains Hindi Idioms and corresponding to it its meaning. This meaning in Hindi is our intermediate representation1 i.e. IR1.The output of this stage for our instance is the meaning of that Hindi idiom in Hindi which can be given as---

  - "सच्चे इंसान को किसी का डर नहीं होता है । "

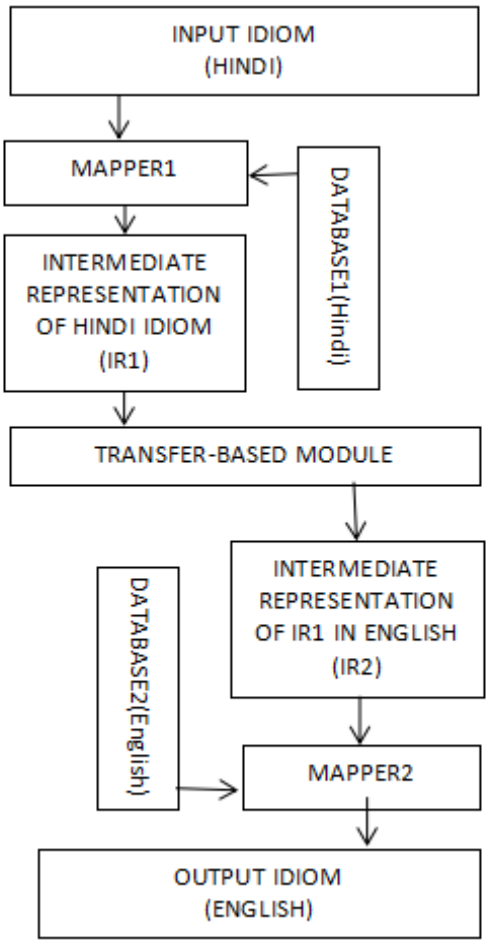

Fig 4. Block Diagram of Interlingual-Based Module [1]

- *Transfer-Based Module*: - In this part, we are re-using our transfer-based module. Now the output of mapper1 which is a sort of intermediate representation(IR) of our input Hindi Idiom is translated into English using above mentioned Transfer-Based MT approach. The output generated by Transfer-Based Module is our intermediate representation2 i.e. IR2. Thus in this module, we are actually using an approach similar to double Interlingual approach. "True person fears nothing " will be the output of this stage.
- *Database2:* - Database2 contains English Idioms together with their meaning in the English language itself.
- *Mapper2*: - This mapper2 takes the output of transfer-based MT module i.e. IR2 and maps it to its correct English Idiom using Database2 which contains English idioms and its meaning in English as IR2 is the meaning of Hindi Idiom in English. " Pure gold does not fear the flame " will be the output of mapper2 which will be pipelined to next stage.
- *Output*: - Output is the machine Translated English Idiom of given Hindi Idiom. We get output equivalent English idiom as--- " Pure gold does not fear the flame " [11][12].

The main architecture of idiom translation covers all types of Idioms structures possible in Hindi and English language using three cases, which are presented during the categorization of idioms. Using above system design, we can use machine translation for translating any type of Hindi Idiom into its equivalent English Idiom.

V. CONCLUSION AND FUTURE WORK

In this paper we have discussed research work in Machine Idiom Translation form 1952 to till date, together with which we come up with a new way of Hindi to English Idiom Translation using Double Interlingual approach encapsulated with classic but improvised Transfer-based approach, which is quite different from applying any one machine translation approach alone. Furthermore, this Hindi to English Idiom Translation system architecture can be extended for Hindi to any language Idiom Translation system. In future, this system can be embedded into other machine translation systems to get better translation results of that system.

REFERENCES

[1] Priyanka , Dr. R.M.K. Sinha, "A System for Identification of Idioms in Hindi", IEEE, 2014.
[2] Dimitra Anastasiou ," Idiom Treatment Experiments in Machine Translation ", The University of Saarland. Germany, 2010.
[3] M. Gaule, Dr. G. S. Josan, "Machine Translation of Idioms from English to Hindi", International Journal Of Computational Engineering Research (ijceronline.com), Vol. 2, Issue. 6
[4] S. P. Singh, A. Kumar, Dr. H. Darbari, A. Gupta, "Improving the quality of Machine translation using Rule Based Tense Synthesizer for Hindi". IEEE. 2015.
[5] J. Nair, Amrutha Krishnan K, Deetha R, "An Efficient English to Hindi Machine Translation System Using Hybrid Mechanism", 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE. 2016, Sept, 21-24.
[6] G. V. Garje and G. K. Kharate. "SURVEY OF MACHINE TRANSLATION SYSTEMS IN INDIA", International Journal on Natural Language Computing (IJNLC), Vol. 2, No.4, 2013, October
[7] N. Wagadiya, P. Ravarta, "English-Hindi Translation system with Scarce resources,. International journal of innovative research and development.
[8] Code project. (10 Aug 2010). Develop your own translation system [Online]. Available: http://www.codeproject.com/Articles/100126/DevelopYourOwnLanguageTranslationSystem
[9] L. R. Nair, David Peter S., "Machine Translation Systems for Indian Languages", International Journal of Computer Applications (0975 – 8887) , Volume 39– No.1, February 2012.
[10] A. Gehlot, V. Sharma, S. Singh, A. Kumar, "Hindi to English Transfer Based Machine Translation System", International Journal of Advanced Computer Research, Volume-5 Issue-19, June 2015.
[11] Prof. R. C. Phatak., "A few English Idioms with their Hindustani Equivanlents " in BHARGAVA'S STANDARD ILLUSTRARED DICTIONARY, vol. 10.
[12] N. K. Aggarwala , "Idioms" in Essentials of English Grammar and Composition, New Delhi,Goyal Brothers Prakashan, 2003.