A Survey of Tools and Techniques for Multiword Expression Detection

Ujwala P. Mahajan
School of Computer sciences
KBC North Maharashtra University
Jalgaon 425001, Maharashtra

Ajay S. Patil
School of Computer sciences
KBC North Maharashtra University
Jalgaon 425001, Maharashtra

Nita V. Patil
School of Computer sciences
KBC North Maharashtra University
Jalgaon 425001, Maharashtra

ABSTRACT
Multiword expressions (MWEs) find application in almost all NLP tasks such as machine translation, information retrieval, question-answering, etc. Multiword expressions are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasy. MWEs in Marathi are quite varied and many of these are of the types that are not encountered in English. This paper presents a survey regarding MWE research done for Indian and Foreign languages. The study and observations related to approaches, techniques, and features required to implement MWE detection system for various languages are reported.

Keywords
MWEs, Machine Translation, Information Retrieval, Multiword expressions.

1. INTRODUCTION
The current era is acknowledged as the information age, defined by the rapid accumulation, manipulation, and distribution of data. The exchange of information in our day happens not only between people but also between machines and humans. Computer science's natural language processing (NLP) branch handles the problems of human-computer interaction by doing things like creating and understanding natural language.

As these tasks deal with words or expressions, proper interpretation plays a key role. The primary challenge is to correctly interpret dialectal expressions and multiword expressions. Multiword expressions (MWEs) are used in almost all natural language processing (NLP) tasks, which includes information retrieval, question-answering, machine translation, and parsing, where the words that make up the MWE are considered as just one word in just one part of speech (POS) tag.

In Natural Language Processing (NLP), a challenging task is the proper treatment of multiword expressions (MWEs). Multiword expressions are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasy. MWEs in Marathi are quite varied and many of these are of the types that are not encountered in English. No comprehensive work has been reported on Marathi MWE. Examples of MWEs in English include conjunctions like 'as long as' (meaning ‘provided that '), idioms like ‘Break a leg ' (meaning ' Good luck '), and phrasal verbs like ' Turn down ' (meaning ' Reject '), etc. Some examples in Marathi are institutionalized adjective-noun pairs like 'prAcIna bhArata' ('ancient India' meaning a period of Indian history), idioms like 'gadhavala gulachi chav kay' ('A right thing in a wrong hand'), complex predicates like 'Swarthi Uddesh' (‘something done for Selfish reasons') and class expressions like 'ikde tikde' ('here there' meaning 'everywhere').

This paper presents a survey of Tools and Techniques implemented for various languages. This paper is organized into four sections. The first section discusses a study of Multiword Expressions. The second section reports tools available for MWE identification and Extraction; third reports techniques used for the implementation of MWE identification and Extraction systems for non-Indian languages. The fourth section reports approaches and techniques used for the implementation of MWE identification and Extraction systems for Indian languages.

2. STUDIES ON MWES
2.1 Types of Multiword Expressions
MWEs include a wide variety of phenomena that involve language. The MWE categories that Sag et al. (Year) had previously identified were further expanded upon by Gheneim et al. (Year). Based on their classification, MWEs come into two primary categories.

2.1.1 Lexicalized Phrases
These are expressions wherein the individual words do not convey their literal meanings. This category can be further subdivided into

2.1.1.1 Fixed Expressions
These are inflexible phrases like "by and large," "in short," and "every which way."

2.1.1.2 Semi-fixed Expressions
Those include proper nouns, compound nouns, and non-decomposable idioms.

2.1.1.3 Syntactically Flexible Expression
It include light verbs, decomposable idioms, and verb-particle constructions. A verb and one or more particles are joined together to form verb-particle structures, such "brush up" and "break up." Particle relocation can occur in some forms; "call him up" and "fight bravely on" are two examples. Syntactic flexibility can be observed in decomposable idioms that include "let the cat out of the bag" and "sweep under the rug." Basic verb formulations with little semantic substance, such "make a mistake" and "make a demo," are examples of light verb formations.

2.1.2 Institutionalized Phrases
Given both syntactic and semantic structure, these statements have characteristics associated with statistics. For example, due to its usual usage, "traffic light" cannot be substituted with other names like "traffic director" or "intersection regulator".

Rather than just simply linguistic, institutionalised phrases have a statistical peculiarity to them. The phrases "telephone booth" and "fresh air" are further examples.
By defining these types of MWEs, scientists are able to better understand the complexity of natural language processing, which open up the possibility to more productive computational analysis of linguistic expressions.

3. MWE SYSTEMS APPROACHES

3.1 Linguistic Approach/ Rule based Approaches (Symbolic, Syntactical or Linguistic pattern based method)

The approach used depends on the fundamental characteristics of the language that is being studied. These approaches extract Multiword Expressions (MWES) from corpora using morphosyntactic property indications, specifically designed to target particular forms of MWES. They usually make use of chunkers, shallow or deep parsers, and Part-of-Speech (POS) taggers.

3.2Statistical Approach (Statistical / Association measure-based method)

Collocation score measures are employed in statistical procedures that are independent of language, to determine MWES. Log Likelihood Ratio (LLR), Phi-Coefficient, Co-Occurrence Measurement, Significance Function, Point wise Mutual Information (PMI), Frequency of Bigrams, t-Score, and Dice's Coefficient are examples of common metrics. Of these, the PMI approach is frequently used as it is flexible and compatible with both languages and MWE types.

3.3 Hybrid Approaches

For MWES extraction, these approaches integrate statistical and linguistic methods. Typically, they include employing statistical techniques to determine real MWES and using POS tagging to find patterns at the word level. MWES can be handled completely with hybrid approaches, which combine the advantages of statistical and linguistic approaches.

4. MWE TOOLS

Table 1. MWEs Tools

MWE Tool	**Websites**
UCS toolkit	http://www.collocations.de/software.html
Bengali Shallow Parser	http://ltrc.iiit.ac.in/analyzer/bengali/
WEKA	www.cs.waikato.ac.nz/ml/weka
Semantic Analysis System (USAS)	https://ucrel.lancs.ac.uk/usas/
The Stanford Parser	http://nlp.stanford.edu/software/lex-parser.shtml
CRF chunker	http://crfchunker.sourceforge.net/
Protégé	http://protege.stanford.edu
Hindi POS Tagger	http://sanskrit.jnu.ac.in/pos/index.jsp
Hindi POS tagger	https://bitbucket.org/sivareddyg/hindi-part-ofspeech-tagger
CRF++	http://crfpp.googlecode.com/svn/trunk/doc/index.html
Yamcha	http://chasen-org/~taku/software/yamcha/
Tiny-0.07 classifier	https://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM
Stanford Segmenter	http://nlp.stanford.edu/software/segmenter.shtml
PyMUSAS	https://pypi.org/project/pymusas/
Giza++ word alignment tool	https://www2.statmt.org/moses/giza/GIZA++.html

A number of the tools shown in Table 1 can be used for free for the purpose to extract and identify Multiword Expressions. These tools are designed for particular languages and have various uses.
A set of scripts and tools for statistically analysing co-occurrence data is known as the UCS toolkit.Weka is Open-source software which allows data preparation, classification, regression, clustering, mining association rules, and presentation for machine learning algorithms.CRF Chunker is known for its fast chunking, this English Conditional Random Fields Phrase Chunker is Java-based. Protégé is an open-source programme for ontologies and terminologies creation and management. The Semantic Analysis System, or USAS, is a multilingual system for computerised semantic analysis of text. The Python Multilingual Ucrel Semantic Analysis System, or PyMUSAS, is a rule-based semantic tagger for tokens and multiword expressions that supports a number of semantic tagsets.
The automatic word alignment tool is called Giza++. The word alignment generated through GIZA++ is widely and easily used by the majority of statistical machine translation systems in the Linux system environment.

5. MWE FOR NON-INDIAN LANGUAGES

Many research studies on MWE have been carried out in a wide range of European languages, including German, English, and others. Attia and associates (2010) Mohammed Attia (2010)[42] uses three data sources: Wikipedia, Princeton WordNet, and Arabic Giga word Fourth Edition. He suggests employing cross-lingual correspondence asymmetries, translation-based approach, and corpus-based approach. There is basically little overlap between the corpus-based and other techniques. They said that it is too difficult to identify MWEs. Similar methods for automatically detecting Chinese MWEs are proposed by Piao et al. (2006), who achieve precision ranging from 61.16% to 93.96% for various categories [47]. A new approach for extracting synonymous collocations utilising semantic information was introduced by Li et al. in 2007[48]. Jian Xu et al. (2010)[49] describe a way to combine statistical tools and similarity metrics that has been suggested for automatically removing English MWEs from the corpus of work reports and white papers from the Chinese government between 1991 and 2010. When combined with the statistical tool, it was seen that the similarity measure produces a high recall in extracting MWEs [].when combined with the

statistical component [].Using a corpus of about 9,000 candidates for MWEs collected from a hundred million bigrams, Katuscak and Genci (2015) work on statistical and linguistic approaches in the identification of MWES in the Slovak language. It has been shown that the PMI approach, with its 85% precision, is the most appropriate statistical tool for identifying MWE in the text.[51]. A domain-independent clustering based approach for the automatic extraction of MWES was proposed by Elena Tutubalina in 2015. The methodology integrates textual data from Wikipedia articles with statistical data from a general-purpose corpus [52].

A technique for extracting multiword expression candidates from texts in an Indonesian corpus was proposed by Dani Gunawan et al. (2016). Uncommon multiword expressions compiled from internet local newspapers [54]. Tsvetkov and Wintner (2011) identified MWEs in a monolingual text by utilising a Bayesian Network framework with a variety of linguistic factors that capture word inflections together with a few statistical parameters [56]. For the purpose of extracting MWEs, Pecina (2008) employed linear logistic regression, linear discriminant analysis (LDA), and neural networks independently on a feature vector made up of 55 association metrics.[58]A rule-based processor for multiword expressions was proposed by Oflazer et al. [60]. In a corpus that has gone through morphological analysis and has parts of speech and inflections was identified the processor obtains multiword expressions.

Table 2. MWEs for Non-Indian Languages

Language	Year	Methods/Techniques	Remark & Result
English	2018	Hybrid technique incorporating Bigram and Trigram MWEs.	Achieved F-score of 39.17% using the "Maximizing Recall Method."
	2016	Filtering method + statistical information	Advanced evaluation on bigram + trigram set from the Leipzig corpora collection.
	2015	Annotation types: Constituency analyses, Dependency analyses, HPSG, LFG	Surveyed MWEs in various Treebank's.
	2011	Dictionary-based and machine learning methods	First corpus to annotate MWEs and named entities simultaneously: Wiki50 Corpus.
	2011	Match, POS rules, and Merge	Significant improvements observed over baseline methods, with potential for further enhancement highlighted by researchers.
	2010	Ranking and filtering method + statistical information	Contributions to Machine Translation and Information Extraction: Genia corpus.
	2009	Sentence surface features, Canonical form, Collins COBUILD Advanced Learner's Dictionary	Baseline accuracy of 82.7% on development set and 87.2% on test set.
	2007	Unsupervised method, Large corpora	First method to extract MWEs based on non-compositionality on a large scale.
	2005	Semantic tagger + statistical algorithm	Semantic tagger achieved high precision in extracting core MWEs.
Arabic	2010	Cross-lingual Correspondence Asymmetries, Translation-Based Approach, Corpus-Based Approach	Identification of MWEs using Wikipedia, Princeton WordNet, and Arabic GigaWord Fourth Edition.
Chinese	2010	Similarity measure and statistical tool	High recall in extracting MWEs when combined with the statistical tool, using Chinese government white papers
	2007	Filtering method + statistical information	Hybrid approach to automatically identify bi-gram noun and verb Chinese collocations.
	2006	Filtering method + Statistical information	Employed a sieve of statistical measures on the Chinese corpus built at CCID tool. Improved recall needed for practical applicability in Machine Translation systems.
French	2011	Tree substitution grammar	Improved baseline for parsing raw French text.
Slovak	2015	Statistical + linguistic methods	Focused on bigrams with 85% precision (using PMI method).
Russian	2015	Domain-independent clustering based approach .	Incorporated association measures from two corpora.
Indonesian language	2016	Collection from local online newspapers	Results usable for text pre-processing to obtain MWE candidates.
Swahili language	2008	Rule-based machine translation	Described various types of MWEs to enhance accurate further processing.
Hebrew	2011	Supervised method + linguistic information	Classifier able to identify MWEs of different types and constructions.
German and Czech	2008	Ranking + supervised method + statistical information	No general association measure recommended, results vary with data.

Korean	2001	Filtering method + statistical information	Defined statistics for Korean collocations and retrieved meaningful bigrams.
Turkish	2010	Ranking method + statistical information	Higher precision and recall with $\chi 2$ and mutual information methods.
	2004	Rule-based method + linguistic information	Achieved nearly 100% precision and around 60% recall.

6. MWE FOR INDIAN LANGUAGES

Detection of Multiword Expressions (MWEs) in Indian languages presents unique challenges due to the diversity of MWE types and the linguistic complexity inherent in these languages. Unlike in English, where specific MWEs are common, Indian languages such as Urdu, Bengali, Manipuri, Hindi, etc., exhibit a wide range of MWE structures.

While researchers have explored various methods for MWE extraction and identification in Indian languages, the process is still complex. Few research have been done, emphasising how challenging it is to use linguistic methods exclusively for MWE detection. As a result, scientists are beginning to agree more and more about how useful statistical techniques are in this situation.

The need for hybrid model-based language-independent MWE detection systems is rising due to the complexity of the problem. These models enhance the robustness and accuracy of MWE recognition in Indian languages by combining statistical methods with linguistic insights. A strategy like this has potential to solve the problems caused by India's large and diverse linguistic environment.

Table:-3 MWEs for Indian Language

Language	**Year**	**Technique/Algorithm**	**Accuracy**
Punjabi	2018	Linguistic property	F-Score:- 27% to 97%
Urdu	2017	Automatic, semi-automatic, and manual approach	-
Assamese and Bodo	2013	Statistical measure + Language specific knowledge	-
Bengali	2015	Linguistic methods and Statistical models	-
	2014	Unsupervised approach with various statistical evidence	P:- 39.64%, R:- 91.29%, F:- 55.28% (top 1000 Candidates)
	2013	Machine learning approach + Random forest	Weighted average: 0.869
	2013	Supervised machine learning approach using Random Forest	F-measure: 85%
	2012	Conditional Random Field (CRF) machine learning technique	CRF-PP: P:- 67.98%, R:- 0.01%, F:- 73.01%
	2011	Semantic Clustering	Precision: 80.90%
	2010	Rule-based approach	P:- 92.82%, R:- 91.50%, F:- 92.15%
Hindi	2017	NLTK processing supporting Unicode format	F-measure score of many types > 90% (27% - 97%)
	2016	CRF	P:- 88.89%, R:- 85.66%, F:- 87.32%, Overall Accuracy: 89.76%
	2015	Linguistic Methods	P:- 89.42% to 91.77%, R:- 90.29% to 91.20%, F:- 89.86% to 91.48%
	2011	Linguistic Knowledge	F-Score: 27% to 97%
	2009	Linguistic property	CPs with average precision of 89% and recall of 90%
	2008	Statistical methods	Recall: 80%, Precision: 23% (at rank 1000)
	2008	Linguistic Methods	Accuracy: 98%
	2007	Classification-based approach for extracting noun +verb collocations	Classification accuracy of maximum entropy model: 77.24%
	2006	Giza++ Probabilities	P: -54%, R:- 44%, F:- 49%, AER:- 0.5155
	2006	Corpus-based lexicon of CPs	Precision: about 82.5%, Recall: 40%
Manipuri	2012	Conditional Random Field (CRF) machine learning techniques	CRF system: P:- 73.15%, R:- 78.22%, F:- 75.60%. With identification of RMWE:

			P:- 74.31%, R:- 80.20%, F:- 77.14%
	2011	Linguistic Rules or a large amount of annotated corpora	Best feature set in the model: P:- 86.84%, R:- 64.08%, F:- 73.74%
	2011	Conditional Random Field (CRF) machine learning techniques	P:- 91.90%, R:- 92.91%, F:- 92.40%
	2011	CRF-based GA in feature selection	P:-86.84%, R:- 64.08%, F:- 73.74%
	2011	Support vector machine (SVM) learning technique	Improved translation output by integrating RMWEs, MNEs, and non-NEs
	2010	Support vector machine (SVM) learning technique	For SVM-based RMWE System: P:- 93.53%, R:- 94.62%, F:- 94.07%. Multiword NE: P:-93.12%,R:- 94.82%, F:- 93.96%
	2010	Conditional Random Field (CRF) machine learning techniques	P:-85.53%, R:-60.39%, F:-70.83%. After RMWE identification: P:- 86.06%, R:- 62.24%, F:- 72.24%

Language-specific knowledge and measurements of statistics (PMI and Chi-square) have been applied by Anup Kumar Barman et al. (2013).Textbook corpus is used by the system [4]. Three methods have been used by Vaishali Gupta (2017): manual, semi-automated, and automatic. Using news websites, the authors gathered a corpus of 50,000 English sentences from the Health & Tourism domain. An expert in the language then translated the corpus into Urdu [1]. English-Punjabi parallel corpus derived from grammar books and news articles were presented by Kapil Dev Goyal in 2017 and 2018, and the linguistic property achieved an F-Score ranging from 27% to

97%[2][3].In 2015, Md. Jaynal Abedin et al. presented an automated technique for locating and retrieving MWEs in Bengali. The author detailed several MWE kinds and MWE extraction techniques in order to build this approach. The authors made the decision to create a MWE extraction method for Bengali in order to solve the problem of POS tagging. In order to develop this approach, they first determined the frequency of bigram phrases in that context by identifying MWEs in Bengali, such as phrases, reduplicative words, idioms, compound nouns, verbal phrases, and so on. The Pointwise Mutual Information and Chi Square test methods were used to find this frequency. Ultimately, a medium-sized corpus was used to test this approach, and the results were satisfactory [5].

In another paper, Chakraborty (2014) looks at Noun-Noun bigram collocations from a medium-sized untagged Bengali corpus of Rabindranath Tagore's articles. It does this by using a straightforward unsupervised approach along with a variety of statistical evidence to demonstrate the affinity of each bigram candidate's constituents as evidence of the MWE and to construct a weighted measurement to differentiate between MWE and non-MWE. The experimental results show that co-occurrence distribution-based functions yield more accurate results than frequency-based measuring techniques. This article stated that complete identification of MWEs in Bengali is still a work in progress due to a lack of lexical resources [6]. A machine learning based method for recognising noun-noun compound MWEs from the Bengali corpus is presented by Gayen and Sarkar et al. (2013).In their suggested system, the F-measure value obtained is 86.9% [7].When identifying compound nouns in Bengali, Gayen and Sarkar (2013) use a random forest technique that makes advantage of certain syntactic elements and association metrics[8]Nominal multiwords were detected by Chakraborty (2012) both with and without post processing using CRF[9].A method for locating bigram noun-noun MWEs from a medium-sized Bengali corpus is presented by Chakraborty et al. (2011). They observe that Bengali MWEs can be discovered with great benefit using English WordNet. At the cut-off of 0.5, WordNet accurately recognises maximum MWEs (precision 80.90 percent)[10].Using a rule-based methodology, Chakraborty and Bandyopadhyay [2010] introduced the study of MWEs in Bengali for the analysis of Rabindranath Tagore's

writings. There were two stages to the task. Semantic classification came after the initial step of finding MWEs for the Bengali language at the expression level. Additionally, they looked for to distinguish between the good and negative meanings of the Bengali language [11].In order to process and analyse MWEs for a Hindi text corpus extracted from the well-known Hindi novel "Karama Bhumi by Munshi PremChand," Rakhi Joon[2017] employed NLTK[12].Word Embeddings and WordNet-based features were used by Dhirendra Singh (2016) to detect the mean verb ending suffix (MWES) of two different types of noun compounds and noun+verb compounds in the Hindi language[13].The Hindi compound adverbs MWES have been identified by Rakhi Joon in 2017.The method was developed and tested with the help of the CFILT Hindi Corpus.A method for advancing mining of multi-word phrases in Hindi utilising linguistic knowledge has been presented by Mahesh & Sinha [2011]. These were found using a Hindi Wordnet and a bilingual English-Hindi corpus [17].Sinha (2009) used a parallel corpus of Hindi and English to extract complex predicates utilising the linguistic feature of light verbs [19].

Venkatapathy et al. (2007) obtained Hindi noun+verb collocations using a classification-based method [23]. The 'Noun+Verb' expression in Hindi was determined by Venkatapathyet al. [2005] and classified as a multi-word expression between 1 and 4[24].For Manipuri, another Indian language, Nongmeikapam et al. [2012] discovered the MWEs and regenerated multiword phrases. Their driving force was to enhance Manipuri language POS tagging. They recognised various kinds of MWEs [25].Kishorjit and Bandyopadhyay [2011] presented a CRF-based transliteration of Manipuri's recognised MWEs. When comparing the transliterated result to the output of Bengali Script Manipuri and Meitei script, the accuracy is 90.01% [26].Using the CRF model, Kishorjit and Bandyopadhyay [2011] developed RMWES [27].

The integrated model developed by Nongmeikapam and Bandyopadhyay [2011] can identify MWE based on CRF analysis, however it has modifications to the feature list and feature selection. The idea of a genetic algorithm is used in the feature selection process. In comparison to the CRF-based

Manipuri MWE identification without GA application, this model shown an improvement with a recall of 64.8%, precision of 86.84%, and F-measure of 73.4% [28].Manipuri Named Entities (NEs) were identified by Doren Singh and Bandyopadhyay [2010], and the SVM-based machine learning approach is used to identify and categorise RMWEs [30].Nongmeikapam and Bandyopadhyay (2010) talked about the detection of Manipuri reduplicated MWE [32].

7. CONCLUSION

This paper presented a review of the literature on Tools and Techniques for Multiword Expression Detection for Indian and non-Indian languages. It brings attention to the present progress in machine learning for Indian languages while identifying the major challenges, including the absence of annotated corpora, a wide range of writing styles, and complex morphology. Understanding the variety of MWE approaches like linguistic, statistical, machine learning, and hybrid methods highlights how important it is to combine methodologies to achieve accurate identification, particularly when dealing with Indian languages.

Furthermore, highlighting the absence of studies on MWE in particular Indian languages, such as Marathi, indicates the necessity of additional research and the creation of specialised methods for MWE in these languages.

8. REFERENCES

[1] Gupta, Vaishali, Nisheeth Joshi, and Iti Mathur. "Approach for multiword expression recognition & annotation in urdu corpora." *2017 Fourth International Conference on Image Information Processing (ICIIP)*. IEEE, 2017.

[2] Goyal, Kapil Dev, and Vishal Goyal.(2018) "Extraction of Named Entities from Punjabi-English Parallel Corpora."

[3] Goyal, Kapil Dev, and Vishal Goya(2017)l" Multiword Expressionsin Indian Languages "

[4] Barman, A. K., Sarmah, J., & Sarma, S. K. (2013, August). Automatic identification of assamese and bodo multiword expressions. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 26-30). IEEE.

[5] Md Jaynal Abedin , Bipul Syam Purkayastha and Kh. Raju Singha 2015 Automated Multiword Expressions Detection in Bengali Md Jaynal Abedin et al. / International Journal of Computer Science Engineering (IJCSE) ISSN: 2319-7323 Vol. 4 No.02 March 2015

[6] Tanmoy Chakraborty 2014 Towards Identification of Nominal Multiword Expressions in Bengali Language Open Access Library Journal, 1: e582. http://dx.doi.org/10.4236/oalib.1100582

[7] Vivekananda Gayen and Kamal Sarkar. 2013. Automatic identification of Bengali noun-noun compounds using random forest. In Proceedings of the 9th Workshop on Multiword Expressions, pages 64–72, Atlanta, Georgia, USA, June. Association for Computational Linguistics

[8] Vivekananda Gayen, Kamal Sarkar, "A Machine Learning Approach for the Identification of Bengali Noun-Noun Compound Multiword Expressions", and Proceedings of ICON-2013: 10th International Conference on Natural Language Processing, http: //ltrc.iiit.ac.in/proceedings /ICON-2013.

[9] Chakraborty, Tanmoy. "Identification of Nominal Multiword Expressions in Bengali using CRF." *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*. IEEE, 2012.

[10] Tanmoy Chakraborty, Dipankar Das, Sivaji Bandyopadhyay, "Semantic Clustering: an Attempt to Identify Multiword Expressions in Bengali", Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA, 23 June 2011, pp. 8–13.

[11] Chakraborty, T., Bandyopadhyay, S.: Identification of reduplication in Bengali corpus and their semantic analysis: a rule-based approach. In: Proceedings of 23rd International Conference on Computational Linguistics, pp. 73–76 (2010)

[12] Joon, R., & Singhal, A. (2017). Analysis of MWES in Hindi text using NLTK. *Int. J. Nat. Lang. Computer*, *6*(1), 13-22.

[13] Dhirendra Singh, Sudha Bhingardive, Pushpak Bhattacharyya, "Multiword Expression Dataset for Indian Languages", LREC 2016. Portoroz, Slovenia, May 23-28 2016.

[14] Singh, R., Ojha, A. K., & Jha, G. N. (2016). Classification and identification of reduplicated multiword expressions in Hindi. *Classification and identification of reduplicated multiword expressions in Hindi, WILDRE*, 18-22.

[15] .Patel, Dhirendra Singh Sudha Bhingardive Kevin, and Pushpak Bhattacharyya, "Detection of Multiword Expressions for Hindi Language using Word Embeddings and WordNet-based Features" In 12th International Conference on Natural Language Processing, p. 291, 2015

[16] Joon, R. & Singhal, A., (2015), "A System for Compound Adverbs MWEs extraction in Hindi." in the proceedings of International Conference on Contemporary Computing (IC3), IEEE Computer Society, Vol. 8, pp336-341.

[17] R. Mahesh K. Sinha "Stepwise Mining of Multi-Word Expressions in Hindi" Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA, 23 June 2011. @ 2011 Association for Computational Linguistics

[18] S. Venkatapathy, and A. Joshi, "Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features", In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Association for Computational Linguistics, pp. 899-906, 2009

[19] R. Mahesh K. Sinha. Mining Complex Predicates In Hindi Using Parallel Hindi-English Corpus, ACL-IJCNLP 2009 Workshop on Multi Word Expression, Singapore. 2009b

[20] Debasri Chakrabarti, Hemang Mandalia, Ritwik Priya, Vaijayanthi Sarma, Pushpak Bhattacharyya (2008) Hindi Compound Verbs and their Automatic Extraction Coling 2008: Companion volume – Posters and Demonstrations, pages 27–30 Manchester, August 2008

[21] Anoop Kunchukuttan, Om P. Damani, " A System for Compound Noun Multiword Expression Extraction for Hindi", Proceedings of ICON-2008: 6th International Conference on Natural Language Processing, Macmillan Publishers, India.

[22] Amitabha Mukerjee, Ankit Soni, and Achla M Raina. 2006. Detecting complex predicates in hindi using pos projection across parallel corpora. In Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, pages 28–35. Association for Computational Linguistics

[23] Venkatapathy, Sriram, and Aravind K. Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In Proceedings of Coling-ACL 2006: Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, pp. 20-27.

[24] Venkat apathy, S., Agrawal, P., Josh, A.K. : Relative compositionality of Noun +Verb multi-word expressions in Hindi. In: Proceedings of ICON Conference on Natural Language Processing, Kanpur (2005)

[25] Nongmeikapam, K., Nonglenjaoba, L., Nirmal, Y., Bandyopadhyay, S.: Reduplicated MWE (RMWE) helps in improving the CRF based Manipuri POS tagger. Int. J. Inf. Technol. Computer Sci. 2(1), 45–59 (2012)

[26] N. Kishorjit, L. Dhiraj, N. Bikramjit Singh, Ng. Mayekleima Chanu, and B. Sivaji, Identification of Reduplicated Multiword Expressions Using CRF, A. Gelbukh (Ed.):CICLing 2011, LNCS vol.6608, Part I, pp. 41–51, Berlin, Germany: Springer-Verlag

[27] Kishorjit Nongmeikapam, NingombamHerojit Singh, Bishworjit Salam and Sivaji Bandyopadhyay. 2011. Transliteration of CRF Based Multiword Expression (MWE) in Manipuri: From Bengali Script Manipuri to Meitei Mayek (Script) Manipuri. International Journal of Computer Science and Information Technology, vol.2 (4) . pp. 1441-1447

[28] Kishorjit Nongmeikapam, Sivaji Bandyopadhyay, "Genetic Algorithm (GA) in Feature Selection for CRF Based Manipuri Multiword Expression (MWE) Identification", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 5, Oct 2011, pp 53-66

[29] Singh, Thoudam Doren, and Sivaji Bandyopadhyay. "Integration of reduplicated multiword expressions and named entities in a phrase based statistical machine translation system." Proceedings of 5th international joint conference on natural language processing. 2011.

[30] N. Kishorjit, and B. Sivaji, Identification of Reduplicated MWEs in Manipuri: A Rule based Approached. In the Proceeding 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL-2010), San Francisco, pp 49-54, 2010

[31] N. Kishorjit and S. Bandyopadhyay, Identification of MWEs Using CRF in Manipuri and Improvement Using Reduplicated MWEs, In the Proceedings of 8th International Conference on Natural Language (ICON-2010), IIT Kharagpur, India, pp 51-57, 201

[32] Thoudam Doren Singh and Sivaji Bandyopadhya Web Based Manipuri Corpus for Multiword NER and Reduplicated MWEs Identification using SVM Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pages 35–42, the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010

[33] Scott Songlin Piao a, Paul Rayson b, Dawn Archer a, Tony McEnery(2005) Comparing and combining a semantic tagger and a statistical tool for MWE extraction Computer Speech and Language 19 (2005) 378–397

[34] Tim Van de Cruys and Begona Villada Moir ˜ on: Semantics-based Multiword Expression Extraction Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 25–32, Prague, June 2007. c 2007 Association for Computational Linguistics

[35] Ram Boukobza, Ari Rappoport, "Multi-Word Expression Identification Using Sentence Surface Features", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6-7 August 2009, pp. 468–477

[36] Ramisch, C., Villavicencio, A., Boitet, C.: mwetoolkit: a Framework for Multiword Expression Identification, LREC, 2010.

[37] Istvan Nagy T, Veronika Vincze, Gabor Berend, "Domain-dependent identification of multiword expressions", proceedings of the Conference: Recent Advances in Natural Language Processing, RANLP 2011, Hissar, Bulgaria, 12-14 September 2011.

[38] Veronika Vincze, Istvan Nagy T., and Gabor Berend,[2011] Multiword Expressions and Named Entities in the Wiki50 Corpus Proceedings of Recent Advances in Natural Language Processing, pages 289–295, Hissar, Bulgaria, 12-14 September 2011

[39] Victoria Rosen, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejcek, Agata Savary, Adam Przepiorkowski, Petya Osenova, and Verginica Barbu Mititelu A Survey of Multiword Expressions in Treebanks Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14) 2015

[40] S.Kumova-Metin, "Neighbour Unpredictability Measure in Multiword Expression Extraction", International Journal of Computer Systems Science and Engineering: 31-3, 2016.

[41] Agrawal, S., Sanyal, R., & Sanyal, S. (2018). Hybrid method for automatic extraction of multiword expressions. *International Journal of Engineering & Technology*, *7*(2.6), 33-38.

[42] Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, Josef van Genabith, "Automatic Extraction of Arabic Multiword Expressions", Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), Beijing, August 2010, pp 19–27

[43] Attia, Mohammed. 2006. Accommodating Multiword Expressions in an Arabic LFG Grammar. In Salakoski, Tapio, Filip Ginter, Sampo Pyysalo, Tapio Pahikkala (Eds.): Advances in Natural Language Processing. Vol. 4139, pp. 87–98. SpringerVerlag: Berlin, Heidelberg.

[44] Boulaknadel S, Daille B &Aboutajdine D. (2008). A multi-word term extraction program for Arabic language. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp. 1485–1488, Marrakech, Morocco

[45] Habash, Nizar, Owen Rambow and Ryan Roth. 2009. A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In the 2 nd International Conference on Arabic Language Resources and Tools (MEDAR

2009), pp. 102–109. Cairo, Egypt

[46] Bar, M. Diab and A. Hawwari, "Arabic Multiword Expressions," in Language, Culture, and Computation. Computational Linguistics and Linguistics: Essays Dedicated to Yaacov Choueka on the Occasion of His 75th Birthday, Part III, N. Dershowitz and E. Nissan, Eds., Berlin, Springer Berlin Heidelberg, 2014, pp. 64-81

[47] Piao, S, Sun, G, Rayson, P and Yuan, Q "Automatic extraction of Chinese multiword expressions with a statistical tool" Paper presented at Workshop on Multi-word-expressions in a Multilingual Context held in conjunction with the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, 2006

[48] W. Li, Q. Lu and J. Liu," Chinese typed collocation extraction using corpus based syntactic collocation patterns", IEEE NLP-KE 2007 - Proceedings of International Conference on Natural Language Processing and Knowledge Engineering,2007.

[49] Jian Xu, Jingsong Yu, Huilin Wang(2010) Automatic Extraction of Multiword Expressions Combining Statistical and Similarity Approaches 2010 Fourth International Conference on Genetic and Evolutionary Computing.

[50] Spence Green, Marie-Catherine de Marneffe, John Bauer, Christoper D. Manning, "Multiword expression identification with tree substitution grammars: A parsing TOUR DE FORCE with French", Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, , John McIntyre Conference Centre, Edinburgh, UK, 27-31 July 2011, pp.725-735

[51] Matej Katuscak, Jan Genci, "Identification of Multiword Expressions for the Slovak Language", Proceedings of the Faculty of Electrical Engineering and Informatics of the Technical University of Kosice, 2015.

[52] Elena Tutubalina (2015) Clustering-based Approach to Multiword Expression Extraction and Ranking Proceedings of NAACL-HLT 2015, pages 39–43, Denver, Colorado, May 31 – June 5, 2015. c 2015 Association for Computational Linguistics

[53] Natalia Loukachevitch1,2(B) and Ekaterina Parkhomenko2 Recognition of Multiword Expressions Using Word Embeddings: 16th Russian Conference, RCAI 2018, Moscow, Russia, September 24-27, 2018, Proceedings

[54] Dani Gunawan, Amalia Amalia, Indra Charisma (2016)Automatic Extraction of Multiword Expression Candidates for Indonesian Language 2016 6th IEEE International Conference on Control System, Computing and Engineering, 25–27 November 2016, Penang, Malaysia

[55] Arvi Hurskainen, Multiword Expressions and Machine Translation Technical Reports in Language Technology Report No 1, 2008

[56] Y. Tsvetkov and S. Wintner, "Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources", Proceedings of the 2011Conference on Empirical Methods in Natural Language Processing, pages:836-845, Edinburgh, Scotland, UK, July 2731,2011

[57] P.Pecina, "A Machine Learning Approach to Multiword Expression Extraction", Proceedings of the LREC 2008 Workshop towards a Shared Task for Multiword Expressions, 2008.

[58] Weller, M., Heid, U., 2010. Extraction of German multiword expressions from parsed corpora using context features. In: Proceedings of LREC 2010

[59] S. Kim, J. Yoon and M. Song,"Automatic Extraction of Collocations from Korean Text", Computers and the Humanities 35: 273–297, 2001.

[60] K.Oflazer, O.Çetinoğlu, and B. Say," Integrating morphology with multi-word expression processing in Turkish", Proceedings of the Workshop on Multiword Expressions: Integrating Processing, .p. 64-71, 2004.

[61] S.Kumova-Metin andB. Karaoğlan,, "Collocation Extraction in Turkish Texts Using Statistical Methods", 7th International Conference on Natural Language Processing (LNCS-ISI) IceTAL, Reykjavik, Iceland,2010