ISSUES AND CHALLENGES OF REDUPLICATION IN HINDI

[1]S. K. Gupta, [2] Kamlesh Dutta, [3]Prince Rana
[1]Associate Professor, [2]Associate Professor, [3]Research Scholar
[1]Department of Computer Science & Engineering, BCET, Gurdaspur
[2]Department of Computer Science & Engineering, NIT, Hamirpur
[3]Research Scholar, IKGPTU,Kapurthala
[1]skgbcetgsp@gmail.com,[2]kdnith@gmail.com,3erprincerana15@gmail.com

ABSTRACT

In this paper, the author presents two different techniques for recognition of reduplication in Hindi language text. The reduplicated words representation in Hindi is broadly categorized into three patterns: (i) repetition of same syllable twice (ii) repetition of syllable partially (iii) both the syllable are totally different. In Hindi reduplicated words are created either by using multiple syllables or by only single syllables or totally different. We have combine corpus based approach with some rules to handle reduplication in Hindi language. These rules basically emphasis on the semantic characteristics of the language. We have created our corpus from number of Hindi books and from Hindi literature. Here we are focusing on two techniques firstly whether the reduplicated word found in the corpus and other whether the rule satisfies these reduplicated words constraints. Finally we have shown the result and application of our process.

KEYWORDS

Reduplication, Morphemes, Inflection, Derivational, Affix

INTRODUCTION

Handling of reduplication is considered to be a problem which requires special methods. Reduplication means when a root of a word is repeated or change in some part of word is called reduplication. It is very must concerned with phonological as well as with morphological patterns. Reduplication comes from Latin reduplicate whose meaning is doubling. The word which is repeated can be fixed at anywhere either at middle, start or at the end. Word is always divided into two parts root and inflection. When we makes plural of a word then this is called inflectional reduplication (e.g. table tables) and when we makes exact word is called derivational reduplication (e.g. hi hi). In derivational either the morpheme is added to root word or the word is stemmed.

Examples of reduplication in different languages:

In Afrikaans:

Amper	nearly	amper·amper	**very** nearly
dik	thick	dik·dik	**very** thick [18]

In Motu:

Tau	man	ta·tau	m**e**n	
mero	boy	me·mero	boy**s**	[18]

In Tagalog:

sumulat	to write	su·sulat	**will** write	
bumasa	to read	ba·basa	**will** read	[18]

Reduplication plays different role in different languages. Reduplication can occur both in written text and spoken words. Lots of reduplication words are used in conversation. The main area where reduplication formed is in books, magazines, novels and articles. These words are very few used in academics text types. Reduplication also occurs in baby's talk like mani for pani. Morphemes also plays important role in reduplication. These are the smallest words which completes a word like es, er, un etc. The process of making words with these morphemes is called morphology. Morphemes plays important role in reduplication. Reduplication handles properly when we know the proper syntax and semantic of a word.

In this paper section 2 surveys previous work carried out in the area of reduplication in Hindi, English and some other languages. Section 3 tells about reduplication occurs in Hindi. Section 4 describes about the approaches that can be used to handle reduplication. Section 5 shows the approach used for handling of these reduplicated words. Section 6 shows the model for this. Section 7 shows the evaluation and results. Section 8 shows the conclusion and future scope.

RELATED WORK

Research in Hindi Reduplication is not done so far as compared to the other languages like English, Spanish.In this paper author discusses the reduplication process in Bengali. It basically emphasis on differentiation of Bengali reduplication words with Hindi and English words to describes the various types of reduplication that can occur in Bengali language. [12]

In this author describes reduplication in Vietnamese language. Vietnamese is a language where words are created with the combination of multiple syllables whose phonics is similar. It gives description about various types of reduplication in Vietnamese and makes use of optimal finite state devices in particular minimal sequential string to string transducers to build computational model for every efficient recognition and production of these words. At last it gives applications of this model. [10]

In this author describes reduplication in Indonesian language. In this it basically deals with morphophonemic fact relating to sound changes in morphemes and how the construction of word formation rules are done to create these derived words exhibiting reduplication. It describes various tools to construct the word. It also gives description about morphological analyzer. [13]

In this paper author discusses the reduplication in kashmiri language. Here it shows the reduplication in Kashmiri at expression level in which noun, pronoun, adjective, verbs, clauses and phrases are defined. It also shows the reduplication at semantic level, word level, echo words and onomatopoeic words. [15]

In this paper author discusses the reduplication in Kinyarwanda language. It describes the reduplication as bounded and unbounded reduplication. They have created a combination of two level rules and replace rules. Bounded reduplication are those which involves just repeating a given part of word and

unbounded in which bounded reduplication involves copying a fixed number of morphemes. Bounded reduplication is solved up to some extent but the unbounded is more challenging. [7]
In this author discusses about the Manipuri language. It is developed for reduplication multiword expression (MWE) and multiword named entity recognition (NER). Here they collect a news corpus from a Manipuri news website. They have collected four and half a million words. Corpus collection and reduplication is based on support vector machine (SVM) learning techniques are reported. [19]

Reduplicaton is used in number of languages. There are **Proto indo European** languages which basically emphasis on partial reduplication. E.g. hald(I hold) becomes haihald(I held). In **Dutch** reduplication also exists e.g. bonbon. In **Afrikans** reduplication is mostly used for repeated words e.g. krap(to scratch oneself) becomes krap-krap-krap(to scratch oneself vigorously). In **Romance** these used reduplication to create new words and word associations. E.g. via via. They commonly used reduplication for changing person names. E.g gopi to gops. In **Nepalese** reduplication is used to create number of nouns by reduplication. E.g. hina mina means scattered and khan asana for food. Reduplication is also used in number of other languages like Punjabi, Bengali, Malay, Indonesia, Hebrew, Bantu etc. In **Papago** language reduplication is used for making the plural of a word. E.g. kuna for husband and kuukuna for husbands. In **motu** language here the meaning of mero means boy and where meromero means little boy. As I already said reduplication work differently in different language.

REDUPLICATION IN HINDI

There are various types of reduplication exists in Hindi Language.
1 Complete Reduplication: When every individual word is important and fully duplicated is called complete reduplication. This type of reduplication found in Hindi.

S.No.	In Word	In Sentence
1.	**साथ साथ** (Saath Saath)	**हम सब** साथ साथ है(Hum sab saath saath hain)
2.	**अलग अलग**(Alag Alag)	सब अलग अलग चलो (Saba alag alag chalo)

Table 1 Example of complete reduplication

2 Partial Reduplication: When one of the word or syllable is changed i.e. either from start, end or on the middle of the word is called partial reduplication.

S.No.	In Word	In Sentence
1.	**उलट पुलट** (Ulat Palat)	सब **उलट पुलट है** (Sab ulat palat hain)
2.	**आस पास** (Aas Pas)	अपना **आस पास** साफ रखो (Apna aas pas saaf rakho)

Table 2 Example of Partial reduplication

3 Semantic Reduplication: When two words are repeated with some change and semantically repeated is called semantic reduplication.

S.No.	In Word	In Sentence
1.	वाग वागीचा (Bag	शिमला वाग वागीचा से भरा पड़ा है (Shimla bag

	Bagicha)	bagicha se bhara padah hain)
2.	चाय पानी (Chai Pani)	चाय पानी पी के जाना (Chai pani p eke jana)

Table 3 Example of Semantic reduplication

4 Syntax Reduplication: When the main emphasis is on syntax. It means that some word are reduplicated by space or some are used with hypen (-) sign in between them.

S.No.	In Word	In Sentence
1.	ईधर उधर (Idhar Udhar)	ईधर उधर मत देखो (Idhar udhar mat dekho)
2.	लेन-देन (Len Den)	लेन-देन की वात कर लो (Len den ki baat kar lo)

Table 4 Example of Syntax reduplication

5 Onomatopoeic Reduplication: When two words produces sound or refers to sound is called onomatopoeic reduplication.

S.No.	In Word	In Sentence
1.	छ्रन छ्रन (Chan Chan)	पानी छ्रन छ्रन करता है (Pani chan chan karta hai)
2.	टिक टिक(Tik Tik)	टिक टिक की आवाज आ रही है (Tik tik ki awaaz aa rahi hai)

Table 5 Example of onomatopoeic reduplication

6 Numeral Reduplication: When two words have some numerals in it called numeral reduplication.

S.No.	In Word	In Sentence
1.	एक एक (Ek Ek)	एक एक कर के आओ (Eke k kar ke aao)
2.	दो दो (Do Do)	दो दो चार होते है (Do do char hote hain)

Table 6 Example of Numeral reduplication

7 Noun Reduplication: When the words present are noun is called noun reduplication.

S.No.	In Word	In Sentence
1.	घर घर (Ghar Ghar)	घर घर मे शोर है (Ghar Ghar mein shor hain)
2.	गली गली (Gali Gali)	गली गली दुकान है (Gali Gali mein dukaan hain)

Table 7 Example of Noun reduplication

8 Adjective Reduplication: When two words showing adjectiveness and they are duplicated is called adjective reduplication.

S.No.	In Word	In Sentence
1.	वड़ी वड़ी (Badi Badi)	वहां वड़ी वड़ी दुकान है (Waha badi badi dukaan hain)
2.	पीले पीले (Peele Peele)	हर तरफ पीले पीले फूल है (Har taraf peele peele phool hain)

Table 8 Example of Adjective reduplication

APPROACHES TO REDUPLICATION

There are various approaches that are used to find and resolve problem of reduplication.
1 Copy and association rule states that the reduplication is just a special type of affixation. Here an empty template is attached to stem which can be defined by CV slots or syllabic shapes. After attaching the CV skeleton the melody of the base must be associated with skeleton and if necessary all non associative elements are deleted.
2 Full copying Model states that full reduplication is the only reduplication exists in language. Here partial reduplication is eliminated and the main emphasis is on the full copying of first part of the word. This phenomenon is valid for only some languages which have only fully reduplicated words.
3 Corpus based Technique here the corpus for a language is created. With the help of this corpus researchers tries to find the matches in a literature or in a book. Corpus of word is formed in which we have to check how many words are identical with its next word. Corpus is also used to check whether a word is different from other by single syllable or fully reduplicated. We can make corpus from some online news and websites or from some literatures. We can also collect this form various books, animal names, fruit names or thing present in our surroundings. This is up to us whether we are created these reduplicated words or we are finding these reduplicated words from number of lines.
4 Rule based Technique here we have to create number of rules for finding and handling reduplicated words in language. Rules depend on the language used. For applying rules on some data we must requires tags for their words. We must have a good collection of rules for getting the good results. Semantic properties also play important role reduplication of words. These tell us the difference between two paired words.
5 Morphology based technique based on the large number of morphemes that are used in reduplicated words. Here we work on various morphemes and with the help of these morphemes we know how the word is different from other. It means that we are going to know that whether the word is plural of a word or having some reduplication in it.
6 Phonology based technique which is a sound based technique and depends on the pattern of sound of Hindi. Here the main emphasis is on the acoustic of the sound. By knowing the sound pattern we are able to find the reduplicated words in speech. Whenever we are combining the morphology with phonological processes it is said to be true phonology.
7 Optimality theories imply that it does not create any rules for any words. It assumes morphological and phonological parameters simultaneously. Here the numbers of outputs are generates from process and then low ranked and which violates the fewest changes is selected as output. Ranking of constraint is language specific.

APPROACH USED

Every language has different semantics based on needs. Different approaches are used for different languages because each and every language is semantically different. Every language gives different meaning when words are paired or comes in plural form. Every language has also its own way of representing and writing of these words. We have used hybrid approach for handling these reduplicated words. We have work on corpus based and rule based approach. This hybrid model is combination of two models: Rule based model and corpus based model. Firstly the recognition process starts through implemented rules then the confirmation of reduplication is done by comparing with saved corpus. We have created 25 rules for finding reduplication word in Hindi. E.g. if the first character of the word is different and others are same it means that there must be partial reduplication exists between them. We have created rules for finding all types of reduplication in Hindi.

THE PROPOSED MODEL

1 Block Diagram of Model: Block diagram of our proposed model is shown below.

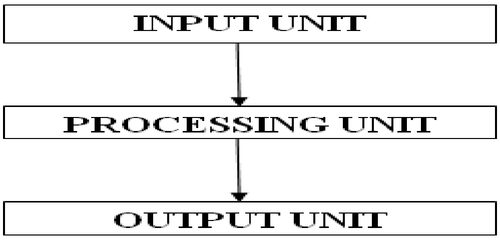

Fig 1 Block Diagram of Proposed Model

1.1 Input Unit: The input to be processed is accepted in input unit. The input is in the form of sentences or a file can also be imported. Here Hindi Keyboard is also attached to make it more user friendly. Backspace and clear control buttons are also implemented to enhance its efficiency We have given some more controls in our model. Clear is used whenever we want to enter a new word for finding reduplication word or checking reduplication word. It clears both the input and output box and the cursor automatically move to the input box. Backspace is used whenever we want to clear any character to the input box. Backspace clears the character in backward direction.

1.2 Processing Unit: Processing of the input will start by checking whether the word is entered or sentence is entered. This checking is done by applying condition if the length of the sentence is greater than or equal to two then this is considered as a word. If the size will be greater than two or more then this is considered to be as sentence. When we click on the Check Reduplication button it starts searching the input word in the database. It starts the recognition by rule based technique. If the pattern matches it will tell us the type of reduplicated word. If the match for the input word occurs, it displays the result in the output box. Otherwise it checks for the database to find the reduplicated words. If the given is sentence then it make a pair of words and matches each pair by pattern matching technique with the words present in the database. It firstly makes a pair of 1st and 2nd word and then 2nd and third and so on. When it found a full stop or comma it makes next word as the starting of new word. After that it stores all the words that are matched in a sentence with the corpus and shows us a result.

1.3 Output Unit: In the output unit model shows the result. The result comes after the processing of word or sentence. In the output box it shows the result after processing.

2 Steps of Checking Reduplication of Word and in Sentences:

The processing of the model is done by number of steps. The steps are shown below.

1. The word or sentence is entered as an input.

2. Firstly the model will check whether we have entered a word or sentence.

3. If we have entered two words it will go for rule base recognition technique. If the word will be found it will show the result else brute force algorithm comes in action to give result. If the sentence is found then it will apply pattern matching technique in which it takes first two words of a line check its reduplication if found stored the result in buffer and move to the next paired words and searches until the line is ends.

4. At last it displays the result in which it shows whether the word is reduplicated in case of only two words and in the case of sentence it will all the reduplicated words found in the line.

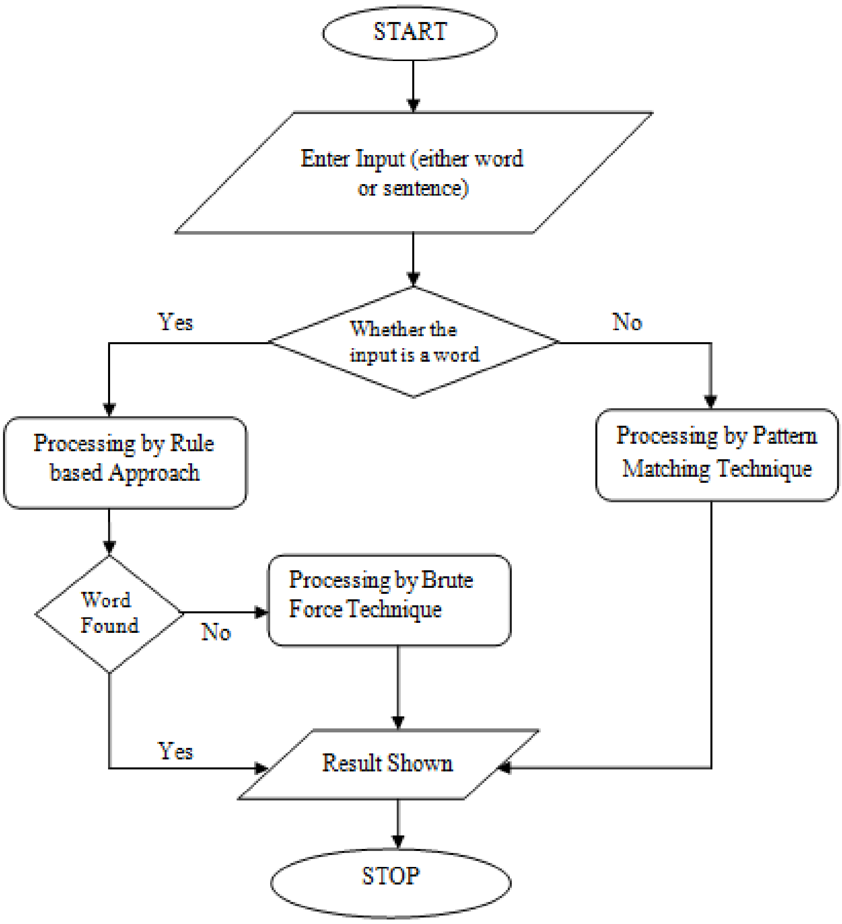
START
Enter Input (either word or sentence)
Whether the input is a word
Yes
No
Processing by Rule based Approach
Processing by Pattern Matching Technique
Word Found
No
Processing by Brute Force Technique
Yes
Result Shown
STOP

Fig 2 Steps of our proposed model

3 Corpus for Model: We have created a corpus of 1000 words that causes reduplication in Hindi. These words are collected from Hindi books, Hindi literature and Hindi newspapers (e.g. Punjab Kesari). We have worked on total 600000 words from which we have collected these reduplicated words. These words contain all type of reduplication that exists in Hindi. These words contains partial, full, echo, onomatopoeic, noun, adjective reduplication. We have created a database for these reduplicated words. We used Ms Access for storing these reduplicated words. We have distinguished all the reduplicated words with hyphen and with space to avoid conflicts when user writes some reduplicated word. We are updating our corpus when we are finding any new reduplicated word during scanning of any document. Our tool gives good response only whenever we have a big corpus.

4 Reduplication of Words: We have created a tool using .net in which we have provided a graphical user interface. We have created space for writing word or sentence. This area is called testing area. When any one enters some text in that area our model checks whether it is a word or sentence. This checks it by condition than if the word size increases than two it consider being the sentence if less than or equal to two then this is a considered as a reduplicated word. It will not count hyphen (-) as a character or word. When the reduplicated word is entered here it scans our word with the corpus. If the word matches in the corpus it will tell us that the entered word is reduplicated word. Our system will respond to all the reduplicated words either they are distinguished by hyphen (-) or with a space. Sometimes writer writes a word and differentiate it with other by hyphen (-) or sometime by space. To avoid this problem we have already created a solution for it. The approach used for finding a word in the corpus is brute force searching technique. Our stemmer uses a brute force approach. Brute force search is also called exhaustive search also known as generate and test this is a systematical approach which search for all the possible solution in the data. This approach employs a lookup table which contains a database of reduplicated words. Process is done by finding a word in the table if the match is found then the output is generated. Brute force requires immense amount of storage to create a database but it solves the problem immediately if the word found in the corpus.

5 Reduplication of Sentences: We already written that we have created space for writing word or sentence. When the sentence in found in the testing area we store first two words in a buffer and matches these words by pattern matching technique in the corpus. If these two words matches with the reduplicated word counter will be incremented. After that it creates a pair of next two words and so on until the end of the line. Line termination is checked by full stop sign. After full stop it creates a new searching by considering first character as a first and the same operation is repeated. At the end it shows the number of reduplicated words and also shows the words to be present in the paragraph or line.

EVALUATION

For creation and evaluation of any model these parameters must be satisfied.

Correctness of finding reduplicated words: Correctness of our model depends upon the word present the corpus. We have created a corpus of 1000 words and this corpus is increasing as we are testing our model with new data. We have measured a good correctness during testing our model.

Effectiveness of proposed model: Effectiveness of the model depends on the behavior of the used approach. Behavior means what model will do whenever an abnormal condition occur. Abnormal condition means whenever somebody tries to enter a word which does not exists. Our model will shows that this word is not reduplicated or in the case of sentence it shows that the text contains no reduplicated word please try with some other words. By using brute force technique we have controlled these errors. If the reduplicated word found in the database then there will be no chance of these errors to occur.

Performance of proposed model: Performance always matters when we are creating any model. The performance of our system goes high if we are able to find or catch maximum reduplicated words in a text. If we are missing number of reduplicated words then our performance goes low.

We have tested our model by ourselves and also by our friends and colleagues. They have created a table in which they have shown their results. We have made 9 tests on reduplicated words and sentences. Every test gives us different results in the term of finding reduplicated words. As we are testing our model our accuracy increases because we found new reduplicated words and stored them in the corpus continuously. We have found accuracy in case of words as well as in the case of sentences. In the case of word results are shown in table 9 and in the case of sentences results are shown in table 10.

S. No.	Number of words	Reduplicated	Found Accurate	Accuracy in

	having length 2	Words in them	words among them	%age
Test 1	400	85	71	83.53
Test 2	500	70	57	81.43
Test 3	600	60	51	85.00
Test 4	400	80	67	83.75
Test 5	300	50	43	86.00
Test 6	500	50	46	92.00
Test 7	300	30	29	96.67
Test 8	400	50	48	96.00
Test 9	200	25	24	96.00

Table 9 Evaluation in case of reduplicated words

S. No.	Number of lines	Reduplicated Words in them	Found Accurate words among them	Accuracy in %age
Test 1	20	12	10	83.33
Test 2	25	13	10	76.92
Test 3	30	14	11	78.57
Test 4	25	12	10	83.33
Test 5	40	15	11	73.33
Test 6	35	14	11	78.57
Test 7	30	13	12	92.31
Test 8	25	11	9	81.82
Test 9	35	16	13	81.25

Table 10 Evaluation in case of reduplicated words in sentences

The overall accuracy provided by our model is 88.93% in the case of words and 81.05% when we are searching reduplicated words from number of lines.

FUTURE SCOPE AND APPLICATIONS

The solution provided in this paper has demonstrated that the approach used by us is not sufficient for handling all type of reduplication in Hindi. We can increase the size of our corpus and we can also apply some different theories to find reduplication in Hindi and also make comparison with different techniques. Reduplication is used in information retrieval, finding proper words, in part of speech tagging and also useful when we are translating Hindi to English. If reduplicated words are properly handled then we can get better output from these processes.

REFERENCES

[1] Anoop Kunchukuttan and Om P. Damani " A system for Compound noun multiword expression extraction for Hindi" International conference on Natural language processing Macmillan publishers, 2008.

[2] Binna lee and Chungmin lee "A focus for contrastive reduplication prototypicality and contrastivity" pp 259-266, 2007.
[3] Chakraborty, Tanmoy and Sivaji Bandyopadhyay, "Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach" Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), pp 73–76,2010.
[4] Eva Zimmermann, Jochen Trommer," Overwriting as optimization" Natural Language & Linguistic Theory, Volume 29, Issue 2, pp 561-580,2011.
[5] Jackson Muhirwe and Trond trosterud"Finite State solutions for reduplication in Kinyarwanda language" Proceedings of the IJCNLP workshop on NLP for less privileged Languages, pp 73-80,2008.
[6] Jackson Muhirwe and Trond trosterud, "Finite State solutions for reduplication in Kinyarwanda language" Proceedings of the IJCNLP workshop on NLP for less privileged Languages, pp 73-80,2008.
[7] Jan Daciuk, Stoyan Mihov, Bruce W. Watson and Richard E. Watson, "Incremental Construction of Minimal Acyclic Finite-State Automata" Computational Linguistics, Vol. 26, No. 1, 2000.
[8]Kishorjit NongmeikapamDhiraj Laishram,Naorem Bikramjit Singh,Ngariyanbam Mayekleima Chan u and Sivaji Bandyopadhyay, "Identification of Reduplicated Multiword Expressions Using CRF" Computational Linguistics and Intelligent Text Processing pp 41-51,2011.
[9] Larry M. Hyman,"The natural history of verb stem reduplication in Bantu" Springer :Morphology, pp 177-206,2009.
[10] Le H. Phuong, Nguyen T. M. Huyen, Roussanaly A., Ho T. Vinh , "A hybrid approach to word segmentation of Hindi texts" Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Tarragona, Spain.Springer,2008.
[11] Le home phoung, Nguyen thi minh huyen, Azim roussanaly "Finite State Description of Vietnamese Reduplication" Proceedings of the 7th workshop on asian language resources ACL-IJCNLP, pp 63-69,2009.
[12] Md. Sohal rana "Reduplication in Bengali Language" Language in India ,Vol. 10, ISBN 1930-2940, pp 88-95,2010.
[13] Meladel Mistica, Avery Andrews, Iwayan arka and Timothy Baldwin "Double Double Morphology and Trouble: Looking into reduplication in Indonesian" Workshop (ALTA 2009), ed. Luiz Pizzato and Rolf Schwitter, Australasian Language Technology Association, Sydney, pp. 44-52,2009.
[14] Motomi Kajitani" Semantic properties of reduplication among the world's languages" LSO working paper in Linguistic 5, pp 93-106,2005.
[15] Omkar n koul "Reduplication in Kashmiri" Indian institute of language studies.
[16] R.M.K. Sinha and Anil Thakur "Dealing with replicative words in Hindi for machine translation to English "10th Machine Translation summit (MT Summit X), Phuket, Thailand, September 13-15, 2005.
[17] Rafiya Begum, Karan Jindal,Ashish Jain, Samar Husain and Dipti Misra sharma," Identification of conjunct verbs in hindi and its effect on parsing accuracy"CICLing'11 proceedings of the 12th international conference on computational linguistic and intelligent text processing(ACM), Volume 1, pp 29-40,2011.
[18] Reduplication-A PDE Perspective from www.google.com.

[19] Thoudam doren Singh and Sivaji Bandyopadhyay ”Web based manipuri corpus for multiword NER and reduplicated MWEs identification using SVM” proceedings of 1st workshop on south and southeast asian natural language processing, pp 35-42,2010.