International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014)

Automatic Classification of Hindi Verbs in Syntactic Perspective

Neha Dixit[1], Narayan Choudhary[2]
[1]MGAHV, Wardha, Maharashtra, India
[2]ezDI, LLC.

***Abstract*—We report of a rule based, knowledge-base driven tool to automatically classify Hindi verbs in syntactic perspective. We also report of developing the largest lexical resource for Hindi verbs along with the information on their class based on valency and some syntactic diagnostic tests as well as their morphological/inflectional type. We use this resource to develop the tool to automatically classify Hindi verbs in any given input text. In this process we also develop an algorithm based on the analysis of this lexical resource and a large Hindi corpus. In this algorithm we spell out the rules to disambiguate the homophonous/homo-gramphaemic words that become the root cause of ambiguity in this task. The tool developed has been tested against two large annotated corpora of Hindi and achieves close to 100% result on a qualitative evaluation and more than 99% on quantitative evaluation if errors from other sources are discounted.**

***Keywords*—Automatic Verb Class Identification, Hindi, NLP, Verb Class Knowledge Base, Verb Class Disambiguation**

I. Introduction

The grammatical category of verb has been categorized in several ways. Apart from the morphological categories which cover the inflectional part of the verb morphology, the syntactic categorization of verbs has been dealt with mainly in the light of transitivity. While such categorization was traditionally done to gain insight into the grammar of a language, recent interest in computational linguistics has brought some other uses of such a classification of the verbs. If the same task is done automatically through a system, it would be useful in various tasks such as grammar checking, language teaching, language generation and so on. In this paper we present a method of identifying the verb classes of any given sentence using a knowledge base developed particularly for this purpose coupled with an algorithm coming out of corpus analysis.

II. Motivation And Literature Review

Extending the traditional classification of verbs into transitive and intransitives, recent researches have classified the Hindi verbs into the categories of ergative and un-accusative [10], [7], [13], [11], [9].

Though this kind of classification is useful at the semantic level and enhances the theoretical understanding as to how the verbs behave in syntactic perspective, the same cannot be used at a formal level and cannot be included inside any automated system. Other works to handle verbs inside the theoretical computational perspectives such as HPSG, LFG and even syntactic parsing have propelled the development of linguistic resources such as PropBank [12] and FrameNet [14]. These resources cover the valency of the verbs and notes down the sub-categorization frames for each of the verbs. But these are resources that cannot be used to identify various other constraints a verb may impose on the syntactic formation of a sentence. Besides, these resources cover too few verbs as of now.

We present an alternative method of identifying the verbs and their classes based on a knowledge base and a rule based algorithm that also does disambiguation in finding out what the class of a given verb.

III. Knowledge Base Preparation

Taking a cue from the existing resources like PropBank and FrameNet, we have prepared an alternative resource with more than 3000 verbs collected from various sources such as PoS annotated Hindi corpus including the LDC corpus [1] and the ILCI corpus [3] freely available dictionaries including Hindi Shabdsagar [5] and other online dictionaries including the resource available at pustak.org website.

While the two corpora included represent the current uses of verbs in Hindi, inclusion of verbs from the dictionaries give a wide-spread coverage to most of the verbs used in Hindi.

A. Structure of Knowledge Base

For each of the lexical entry in the knowledge base, there is a corresponding morphological type and its verb class included. So, the structure of the knowledge base looks as illustrated in table I.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014)

Here ID stands for the serial number of the entry in the knowledge base, word is the actual Hindi verbal word, morph type denotes the unique morphological type of the word in the verb paradigm and verb class is the class of the verb assigned.

TABLE I
STRUCTURE OF THE KNOWLEDGE BASE

ID	Word	morph_type	verb_class
86727	पीना	inf_msg	Transitive
86728	पीनी	inf_fsg	Transitive
86729	पीने	inf_pl	Transitive
86730	पीता	impf_msg	Transitive
86731	पीती	impf_fsg	Transitive
86732	पीतीं	impf_fsgh	Transitive
86733	पीते	impf_pl	Transitive
86734	पिया	pft_msg	Transitive
86735	पियी	pft_fsg_var1	Transitive

As is evident from the table above, for each lexical entry of a verb, we have included all its morphological types based on gender, number and person (GNP) as well as tense, aspect and mood (TAM). The verb class for each verb remains the same across all its inflectional forms.

B. Deriving the Inflectional Forms for Each of the Verbs

Though we have included only 3241 entries for verb roots or lexical entries (as per lexicographic eligibility), the total number of entries in our knowledge base for verb stands at 149,518. Although actual, grammatically valid inflectional forms for a regular verb in Hindi stands at 25 [4], [6], if we also consider the spelling variations (as is a common practice in Hindi), for some verbs it reaches up to 75. These variations are mainly because of the reasons as listed below.

- People tend to use both vowel '-i:' and the semi-vowel '-yi:' as an inflectional suffix while making perfective forms of the verbs. Examples include:

Standard Form	Variation	Gloss
खिलाई/khila:yi:	खिलायी/khila:yi:	eat-PFT
गए/gəe	गये/gəye	go-PFT

- Nasalization (chandrabindu) and nasal assimilation (anuswara) are represented almost interchangeably. For example:

Standard Form	Variation	Gloss
खिलाएं/khila:ẽ	खिलाएँ/khila:yẽ	eat-IMP
गईं/gəĩ	गईँ/gəyĩ	go-IMP

- In the cases of conjunctive participles, both –कर /kər/ and –के/-ke suffix are used for some of the verbs. For example:

Standard Form	Variation	Gloss
खाकर/kha:kər	खाके/kha:ke	eat-PFT
जाकर/ʤa:kər	जाके/ʤa:ke	go-PFT

While the words collected from the annotated corpora had all the inflectional varieties with them, the verbs taken from the dictionaries are actual lexical entries with no inflectional forms included. So, we had to include those inflection forms of the verbs. We did this using semi-automatic method. For a morphologically highly productive language like Hindi, it was possible for us to set out the rules to form inflected forms of the verbs. These rules are highly productive and work on the phonological cues provided by the ending of the verb root. For example, verb roots ending in consonant follow the same pattern. Similarly, verb roots ending in particular vowels follow the same pattern. There are five types of vowels that can end a verb root in Hindi. A frequency analysis done on all of these 3200 words reveal that while –a: ending verb roots are abundant in Hindi, verb roots in ending in other vowels are only a few. Namely, the verb roots ending in –i:, -u:, -e and -o are respectively 3, 2, 3 and 11 in number. Given that these verb roots are small in number, their morphological forms were manually created and added into the knowledge base. For the rest of the two types of verb roots, rules were applied to derive the inflected forms quickly. These suffixation rules have been outlined in the table below. Also to note here is that these rules have been based on Devanagari Unicode characters which is the default script for Hindi.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014)

TABLE II
SUFFIXATION RULE FOR HINDI VERB INFLECTIONAL PARADIGM

Inflection Forms	Consonant Ending	a_ending
root	-NULL	-NULL
Infinitive MSG	-ना/-na:	-ना/-na:
Infinitive FSG	-नी/-ni:	-नी/-ni:
Infinitive PL	-ने/-ne	-ने/-ne
Imperfect MSG	-ता/-ta:	-ता/-ta:
Imperfect FSG	-ती/-ti:	-ती/-ti:
Imperfect FSGH	-तीं/-tĩ	-तीं/-tĩ
Imperfect PL	-ते/-te	-ते/-te
Perfect MSG	-ा/-a:	-या/-ja:
Perfect FSG	-ी/-i:	-यी/ई/-yi:/-i:
Perfect FSGH	-ीं/-ĩ:	-यीं/ईं/-yĩ/-ĩ:
Perfect PL	-े/-e	-ए/ये/-e/-je
Future 3MSG	-ेगा/-ega:	-एगा/येगा/-ega:/-jega:
Future 2MSGH	-िएगा/-iega:	-इएगा/इयेगा/-iega:/ijega:
Future 3FSG	-ेगी/-egi:	-एगी/येगी/-egi:/-jegi:
Future 3FSGH	-ेंगी/-ẽgi	-एंगी/येंगी/-ẽgi:/-jẽgi:
Future PL	-ेंगे/-ẽge	-एंगे/येंगे/-ẽge//-jẽge
Future 2MSGNH	-ोगे/-oge	-ओगे/-oge
Future 2FSGNH	-ोगी/ogi:	-ओगी/-ogi
Future 1MSG	-ुंगा/-ũga:	-ुंगा/-ũga:
Future 1FSG	-ुंगी/-ũgi:	-ुंगी/-ũgi
Optative 1SG	-ूं/-ũ:	-ूँ/-ũ:
Imperative PL_Honorific	-ेंं/-ẽ	-एं/-ẽ
Imperative Honorific	-िए/-ie	-इए/-ie
Imperativev 2SGNH	-ो/-o	-ओ/-o

IV. CLASSIFICATION OF VERBS

Verbs are classified into a total of 12 categories. The classification we use emanates from the practical use envisaged for such a knowledge base. While the major categories are traditional four in number (namely, intransitive, transitive, causative and double causative), we further classify the intransitives into 7 sub-classes based on some diagnostic tests which govern their syntactic function and affect or validate what constructions they allow in a sentence.

There are three diagnostic tests applied to classify the intransitive verbs. These diagnostic tests are as follows:

- Allows –ne marking with the subject
- Perfective form of the verb root can be used as an adjective
- Passivization is allowed

All the verbs included in the knowledge base have undergone this test. The tests were done subjectively by the first author of this paper, a native speaker of Hindi, and were validated by another native speaker not familiar with linguistics.

V. AUTOMATING THE TASK OF VERB CLASS IDENTIFICATION

If we use just this knowledge base to identify the verb class, we face some problems with verbs which have two meaning and occur twice in the knowledge base itself. This renders a case of ambiguity inherent in the knowledge base itself which must be resolved or the output for these words will be either incorrect or the system itself will fail to run.

If we analyse the total words in the knowledge base, we find that there are 1260 words that appear twice and one word appears thrice. If we take into account the inflectional type of the verb, the following table sums it up:

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014)

TABLE III
HOMOPHONOUS/HOMO-GRAPHAEMIC VERBS IN HINDI

Morph Type	Frequency	Examples
PFT_MSG	569	जगा/ʤəga:, उगा/uga:, उठा/uʈʰa: …
ROOT	566	जगा/ʤəga:, उगा/uga:, उठा/uʈʰa: …
PFT_PL	53	अटके/əʈke, टपके/ʈəpke, झटके/ʤʰəʈke, ...
CP_VAR1	48	अटके/əʈke, टपके/ʈəpke, झटके/ʤʰəʈke, रोके/roke ...
PFT_FSG	5	छानी/cʰa:ni:, जानी/ʤa:ni:, उसानी/usa:ni:, जीती/ʤi:tI, बरती/bərti:
INF_MSG	3	उसाना/usa:na:, छाना/cʰa:na:, जाना/ʤa:na:
INF_FSG	3	उसानी/usa:ni:, छानी/cʰani:, जानी/ʤa:ni:
INF_PL	3	उसाने/usa:ne, छाने/cʰane, जाने/ʤa:ne
PFT_FSGH	2	जीतीं/ʤi:tĩ:, बरतीं/bərtĩ
IMPF_MSG	2	जीता/ʤi:ta:, बरता/bərta
IMPF_FSG	2	जीती/ʤi:ti:, बरती/bərti:
IMPF_FSGH	2	/ʤi:tĩ:, बरतीं/bərtĩ
IMPF_PL	2	जीते/ʤi:te, बरते/bərte
IMP_2SGNH	1	बिलो/bilo

As we can see in the table above, most of the ambiguity arises between the verb roots and their perfective form. But there are also other types of ambiguity that emanate from within the knowledge base. These ambiguities have been classified into five categories as illustrated in the table below:

TABLE IV
TYPES OF AMBIGUITIES IN HINDI VERB CLASSIFICATION

No.	Ambiguity Type	Example Verbs
1	Verb Root vs. Perfective Verb	जगा/ʤəga:, उठा/uʈʰa:, ...
2	Conjunctive Participle vs. Perfective Verb	रोके/roke, कसके/kəske ...
3	Perfective Verb vs. Infinitive Verb	जाना/ʤa:na:, छाना/cʰa:na:, …
4	Perfective Verb vs. Imperfective Verb	बरता/bərta:, जीता/ʤi:ta: …
5	Verbs in multiple classes	ऐंठना/ɛ̃ʈʰna:

A. *Ambiguity: Verb Root vs. Perfective Verb*

This occurs mainly because of a derivational process used in Hindi and several other Indian languages where a valency is added by vowel lengthening. For example while the verb जग/ʤəg, an intransitive verb, means "to wake up" the verb root जगा/ʤəga, a transitive verb, means "to awaken". While जगा/ʤəga is a verb root, it is also the perfective inflectional form of the verb root जग/ʤəg and this way जगा/ʤəga gets two verb classes which need to be resolved.

While a solution to disambiguate this type of ambiguity has been implemented as described in the section below, the other types of ambiguities (described below) have to be taken care of at a word level.

B. *Ambiguity: Conjunctive Participle vs. Perfective Verb*

Another type of ambiguity which has a chance of becoming frequent if the genre of the corpus under test is of non-formal kind is that conjunctive participles can get confused with the perfective of the verbs ending on consonant -क/-k. Conjunctive participles are usually formed by adding the auxiliary verb –कर/-kər to any verb root and give a sense of perfective aspect to the verb. While the formal way of creating the conjunctive participle is to either attach -कर-/kər to the verb root itself or juxtaposing it afterwards, informally the variant -के/-ke is used. Thus we can have खाकर/kʰa:kər and खाके/kʰa:ke having the same sense and used interchangeably. Except for a few frequent use of this variant such as खाके//kʰa:ke, रोके/roke, कसके/kəske, etc. most of the time this variant is not used. And this is why we have ignored finding out a rule-based solution to disambiguate this.

C. *Ambiguity: Perfective Verb vs. Infinitive Verb*

Some verbs that end with consonant –न/-n as in छान/cʰa:n, मान/ma:n, जान/ʤa:n etc. may be sharing the same grapheme and may be homophonous with some other verb's infinitive form. Thus a verb like छाना/cʰa:na: may have two meanings, the first being the perfective of the verb root छान/cʰa:n (to filter) and another as the infinitive form of the verb root छा/cʰa: (which means "to cover the roof").

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014)

However, this type of ambiguity is also limited and count only 3 in Hindi. For this very reason, we have also ignored disambiguating this for the time being.

D. Ambiguity: Perfective Verb vs. Imperfective Verb

There are also a couple of verbs which can be interpreted as imperfective of a verb root and perfective of another. There are two verb root pairs that create this problem. The first pair is जीतना/ʤi:tna: and जीना/ʤina:, meaning respectively "to win" and "to live". The second pair is बरतना/bərətna बरना/bərna, meaning respectively "to follow" and "to choose".

E. Ambiguity: Verbs in multiple classes

While it is very common in other languages such as English that the same verb is used both as transitive and intransitive, the same is very rare in Indo-Aryan languages like Hindi. Out of all the verbs that we have analysed, we found only one verb that can be used both as transitive and intransitive. This verb is ऐंठना/ɛ̃ʈʰna: (meaning "to writhe" or "to snatch by deceit").

VI. Disambiguating Whether It Is Perfective Or Verb Root

Taking a cue from the work done on identification of verb groups in Hindi by Choudhary et.al. [4], we perform an analysis of the total verb group templates as defined here. Choudhary identifies a total of 675 templates covering all the verb groups possible in Hindi, including the compound verb constructions. As we know that for each of the verb groups found in Hindi, the class is defined by the main verb and this main verb occurs at the start of the verb groups. If we know the morphological type of the main verb (knowing whether it is verb root or a perfective form), we can identify the class of the verb with the help of the verb groups.

An analysis of the 675 verb group templates shows that there only 30 such templates where both verb root (VR) and perfective verb (VR_pft) can stay. This means we basically have to find out the disambiguation rules for only these verb groups. The rest are taken care of automatically as proper, grammatical structure of Hindi would not allow them. These verb group templates are noted in the table below:

TABLE V
PROBABLE VERB GROUP TEMPLATES CREATING AMBIGUITY

Template with First word as either VR or VR_pft	Ambiguity
VR_pft+ja_fut	Yes
VR_pft+ja_impf	Yes
VR_pft+ja_impf+prs_aux	Yes
VR_pft+ja_inf+cahiye	Yes
VR_pft+ja_inf+cahiye+pst_aux	Yes
VR_pft+ja_opt	Yes
VR_pft+ja_pft	Yes
VR_pft+ja_pft+ho_impf+prs_aux	Yes
VR_pft+ja_pft+prs_aux	Yes
VR_pft+ja_pft+pst_aux	Yes
VR_pft+ja+rah_pft	Yes
VR_pft+ja+rah_pft+ho_fut	Yes
VR_pft+ja+rah_pft+prs_aux	Yes
VR_pft+ja+rah_pft+pst_aux	Yes
VR_pft+par_pft+prs_aux	Yes
VR_pft+par_pft+pst_aux	Yes
VR_pft+rah_pft	Yes
VR_pft+VINF	Yes
VR_pft+VINF_imp	Yes
VR_pft+ho_fut	No
VR_pft+ho_impf	No
VR_pft+ho_opt	No
VR_pft+ho_pft	No
VR_pft+kar_fut	No
VR_pft+kar_imp	No
VR_pft+kar_impf	No
VR_pft+rah_fut	No
VR_pft+rah_imp	No
VR_pft+rah_opt	No
VR_pft+rakh_pft	No

If we further analyse these templates we find that for some of these templates, ambiguity will never arise. This has been marked with "No" in the table above. Thus we get a total of 19 templates where there is a chance of ambiguity coming up.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014)

If we look at the first auxiliary verbs of these 19 templates, we find that they actually cannot allow a verb root to occur as main verb. The first auxiliary verbs for these templates are as follows:

TABLE VI
FIRST AUXILIARY VERBS OF AMBIGUOUS TEMPLATES AND THEIR VALUES

Variable	Inflectional Type	Values
ja	Verb Root	जा/ʤa:
ja_fut	Future forms of the verb जा/ʤa:	जाएगा/ʤa:ega:, जाएगी/ʤa:egi:, जाएंगे/ʤa:ẽge, जाउंगा/ʤa:ũga:, जाओगे//ʤa:oge
ja_impf	Imperfect Forms of the verb जा/ʤa:	जाता/ʤa:ta:, जाती/ʤa:ti:, जाते/ʤa:te, जातीं/ʤa:tĩ
ja_inf	Infinitive Forms of the verb जा/ʤa:	जाना/ʤa:na:, जानी/ʤa:ni:, जाने/ʤa:ne:
ja_opt	Optative Forms of the verb जा/ʤa:	जाए/ʤa:e, जाय/ʤa:j
ja_pft	Perfective Forms of the verb जा/ʤa:	गया/gəja:, गई/gəji:, गए/gəje, गईं/gəjĩ:
par_pft	Perfective Forms of the verb पड़/pəɖ	पड़ा/pəɖa:, पड़ी/pəɖi:, पड़े/pəɖe, पड़ीं/pəɖĩ:
rah_pft	Perfective Forms of the verb रह/rəh	रहा/rəha:, रही/rəhi:, रहे/rəhe, रहीं/rəhĩ:
VINF	Infinitive form of any verb	VR+ना/-na:, VR+ने/-ne, VR+नी/-ni
VINF_imp	Infinitive Form of a verb used in the Imperative Mood	VINF_imp

Now, if we closely analyse these auxiliary verbs in the given templates, we find that these auxiliaries actually do not allow perfective verbs to occur as their main verbs. This conclusion is based on a corpus study done on the EMILLE [8] corpus and the GyanNidhi corpus as well as some exact searches done on a prominent search engine.

For example, in the phrase जगा जाएगी/ʤəga: ʤa:egi:, the main verb जगा/ʤəga: or for that matter any other verb can never be inflected for the perfective aspect. Similar is the case with the place of other main verbs in the templates having ambiguity (noted with "yes" in table V above).

VII. ALGORITHM AND TOOL FOR VERB CLASS FINDER

Following the processes outlined above so far, a tool was developed to find the verb class for all the words marked as main verb in a parts of speech annotated input text. A step by step flow for the algorithm has been illustrated in the following flowchart:

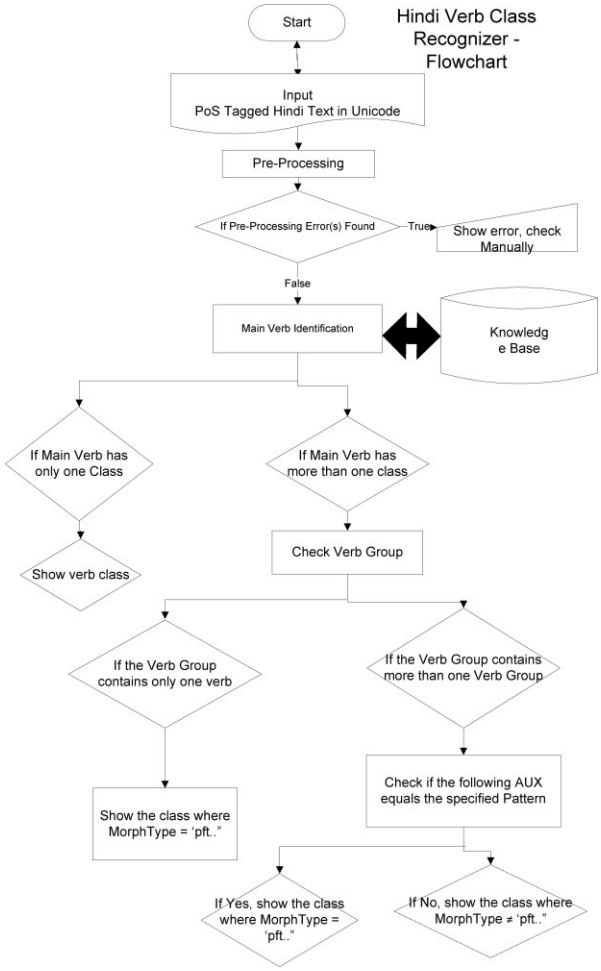

Figure 1: Hindi Verb Class Recognizer Flowchart

Based on this algorithm, an application was written in first in Java and then developed into a web application which can be viewed and tested at the following URL:

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014)

http://app.langlex.com/vclassfinder/

The application is run via an instance of Apache Tomcat server hosted via a proxy on an Apache HTTPD server.

VIII. Evaluation

Evaluation was done in two phases. The results obtained in the first phase was analysed for errors and the analysis was used to improve the tool itself.

A. First Phase Evaluation: the LDC Corpus

Output was taken for the LDC corpus which has a total of 4,839 sentences annotated using the IL-PoST tagset [2]. Evaluation was done two fold. One was quantitative evaluation which showed how good was the knowledge base and whether we missed any word. It also served as an evaluation of the corpus undergoing the test. The other was a qualitative evaluation where we evaluated the class assigned by the tool for each of the verbs identified. The results have been shown in the table below.

TABLE VII
Quantitative Evaluation Results for LDC Corpus

Total Main Verbs Found	8386
Total Main Verbs Classified	8236
Unclassified Main Verbs	145
Error Percentage	145*100/8386 = 1.729072263

An error analysis done on the errors found revealed that errors occurred due to four different reasons as illustrated in the table below:

TABLE VIII
Error Analysis of the LDC Corpus

Error Type	Error Count	Part of Error
Spelling Mistake	88	60.68966
Annotation Error	27	18.62069
Echo-Word	22	15.17241
Pre-processing Error	8	5.517241

As we can see here, the LDC corpus had a lot of spelling mistakes mainly because the corpus is made of text taken from the web and had a different kind of encoding which was later converted to Unicode. This encoding error has generated a lot of errors. Another type of frequent error was due to annotation. In this case, the annotation for one of the verbs was not correct i.e. either the main verb or the auxiliary verb was not tagged properly.

Echo-words have not been accounted for in the tool we have developed. As echo-words are very frequent in most of the South Asian languages, this should be given a dedicated attention and a framework needs to be developed to handle various types of echo-word formations appearing in the content class words such as nouns, verbs, adjectives and adverbs. A few pre-processing errors were also encountered as our tool does not have any pre-processing module yet.

Qualitative evaluation was done manually by engaging two linguists and native speakers of Hindi. There was no need of evaluating all the sentences as we know that ambiguities will arise only in a few sentences (1261 words from the knowledge base as mentioned above). So, we selected all the words given a class by the tool which may possibly have an ambiguity. Thus, we got a total of 70 unique verbs that appeared in the corpus and could possibly have more than one class in a given context. These 70 unique verbs were spread in a total of 330 sentences. The output of these sentences was evaluated by two linguists.

No errors in class assignment were found as all the ambiguities encountered were of the first type i.e. they belonged to the category of perfective vs. verb root which have been covered in the algorithm. Thus, with the LDC corpus, we are getting 100% accuracy at the qualitative level of evaluation.

B. Second Phase Evaluation: ILCI Corpus

A similar evaluation was done by running the tool against the ILCI corpus [4]. Resourced from from the TDIL[1] for this research, this corpus has a total of 50,000 sentences and more than 800,000 words belonging to two domains of health and tourism.

The table below shows the results obtained on the ILCI corpus.

TABLE IX
Quantitative Evaluation Results for ILCI Corpus

Total Main Verbs in ILCI	87801
Number of Main Verbs Classified in ILCI	82232
Number of Main Verbs Unclassified in ILCI	5569
Accuracy on the ILCI Corpus	93.66
Error Percentage	6.34

[1] http://tdil-dc.in

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014)

As we can see here, the error rate is high and requires attention. For this corpus also we did the error analysis following the same method done for the ILCI corpus. This revealed several types of phenomenon contributing this high error rate which has been illustrated in the table below:

TABLE X
ERROR ANALYSIS OF THE LDC CORPUS

Error Type	Tourism	Health	Overall	% of Errors
Annotation Error	3389	1711	5100	91.58
Echo word	99	137	236	4.24
Spelling Error	29	93	122	2.19
KB Error	74	8	82	1.47
Echo-word + Ann Error	5	7	12	0.21
Pre-Processing Error	3	9	12	0.21
Spelling Error + Annotation Error	2	3	5	0.1

As we can see from the table above, this application shows more than 93% accuracy on the ILCI corpus. Although this is not a very good result, if we look into the reasons of these errors, it becomes evident most of these errors are not because of the knowledge base or because of the disambiguation module where the novelty of this research lies. These errors cannot be attributed to the system developed and presented here. Only the errors coming due to the knowledge base and can be attributed to the system. Error count generated due to the failure of the knowledge base is 82. If we count accuracy by ignoring other others, the accuracy soars to 99.9%.

IX. CONCLUSION

This paper sums up the processes undertaken to develop a tool that identifies the class of a given verb in Hindi in the syntactic perspective. Using a detailed classification criteria based on important diagnostic tests, a large lexical resource of Hindi verbs have been developed over the last three years. The knowledge base contains most of the verbs along with their possible spelling variations and all their inflectional forms. This knowledge base also claims to be the biggest collection of Hindi verbs ever placed in a single resource. Using this knowledge base, we develop a rule based tool to disambiguate the actual class for the verbs occurring twice or thrice in the knowledge base.

As this is a rule based tool, we also report an accuracy of more than 99% quantitatively and 100% using two specific corpora qualitatively.

REFERENCES

[1] Bali, Kalika et.al. 2010. Indian Language Part-of-Speech Tagset: Hindi. Linguistic Data Consortium, Philadelphia.

[2] Baskaran, Sankaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha and K.V. Subbarao. 2008. A Common Parts-of-Speech Tagset Framework for Indian Languages. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Daniel Tapias (Eds.) Proceedings of the Sixth Inter-national Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

[3] Choudhary, Narayan and Girish Nath Jha. 2011. Creating Multilingual Parallel Corpora in Indian Languages. In: Proceedings of 5th Language Technology Conference, Fundacja Uniwersytety im. A. Mickiewicza, Poznan. Pp. 85-89 ISBN No.: 978-83-932640-1-8

[4] Choudhary, Narayan, Girish Nath Jha, Pramod Pandey. 2011. A Rule based Method for the Identification of TAM features in a PoS Tagged Corpus. In: Proceedings of 5th Language Technology Conference, Fundacja Uniwersytety im. A. Mickiewicza, Poznan. Pp. 334-338 ISBN No.: 978-83-932640-1-8

[5] Dasa, Syamasundara. Hindi sabdasagara. Navina samskarana. Kasi: Nagari Pracarini Sabha, 1965-1975.

[6] Kachru, Yamuna. 2006. Hindi. John Benjamins: Amsterdam/Philadelphia.

[7] Levin, Beth & M. Rappaport Hovav. 1995. Unaccusativity at the Syntax-Semantics Interface. Cambridge. Mass: .MIT Press.

[8] McEnery, A. M., P. Baker, R. Gaizauskas, and H. Cunningham. 2000. EMILLE: Building a Corpus of South Asian Languages. In: Vivek: A Quarterly in Artiificial Intelligence. 13(3): pp: 23–32.

[9] Nitesh Surtani, Khushboo Jha and Soma Paul. 2011. Issues with the Unergative/Unaccusative Classification of the Intransitive Verbs . International Confer-ence on Asian Language Processing (IALP), Penang, Malaysia

[10] Perlmutter, D. 1978. Impersonal Passives and the Unaccusative Hypothesis. In Proceedings of the Berkley Linguistics Society, 4. Berkley Linguistics Society, University of California, Berkley.

[11] R Bhatt and D. Embick, Causative derivations in Hindi, 2003, http://people.umass.edu/bhatt/papers/bhatt-embick-caus.pdf

[12] Rajesh Bhatt; Bhuvana Narasimhan; Martha Palmer; Owen Rambow; Dipti Sharma; Fei Xia, A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu, In the Proceedings of the Third Linguistic Annotation Workshop, held in conjunction with ACL-IJCNLP 2009, Singapore, August, 2009.

[13] Richa. 2008. Unaccusativity, Unergativity and the Causative Alternation in Hindi: A Minimalist Analysis. PhD Thesis. JNU, New Delhi.

[14] Ruppenhofer, J. et. al. 2006. FrameNet II: Extended Theory and Practice. http://framenet.icsi.berkeley.edu/book/book.pdf