Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora

Amitabha Mukerjee, Ankit Soni, and
Dept. of Computer Science and Engg
Indian Institute of Technology Kanpur
Kanpur -208016, India
amit@iitk.ac.in,
ankit@iitk.ac.in

Achla M Raina
Dept. of Humanities and Social Sciences
Indian Institute of Technology Kanpur
Kanpur -208016, India
achla@iitk.ac.in

Abstract

Complex Predicates or CPs are multi-word complexes functioning as single verbal units. CPs are particularly pervasive in Hindi and other Indo-Aryan languages, but an usage account driven by corpus-based identification of these constructs has not been possible since single-language systems based on rules and statistical approaches require reliable tools (POS taggers, parsers, etc.) that are unavailable for Hindi. This paper highlights the development of first such database based on the simple idea of projecting POS tags across an English-Hindi parallel corpus. The CP types considered include adjective-verb (AV), noun-verb (NV), adverb-verb (Adv-V), and verb-verb (VV) composites. CPs are hypothesized where a verb in English is projected onto a multi-word sequence in Hindi. While this process misses some CPs, those that are detected appear to be more reliable (83% precision, 46% recall). The resulting database lists usage instances of 1439 CPs in 4400 sentences.

1 Introduction

A "pain in the neck" (Sag et al., 2002) for NLP in languages of the Indo-Aryan family (e.g. Hindi-Urdu, Bangla and Kashmiri) is the fact that most verbs (nearly half of all instances in Hindi) occur as *complex predicates* - multi-word complexes which function as a single verbal unit in terms of argument and event structure (Hook, 1993; Butt and Geuder, 2003; Raina and Mukerjee, 2005). Moreover, most of these languages being resource-poor, even a proper corpus-based characterization of such CPs has remained an elusive goal.

In this paper we construct the first corpus-based lexicon of CPs in Hindi based on projecting POS tags across parallel English-Hindi corpora. While such approaches sometimes leave out some CPs, the ones that are identified are seen to be quite robust. As a result, this appears to be a good first approach for identifying the majority of CPs along with usage data. Moreover, since the language specific input in the procedure is minimal, it can be easily extended to other languages with similar multi word expressions.

2 Complex Predicates

CPs are characterized by a predicate or host - typically a noun (N), adjective (A), verb (V), or adverb (Adv) - followed by a *light verb* (LV), a grammaticalized version of a main verb, which contributes little telic significance to the composite predicate. As an example, the English verb "describe" may be rendered in Hindi as the Noun-Verb complex 'वर्णन + कर', *varNan kar*, "description + do". Analysis based on a non-CP lexicon might assign the verbal head as *kar* (do), whereas functional aspects such as the argument structure are determined by the noun host *varNan* "description". An example of a V-V CP may

Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, pages 28–35,
Sydney, July 2006.

be ‘कर + दे’, kar *de* "do+give", where the light verb *de* “give” imposes a completive aspect on the action kar “do”.

Identifying such constructs is a significant hurdle for NLP tasks ranging from phrasal parsing (Ray et al., 2003, Shrivastava et al., 2005), translation (where each complex may be treated as a lexical unit in the target language), predicate-argument analysis, to semantic delineation. In addition to the computational aspects, a mere listing of all CPs occurring in the corpus would provide an important resource for tasks such as constructing WordNets (Narayan et al.,2002) and linguistic analysis of CPs (Butt and Geuder, 2003).

Rule-based approaches to identifying CPs are not very effective since there do not seem to be any clear set of rules that can be used to distinguish CPs from non-CP constructs (contrast, for example, the composite CP ‘अनुमति दे’ *anumati de* "permission+give" with the non-composite N-V structure ‘किताब दे’ *kitaab de* "give the book"). Even where such rules do exist, they depend on semantic properties such as the fact that book is a physical object which can be given in the physical sense (Raina and Mukerjee, 2005). However, in the translated form, the former may show up as a verb, whereas the latter invariably will be a N+V, so the tag projection would rule out the latter as a CP.

Here we adopt a parallel corpus-based approach to creating a database of complex predicates in Hindi. The procedure can potentially be duplicated to most Indo-Aryan languages. The motivation is that a CP may be translated as a direct verb in other languages, and POS Projection across Parallel Corpora then project a tag of Verb for this expression in the source language. Additional linguistic constraints are used to determine if the multi-word cluster qualifies as a CP. These include a check list of LVs that can occur with A, N, V and Adv constituents of a multi word predicate.

Let us consider some examples from the CP lexicon constructed from the EMILLE parallel corpus (McEnery et al., 2000) of 200,000 words, collected from leaflets prepared by the UK government for immigrants. Examples of these different complexes may be:

(1) N+V: वर्णन + कर *varNan kar*
“description + do”:

पैकेज या प्रस्तुत इश्तेहार में जैसे
paikej yaa prastut ishtehaar mein jaise
package or present advertisement in as

वर्णन किया गया हो, ठीक वैसा
varNan kiyaa gayaa ho ThIk vaisaa
description do-past go-past be-pres exact same

ही होगा
hii hogaa
emph be-fut

“It will be exactly as described on the package or the display advertisement.”

(2) A+V: उपलब्ध है *upalabdh hai*
“available+ be”:

सहायता समीप ही उपलब्ध है।
Sahaytaa samiip hii upalabdh hai
Help near emph available be-pres

“Help is available nearby.”

(3) V+V : सोच ले *soch le* “think+take”:

पहले हर पहलू के बारे में अच्छी तरह
Pahle har pehluu ke baare-mein achchhi tarah
First every aspect-poss about good way

सोच लीजिए।
soch liijiye
think take-imp-hon

“Think it through first.”

(4) Adv+V *vaapas paa* “return+obtain”

आप सामान बदलने में अपने पूरे पैसे
Aap saamaan badalne mein apne puure paise
You goods exchange-nom in your all money

वापस पाने का अधिकार खो देते हैं।
vaapas paane kaa adhikar kho dete hai
return obtain-nom of right lose give be-pres

“You loose your right to get your full money back in exchanging the goods. “

Of the four classes cited above, the NV and AV classes are the most productive. The AdvV class is highly restricted, confined to a few adverbs. The VV class is highly selective for its constituents, apparently driven by semantic considerations.

Identifying CPs in text is crucial to processing since it serves as a clausal head, and other elements in the phrase are licensed by the complex as a whole and not by the verbal head. The semantic import of the host-verb complex varies along a composability continuum, at one end of which we have purely idiomatic CPs, while at the other end, the CPs may be recoverable from its constituents. For example, 'व्यवहार+कर', vyavhaar *kar*, "behave+do" has a sense of "use,treat" in English, reflecting clearly an idiomatic usage.

Detecting CPs is made difficult by the differing degrees of productivity for different classes of open-class host, which reflects the applicability of unrestricted rules. Also, verbs participating in CPs are very selective; e.g. in NV and AV CPs the verb is typically restricted to *ho, kar* and the like, whereas in VV constructs ho reflects auxiliary usage, but a different set of verbs appear. The open class word (host) tends to be uninflected, and only the light verb (LV) carries tense, agreement and aspect markers. Even the host V participating in a VV CP is always uninflected. As an instance of the difficulty in detecting CPs, consider the so called permissive CP (Hook, 1993; Butt and Geuder, 2003), as in the *karne+de* "do-nom +give" example here, where the host verb appears to be inflected:

(5) Raam ne sitaa ko kaam karne diyaa
Ram-erg sita-acc work do-nom give-past
"Ram let Sita do the work"

However, this does not actually reflect CP usage, and is better parsed as:

(6) [$_S$ [$_{NP}$ raam ne] [$_{VP}$ [$_{NP}$ sitaa ko] [$_{VP}$ kaam karne] [$_V$ diyaa] $_{VP}$] $_S$]

Another challenge for CP identification is that the constituents may be separated – sometimes quite widely.

3 CPs from Parallel Projection

Identifying MWEs from corpora is clearly an area of increasing research emphasis. For resource-rich languages, one may use a parse tree and look for mutual information statistics in head-complement collocations, and also compare it with other "similar" collocations to determine if something is unusual about a given construct (Lin, 1999). As of now however, even POS-tagging remains a challenge for languages such as Hindi, thereby making it necessary to seek alternate methods.

Parallel corpus based approaches to inducing monolingual part-of-speech taggers, base noun-phrase bracketers, named-entity taggers and morphological analyzers for French, Chinese and other languages have shown quite promising results (Yarowsky et al., 2001). These approaches use minimal linguistic input and have been increasingly effective with the growth in the availability of large parallel corpuses. The algorithm essentially attempts to word-align the target language sentences with the source language sentences and then use a probabilistic model try to project the linguistic information from the source language. Since these are statistical algorithms, the accuracy of results depends on the size of the corpus used.

In our approach, we first use a similar approach to word-align an English-Hindi parallel corpus. The English sentences are tagged and the tags are projected to Hindi sentences. We observe that words which are tagged as verbs by projection and have POS tag as N, A, Adv or V in the Hindi lexicon, and are followed by an LV, are usually CPs.

Clearly the CP detection is limited to those instances where a CP in the target language is translated as a single verb in English. For example, a phrase such as जवाब दे, *jawaab de,* "answer give", may be rendered in English either as the verb "answer" or as the English CP "give answer". In the latter case (an example appearing quite frequently in this corpus), the correct POS projection would label jawaab as [N answer], thus failing to detect the CP. While this may not be significant in certain tasks (e.g. translation), it may be relevant in others (e.g. semantic processing).

Furthermore, the POS tagging process is inherently biased towards projecting tags for frequently encountered constituents first, and this may lead to some constituents in certain CPs being flagged with their normal POS tags, resulting in missed CPs. However, this does not result in false positives, since non-CP constructs often fail on other criteria (e.g. list of LVs).

For reasons discussed above, many CPs are not identifiable through parallel corpus methods. Some examples include 'अधिकार होते', 'पैदा करने', 'हानि होती'. Our database is therefore correspondingly thin for these types of CPs.

With VV CPs, it is difficult to distinguish between CPs and other related structures such as the passive construct or serial verbs. These are illustrated below.

(7) Passive

ऐसा	भी	हो	सकता	है	कि	क्रेडिट	नोट
Aisa	*bhii*	*ho*	*saktaa*	*hai*	*ki*	*credit*	*note*
It	emph	be	can	aux	that	credit	note

सिर्फ	कुछ	ही	दिनों	तक	काम	में
siraf	*kuch*	*hii*	*dino*	*tak*	*kaam*	*me*
only	few	emph	days	for	use	in

लाया	जा	सकता	हो।
laaya	*jaa*	*sakta*	*ho*
bring	go	can	be

"It is quite possible that the credit note can be put to use only for a few days."

(8) Serial verb

वह	लडका	मुझे	अपनी	किताब	दे	गया।
voh	*laDkaa*	*mujhe*	*apni*	*kitaab*	*de*	*gayaa*
That	boy	me	own	book	give	go-past

"That boy gave me his book and went away."

It appears that passive can be reliably ruled out using the root verb criterion for VVs, since the main verb in passive is always in an inflected form. No comparable formal criterion exists for the serial verb, where also the POS tagger will identify both constituents as verbs. However, these verbs are relatively rare compared to CPs.

4 Hindi-English POS Projection

4.1 Data Resources and Preprocessing

We used the EMILLE[1] corpus Hindi-English parallel corpus, with approximately 200,000 words in non-sentenced aligned translations in Unicode 16 format (McEnery et al., 2000). The texts consist of different types of information leaflets originally in English, along with translations in Hindi, Bangla, Gujarati and a number of South Asian languages. Closer analysis of the corpus reveals that the corpus is not completely sentence aligned and also that the translations are not very correct in many cases. Hindi versions of the manuals tend to be more verbose than their English translations.

For the word alignment algorithm we needed a sentence aligned corpus but due to the small size of the parallel corpus, the standard sentence alignment systems did not give very high accuracy levels. Therefore, the whole data was manually sentence aligned to produce a sentence aligned parallel corpus of about nine thousand sentences and 140 thousand words which is used in this work.

4.2 Word alignment

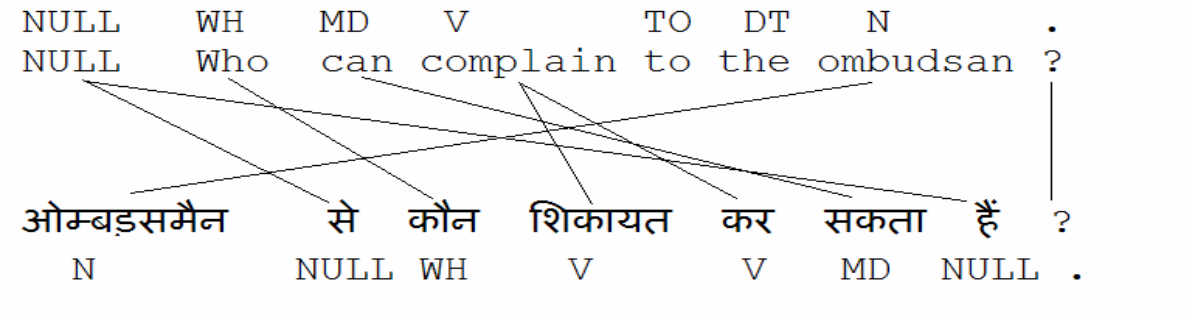
We have used IBM models proposed by Brown (Brown et al., 1993) for word aligning the parallel corpus. The IBM models have been widely used in statistical machine translation. Given a Hindi sentence h, we seek the English sentence e that maximizes P(e | h); the "most likely" translation.

Now P (e | h) = P (e) * P (h | e) / P (h)
argmax-e P(e | h) = argmax-e P(e) * P(h | e).

P (e) is modeled by the N-gram model .We are interested in P (h | e). We used the Giza++ tool kit (Och and Ney, 2000), based on the Expectation Maximization (EM) algorithm, to calculate these probability measures. At the end of this step, we have a word-to-word mapping between the English and Hindi sentences. A "NULL" is used in the English sentences to account for the unaligned Hindi words from the corresponding Hindi sentence.

[1] http://bowland-files.lancs.ac.uk/corplang/emille/

Figure 1. Example of projection of POS tags from English to Hindi. Here the phrase "shikaayat kar" is projected from the English "complain" and is tagged as V+V. Since shikaayat is a N in the Hindi lexicon, this phrase is identified as an CP of N+V type.



4.3 Tagging English Sentences

The English sentences are POS-tagged using the Brill Tagger (Brill, 1994), a rule based tagger which uses more or less the same tags as the Penn Treebank project (Marcus, 1994). Since for our purposes, we did not need a very detailed subcategorization of the tag set for Hindi, the English tag set was reduced by merging the subcategorization tags of a few categories. Thus all noun distinctions in the Pen Treebank tagset based on number, person etc were merged in our treatment of the Noun class. Similarly in the case of verbs, we merged distinctions based on tense, person, aspect and participles etc. Subclasses of adverbs and case forms of pronouns were also merged. Rest of the POS categories were retained. The "NULL" word in the English sentences, used for unaligned Hindi words in the parallel corpus, was given a "NULL" tag.

4.4 Projection of Tags to Hindi

The reduced English tags were projected to Hindi words based on the word alignments obtained earlier. A sample alignment and tagged projection is shown in Figure 1. As the figure shows, postpositional markers, which are relatively more frequent in Hindi are mapped to the "NULL" word in the English sentence.

Since the amount of training data is very small, the statistical word alignment algorithm is not adequate enough to align all words correctly. To overcome this weakness, we apply some filtering conditions to remove alignment errors, especially in smaller sentences. This filtering is based on two parameters: a) Fertility count (r_f), which is defined as the number of Hindi words an English word maps to, and b) Acceptance level (k), defined as the number of words acceptable in a sentence with fertility count greater than equal to r_f. These two parameters are selected to minimize errors in the groundtruth sample-set, and the resulting filtering heuristics used are presented in Table 1.

Table-1. Filtering Criteria

	Sentence Length	Fertility Count(r_f)	Acceptance Level(k)
1.	1-5	2	1
2.	5-10	3	2
3.	10-15	3	3
4.	15-20	4	3
5.	20-25	4	3
6.	25-35	4	3
7.	35+	4	3

4.5 Identification of CP's

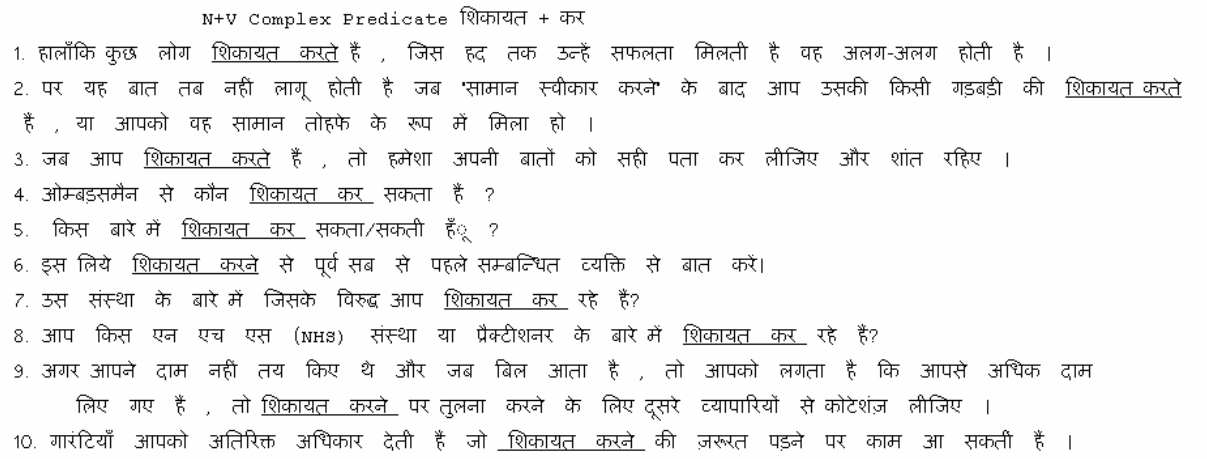
After the filtering is done we observe that the CP's are usually translated as a direct verb in English. So if the projected tag of a Hindi word is Verb and the normal POS tag of the word in the Hindi dictionary is N, A, V or Adv and the word is followed by one of the members from the LV set, then we classify the multi word expression as N+V, A+V, V+V, or Adv+V CP respectively.

4.6 Fragments of the CP Lexicon

A sample fragment of the CP lexicon is shown in Figure-2. The whole corpus is available online[2]. Since we do not have a very comprehensive Hindi dictionary we are not able to classify many CP's that are identified in their respective class. On a test with 4400 sentences we identified a total of 1439 CPs

[2] The lexicon is available online at http://www.cse.iitk.ac.in/users/language/CP-database.htm

Figure 2. Example of the CP lexicon for “shikaayat kr”



with the following distribution: N+V: 788, A+V: 107, Adv+V: 18 and V+V: 526.

4.7 Errors in CP identification

CP identification in the test data set involved certain ground truth decisions such as excluding verbal composites with regular auxiliary verb है, *hai* corresponding to the English finite verb ‘be’ and the progressive ‘रहा’ raha ‘-ing (progressive)’. CPs with idiomatic usage were included, and so were the CPs with a passive verb, although the latter were not counted in computational scores. The testing was done on a small set of about 120 groundtruth sentences in which the CP’s were carefully identified manually. We get a precision of about 82.5% and a recall of 40% with our CP finding algorithm. If the idiomatic CPs is not considered the recall goes upto 46%.

Several types of errors are observed in the corpus-derived results. A False Negative (missed CP) error arising due to the English complex predicate is shown in Figure 3. A number of False Positives arise due to inadequacy in the Hindi dictionary – the online dictionary of Hindi we used was missing many lexemes. A further problem is homography – e.g. the word *kii* (do-past) appears both as an possessive marker, as well as the past-tense form for the verb *kara* (do), occurring frequently (with jaa, go) in adjectival clause constructions. This has been mis-tagged in about one in ten instances (approx 0.2% cases), with hosts such as *shikaayat* (complaint), *baat* (talk), *dekhvaal* (looking-after), *madad* (help) etc. Similarly, the word *un* can appear as a noun (wool) or a pronoun (he). Furthermore, while considerable care was taken to manually sentence align the parallel corpus, a number of typos and other problems remain, some of them show up as false positives.

4.8 Discontinuous CP identification

In the results above, we have made no attempt to identify discontinuous CPs, i.e., instances where other phonological material intervenes between the constituents of a CP, As an example, consider

(9) जाँच हो, jaanch ho, “inspection-be”

अगर कार की जाँच पहले ही हो
agar kaar kii jaanch pahale hii ho
if car poss inspection earlier emph happen

चुकी है , तो रिपोर्ट माँगिए ।
chuki hai to report mangiye
comp. be-present then report ask-imp-hon

“If the car has already been inspected please ask to see the report.”

These separated multi-word expressions constitute some of the most difficult problems for any language – for example, one may compare these with English phrasal verbs like “give up”, which can sometimes occur in discontinuity. However, owing to the relatively free word order in Hindi, the discontinuous CPs in Hindi are separated by a variety of structures ranging from simple emphatic or focal particles and negation markers to clausal

Figure 3. Here the projection process fails to detect the CP "shikaayat karna" since the English translation is also CP "make complaint". Improvements in MWE detection in English can possibly help reduce such errors.

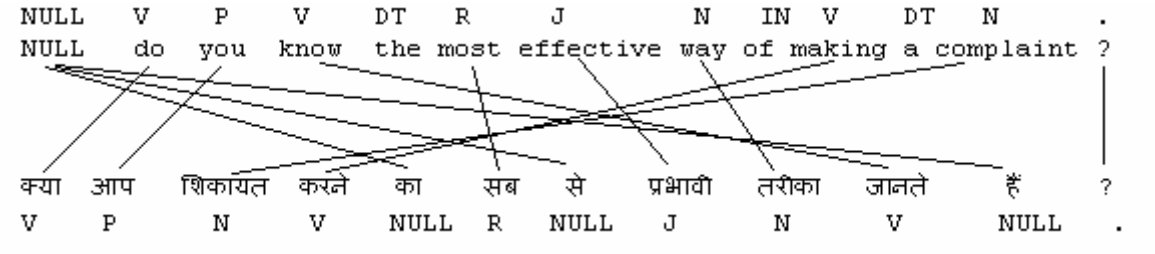


Figure 4. A verb in the source language, "inspected" projects to *jaanch* (inspection)+ *ho* (be) + *chukaa hai* (aux), although they are separated by the phrase *pahale-bhi* (already). Thus, using source and target languages together. the parallel projection method may have the potential for discovering discontinuous CPs as well.

constituents. How these structures are to be encoded in a computational lexicon is a complex matter that takes us beyond CP identification (Villavicencio et al. 2004). But while rule-based identification of such constructs is problematic, we feel that POS-tag projection holds considerable promise in this direction.

In the algorithm above we have only considered the target language (Hindi) tags after the parallel tagging is completed. If in addition, we also consider the source language tag and its radiation the CP probabilities may be redefined in a manner that helps capture some discontinuous CPs as well. Thus, if English "complain" radiates to *shikaayat* and *kara*, the inherent CP can be detected even in the presence of an intermediate phrase. An example from the POS-tagged data exhibiting discontinous CP detection is presented in Figure 4.

5 Conclusion

In this work we have presented a preliminary approach to a corpus-based lexicon of CPs in Hindi based on projecting POS tags across parallel English-Hindi corpora. Since the approach involves minimal linguistic analysis, it is easily extendable to other languages which exhibit similar CP constructs, provided the availability of a POS lexicon.

Clearly, a number of problems will remain with any such approach. The limitiations of the parallel POS tagging is that certain kinds of maps may never be found (as in parallel CPs in source and target languages). On the other hand, some of our accuracies, we feel, would improve considerably given a larger parallel corpus and more refined use of a Hindi lexicon.

In addition to the handling of discontinuous CPs hinted at above, another aspect that we would like to consider next is to tune some of the parameters of the parallel tagging algorithm, such as specifically tuning the distortion and fertility probabilities in situations (e.g. English verbs) that are likely to manifest CPs in Hindi.

We feel that beyond the usefulness of this initial approach, the database of CPs constructed in this work may in itself be an important linguistic resource for Hindi. Furthermore, the approach can possibly be used to detect MWEs that radiate to a single lexical structure in another language, e.g. phrasal verbs in English.

Acknowledgements We acknowledge a comment from an anonymous reviewer regarding discontinuous CPs which led us to investigate them (Figure 4 above). However, it was not possible to report this important exception for the entire database.

References

Eric Brill.1994. *Some advances in transformation-based part of speech tagging*, National Conference on Artificial Intelligence,p 722-727.

Peter F. Brown, Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L.. 1993. Computational Linguistics 19(2), 263-311.

Miriam Butt and Wilhelm Geuder. 2003. *Light Verbs in Urdu and Grammaticalization.,* Trends in Linguistics Studies and Monographs, Vol 143, p295-350.

Peter E. Hook. 1993. *Aspectogenesis and the Compound Verb in Indo-Aryan.* Complex Predicates in South Asian Languages.

Dekang Lin.1999. *Automatic Identification of Non-compositional Phrases*, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 317--324.

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1994. *Building a large annotated corpus of English: the Penn Treebank,* Computational Linguistics 19(2), 313–330.

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

D Narayan, D Chakrabarty, P Pande and P Bhattacharyya. 2002. *Experiences in Building the Indo Wordnet: A Wordnet for Hindi* International Conference on Global WordNet

Franz Josef Och and Hermann Ney. 2000. *Improved statistical alignment models*, in ACL00 p 440–447.

Achla M. Raina and Amitabha Mukerjee. 2005. *Complex predicates in the generative lexicon*, Proceedings of GL'2005, Third International Workshop on Generative Approaches to the Lexicon, p210-221.

Pradipta Ranjan Ray, Harish V. Sudeshna Sarkar and Anupam Basu.. 2003. *Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi*. In Proceedings of (ICON) 2003.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. *Multiword expressions: A pain in the neck for NLP* ,Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002) ,p1-15.

Manish Shrivastava, Nitin Agrawal, Smriti Singh and Pushpak Bhattacharya. 2005. *Harnessing Morphological Analysis in POS Tagging Task*, In Proceedings ICON 2005.

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. *The Lexical Encoding of MWEs* , Proceedings Second ACL Workshop on Multiword Expressions: Integrating Processing, p80-87.

David Yarowsky, G. Ngai, and R. Wicentowski. 2001. *Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora*, Proceedings of Human Language Technology Conference .p1 - 8.