Towards Identifying Hindi/Urdu Noun Templates in Support of a Large-Scale LFG Grammar

Sebastian Sulger
Department of Linguistics
University of Konstanz
Germany
sebastian.sulger@uni-konstanz.de

Ashwini Vaidya
University of Colorado
Boulder, CO
80309 USA
vaidyaa@colorado.edu

Abstract

Complex predicates (CPs) are a highly productive predicational phenomenon in Hindi and Urdu and present a challenge for deep syntactic parsing. For CPs, a combination of a noun and light verb express a single event. The combinatorial preferences of nouns with one (or more) light verb is useful for predicting an instance of a CP. In this paper, we present a semi-automatic method to obtain noun groups based on their co-occurrences with light verbs. These noun groups represent the likelihood of a particular noun-verb combination in a large corpus. Finally, in order to encode this in an LFG grammar, we propose linking nouns with templates that describe preferable combinations with light verbs.

1 Introduction

A problem that crops up repeatedly in shallow and deep syntactic parsing approaches to South Asian languages like Urdu and Hindi[1] is the proper treatment of complex predicates (CPs). In CPs, combinations of more than one element are used to express an event (e.g., *memory* + *do* = *remember*). In Urdu/Hindi, only about 700 simple verbs exist (Humayoun, 2006); the remaining verbal inventory consists of CPs. CPs are encountered frequently in general language use, as well as in newspaper corpora. Thus, any NLP application, whether shallow or deep, whether its goal be parsing, generation, question-answering or the construction of lexical resources like WordNet (Bhattacharyya, 2010) encounters CPs sooner rather than later.

There is a range of different elements that may combine with verbs to form a CP: verbs, nouns, prepositions, adjectives all occur in CPs. The constraints and productive mechanisms in verb-verb CPs are comparatively well-understood (e.g, see Hook (1974), Butt (1995), Butt (2010) and references therein). The domain of noun-verb CPs (N-V CPs) is less well understood, the standard theoretical reference being Mohanan (1994). It is only recently that researchers have tried to come up with linguistic generalizations regarding N-V CPs, some by using manual methods and linguistic introspection (Ahmed and Butt, 2011), others using a combination of manual and statistical methods (Butt et al., 2012).

Ahmed and Butt (2011) have suggested that the combinatory possibilities of N-V combinations are in part governed by the lexical semantic compatibility of the noun with the verb. Similar observations have been made for English (Barrett and Davis, 2003; North, 2005). If this is true, then lexical resources such as WordNet could be augmented with semantic specifications or feature information that can then be used to determine dynamically whether a given N-V combination is licit or not.

Knowledge about this kind of lexical-semantic information is essential in computational grammars. For example,lexicon entries and templates are required to define predicational classes. Implementing such a grammar for a language that makes heavy use of CPs calls for two requirements. First, the lexical items taking part in CP formation need to be present in the lexicon of the grammar; and second, the grammar needs to be engineered in a way that represents the correct linguistic generalizations. Any ap-

[1] Urdu is an Indo-Aryan language spoken primarily in Pakistan and parts of India, as well as in the South Asian diaspora. It is structurally almost identical to Hindi, although the lexicon and orthography differs considerably.

proach that is short of either of these requirements will result either in loss in coverage or overgeneration of the grammar.

The Hindi/Urdu ParGram Grammar forms part of a larger international research effort, the ParGram (Parallel Grammars) project (Butt et al., 1999; Butt et al., 2002; Butt and King, 2007). All of the grammars in the ParGram project are couched within the LFG framework and are implemented using the development platform XLE (Crouch et al., 2012). The grammars are developed manually and not via learning methods, which allows for a theoretically sound analysis that is also efficient from a computational point of view. The Hindi/Urdu ParGram Grammar aims at covering both Hindi and Urdu, which is a design decision that suggests itself due to the many structural conformities of the two languages (Butt et al., 2002). One weakness of the grammar is its currently relatively small lexicon, compared to other ParGram grammars. Adding to the lexicon is a critical step in extending the grammar coverage. This is even more true for N-V CPs due to the high frequency of such constructions in running text. Thus, we see the Hindi/Urdu ParGram Grammar as an ideal test bed for developing a lexical resource of Hindi nouns.

This paper is a first step in terms of constructing such a lexical resource for Hindi nouns. Following up on previous work, we assume that there are distinct groups of nouns; nouns that are part of a certain group tend to co-occur with the same light verb(s) and differ in their usage from members of other groups. Contrary to what has been done before, though, we do not dive into available corpora blindly to identify the groupings. Instead, we make use of a manually annotated treebank for Hindi, the Hindi and Urdu Treebank (HUTB, Bhatt et al. (2009)). Thus, we construct a *seed list* of nouns known to partake in CP formation in the HUTB. Since the HUTB is limited in its coverage, we then turn to a large Hindi corpus collected specifically for the present study and use clustering algorithms to put the nouns in the seed list into groups, based on light verb co-occurrence.[2]

Our aim is to arrive at a broad notion of noun similarity. If we can find groups of nouns that behave alike with respect to their light verbs, these groups can be included in an application such as the Hindi/Urdu ParGram Grammar to boost coverage as well as precision. Note that this notion of noun similarity is not the same as semantic classes in the sense of Levin (1993); however, it can serve as input for future research into semantic noun classification.

2 N-V Complex Predicates in Hindi and Urdu

As mentioned above, CPs are an important means of forming verbal predication in Hindi and Urdu. There is no single way of forming CPs; it is possible to find V-V CPs (Butt, 1995), ADJ-V combinations, P-V CPs (Raza, 2011) and N-V CPs (Mohanan, 1994) (see Ahmed et al. (2012) for some examples of each CP type.). In the present paper, we focus on identifying patterns of N-V CP formation. Here, the noun contributes the main predicational content. The verb in such constructions is usually called a light verb (Mohanan, 1994; Butt, 2003). The term represents the fact that these verbs are semantically bleached and specify additional information about the predication, such as whether the predicate has an agentive, telic or stative flavor. The light verb also determines the case marking on the subject, controls agreement patterns and contributes tense and aspect information. This is illustrated in (1).

(1) a. nadya=ne kɑhani yad k-i
Nadya.F.Sg=Erg story.F.Sg memory.F.Sg do-Perf.F.Sg
'Nadya remembered a/the story (agentively).' (lit. 'Nadya did memory of a/the story.')

b. nadya=ko kɑhani yad hɛ
Nadya.F.Sg=Dat story.F.Sg memory.F.Sg be.Pres.3.Sg
'Nadya remembers/knows a/the story.' (lit. 'At Nadya is memory of a/the story.')

[2] Note that despite the many structural conformities between Hindi and Urdu, the main difference between the two languages is in the lexicon; Modern Standard Hindi vocabulary is based on Sanskrit, while Urdu draws from a Persio-Arabic lexicon. This means that in principle, the methodology presented in this paper applied to Hindi needs to be applied to both languages separately. The equivalent Urdu study is pending future work and currently faces two major obstacles. First, the Urdu portion of the HUTB has not yet been released. Second, there is a major shortage of Urdu resources, with comparatively small corpora becoming available only recently (Urooj et al., 2012). Readily available Urdu sources (e.g., Wikipedia) are of minor quality.

c. nadya=ko kɑhani yad a-yi
Nadya.F.Sg=Dat story.F.Sg memory.F.Sg come-Perf.F.Sg
'Nadya remembered a/the story.' (lit. 'The memory of a/the story came to Nadya.')

In all of the examples in (1), it is evident that the noun and the verb form a single predicational element. The object *kahani* 'story' is thematically licensed by the noun *yad* 'memory', but it is not realized as a genitive, as would be typical for arguments of nouns (and as in the English literal translations). Rather, *kahani* 'story' functions as the syntactic object of the joint predication (see Mohanan (1994) for details on the argument structure and agreement patterns).

3 Previous Work

A recent study on the semantic classes of Persian N-V CPs using distributional vector-space methods has shown that verb vectors are a very useful indicator of noun similarity (Taslimipoor et al., 2012). The reported results are significantly better using the light verb dimension; Taslimipoor et al. (2012) state that this affirms their original intuition that a verb-based vector space model can better capture similarities across CPs. This finding is in agreement with our intuition that features based on light verbs best capture generalizations about N-V CPs.

There have been two studies on noun similarity based on co-occurrence of noun and light verb. Ahmed and Butt (2011) look at the light verbs *kar* 'do', *ho* 'be', *hu-* 'become' and identify three classes of nouns based on co-occurrence patters. The first consists of psychological nouns that occur with all three light verbs. The examples shown in (1) represent the class that is compatible with all of the light verbs surveyed. Other CP classes may only be compatible with a subset of light verbs. The second and third classes consist of nouns that are classified as more or less agentive in nature- based on their capacity to form CPs with *hu-* 'become'. For example, the noun *tamir* 'construction' is only compatible with the light verb *kar* 'do' but disallows *hu-* 'become'.

(2) a. bɪlal=ne mɑkan tɑmir kɪ-ya
Bilal.M.Sg=Erg house.M.Sg construction.F.Sg do-Perf.M.Sg
'Bilal built a/the house.'

b. *bɪlal=ko mɑkan tɑmir hɛ/hu-a
Bilal.M.Sg=Dat house.M.Sg construction.F.Sg be.Pres.3.Sg/be.Part-Perf.M.Sg
'Bilal built a/the house.'

In a follow-up study, Butt et al. (2012) attempted to identify Urdu N-V CPs automatically. After filtering out the irrelevant combinations, they found that most nouns were either psychological nouns (and occurred with all three light verbs) or nouns that were highly agentive and disallowed *hu-* 'become'. However, one of the drawbacks of their method was the use of an untagged corpus, which required extensive filtering in order to separate the light and non-light instances of these verbs.

We will draw upon the results of these two studies to motivate this present work. The classes identified by Ahmed and Butt (2011) seem promising, but the corpus work was done manually, and the total size of their data set is limited to 45 nouns. This can hardly serve as input to the development of a large-scale noun lexicon for a grammar. In constructing a lexical resource, we thus take a different route in that we try to expand the search space by using an external, manually-crafted resource and a larger set of light verbs to come up with more substantial noun groups.[3] In addition, we circumvent the problems faced in Butt et al. (2012)'s paper by filtering the list of nominal predicates in advance and by making use of a tagged corpus.

[3] One might argue that there are other features, beyond the light verbs, that one could use in identifying noun classes/groups, e.g., additional arguments licensed, case marking, etc. The reason why we (and other researchers before us) limit ourselves to the light verb occurrences is that Hindi and Urdu make rampant use of pro-drop (Kachru, 2006; Schmidt, 1999; Mohanan, 1994), which means that often, not all arguments are present in a sentence. Thus, the only reliable source of information about the noun is in fact the light verb, since this is the only obligatory element aside from the noun itself.

4 Methodology

In order to build a lexical resource of semantically similar nouns, we first need to identify whether these occur as part of a N-V CP. As we want to improve upon previous work, our aim was to include a large number of nouns. The Hindi portion of the Hindi and Urdu Treebank (Bhatt et al., 2009) includes N-V CPs that have been manually tagged with the dependency label POF (which stands for "part of"). The diagnostic criteria used for identifying CPs in the treebank is based on native speaker intuition. The POF label is used for adjectives and adverbs as well as nouns. We extracted only POF cases that were nouns only. This gave us an initial list of candidate nouns that were further filtered for spelling variations and annotation errors. After this stage, we had a list of 1207 nouns, which we will refer to as our *seed list*. The seed list consists of nouns that are a part of N-V CPs in the treebank.

Our aim was to include nouns that had at least 50 or more occurrences in order to ensure that we were looking at the most well-attested noun and light verb co-occurrences. For this task, the Hindi Treebank corpus (400,000 words) by itself would not be sufficient. For instance, if we applied our cutoff of 50 occurrences to the instances in the treebank, we would be left with only 20 nouns, which would not give us any meaningful groups. Therefore, we chose to use a larger corpus (including the treebank) in order to give us co-occurrence patterns for a noun from the seedlist. At the same time, we did not look at co-occurrence patterns with *any* verb. Instead, we chose a list of the most frequent light verbs from the treebank. This list is given below:

(3) *ho* 'be', *kar* 'do', *de* 'give', *le* 'take', *rakh* 'put', *lag* 'attach', *a* 'come'

Given the seed list and a short list of light verbs, our next step was the extraction of co-occurrences from a larger corpus.

4.1 Extracting Co-occurrences from a Large Corpus

In order to obtain a larger corpus, we scraped two large online sources of Hindi: the BBC Hindi website[4] as well as the Hindi Wikipedia.[5] Along with the Hindi Treebank, this corpus contains about 21 million tokens (BBC Hindi: ∼7 million, Hindi Wikipedia: ∼10 million, Hindi Treebank ∼4 million); by including the Wikipedia part, the resulting corpus extends beyond the newspaper domain. In a second step, the corpus was automatically POS tagged using the tagger described in Reddy and Sharoff (2011).

We were interested in extracting co-occurrences that had the following pattern: *seed list item* + *light verb*. A match would only occur if one of the light verbs occurred directly to the right of the noun (i.e., an item tagged as NN by the POS tagger). Our method therefore did not take into account any N-V CPs that were syntactically flexible, i.e., when the noun and the light verb did not occur next to each other. Those cases where the noun may be scrambled away from the light verb (e.g., topicalization of the noun) are not numerous and occur rarely in the Hindi Treebank (only about 1% of the time).[6]

4.2 Clustering & Evaluation

In the next step, a clustering algorithm was applied to the data. This was done using the clustering tool described in Lamprecht et al. (2013). At the moment, the tool features two clustering algorithms: the k-means algorithm (MacQueen, 1967) as well as the Greedy Variance Minimization (GVM) algorithm.[7] We made use of an automatic method using Hindi WordNet (Bhattacharyya, 2010) to choose the best partition value. We followed the technique described in Van de Cruys (2006), which uses WordNet relations to arrive at the most semantically coherent clusters. We define semantic coherence as the similarity between items in a cluster, based on an overlap between their WordNet relations. Specifically, for each $k = 2-10$, we iterated through the automatically generated clusters and performed the following steps:

1. Using WordNet, we extracted synonyms, hypernyms and hyponyms for every word in a cluster.

[4] `http://www.bbc.co.uk/hindi`

[5] `http://dumps.wikimedia.org/hiwiki`

[6] Mohanan (1994) even goes so far as to call the topicalization of nouns in N-V CPs ungrammatical.

[7] `http://code.google.com/p/tomgibara/`

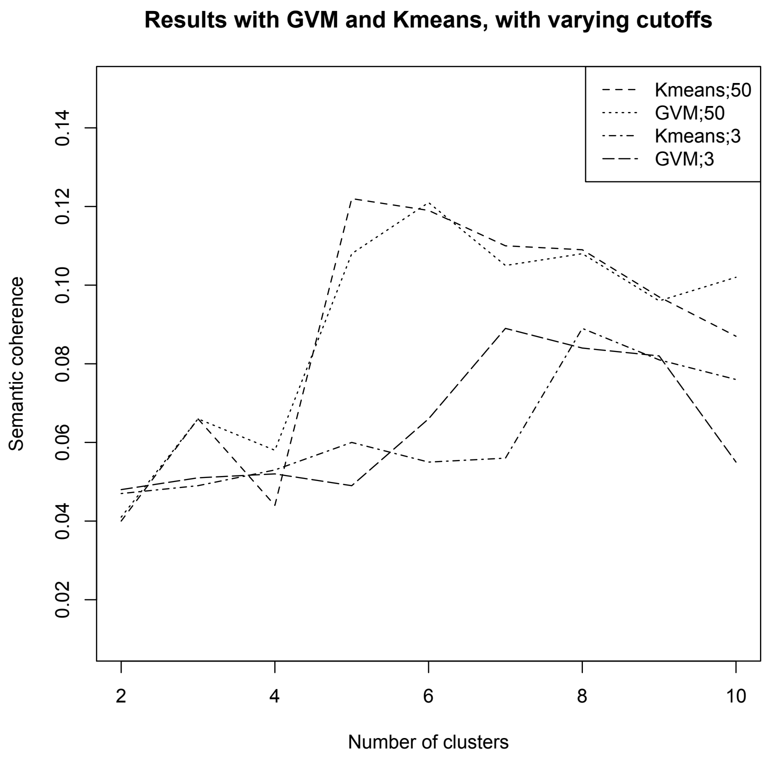

Figure 1: Choosing the best value for k. k-means with a frequency cutoff of 50 gives us the most semantically coherent clusters for $k = 5$

2. A word that had the most semantic relations with every word in the cluster was chosen as its centroid.
3. The co-hyponyms i.e., the hyponyms of the hypernyms for this centroid were extracted from WordNet (along with its synonyms, hypernyms and hyponyms).
4. In order to calculate precision for each cluster, we counted the number of words in that cluster that overlapped with the words in the centroid's relations.

We averaged the precision across all clusters for every k value. We found that precision for each cluster gradually improved until we got the most semantically coherent partitions for $k = 5$ using k-means, for 522 nouns occurring with a frequency of 50 and above. Table 1 shows the values for k using our WordNet evaluation method, for $k = 5 - 9$.

	Frequency = 3		**Frequency = 50**	
Size of k	**GVM**	**k-means**	**GVM**	**k-means**
5	0.049	0.060	0.107	0.122
6	0.066	0.055	0.121	0.119
7	0.089	0.056	0.104	0.110
8	0.084	0.089	0.108	0.109
9	0.082	0.081	0.095	0.097

Table 1: Semantic coherence values for $k = 5 - 9$ for clustering algorithms GVM and k-means

In Figure 1, we have plotted the semantic coherence values against the number of clusters to show the best results. The k-means algorithm performed only slightly better than GVM, and after $k = 5$, the semantic coherence of the clusters declined again. As a point of comparison, we also plotted k-means and GVM results for nouns that occurred more than 3 times in the data (i.e., using a far smaller cutoff). In this configuration, the best results are achieved for a higher k value (i.e., between 7 or 8), but we rejected this on the basis of a better semantic coherence value for k-means with a cutoff of 50.

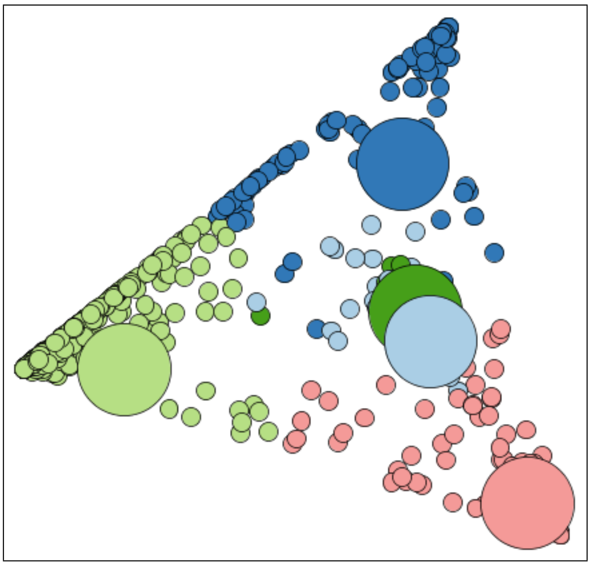

Figure 2: Visualization for $k = 5$ clusters

5 Analysis

Lamprecht et al. (2013)'s tool is useful for visual cluster inspection; e.g., the tool created the visual clustering in Figure 2 using the k-means algorithm with $k = 5$. The visualization enables the user to inspect the data points and derive initial generalizations. For example, Figure 2 shows a visualization with colored circles that encode membership within a cluster. The larger circles represent cluster centroids. The visualization enables us to see three most frequently occurring light verbs, viz. *kar* 'do', *ho* 'be' and *de* 'give', represented by light green, dark blue and pink respectively. Many nouns alternate with 'do' and 'be', hence there is a visible continuum between the light green and dark blue data points. The two clusters in the centre show a dark green cluster, consisting of only a handful of nouns that alternate with the light verbs *rakh* 'keep', *lag* 'attach' and *a* 'come'. The light blue cluster on the other hand is larger and is dominated by the light verb *le* 'take'.

In order to further interpret the results of our study, we also referred to a secondary result from our WordNet evaluation. While extracting the extent of overlap of the semantic relations, we also extracted the 'semantic' centroids i.e. words that had the most semantic relations with every other word in the cluster (see Section 4.2). For our best result of $k = 5$, these centroids also revealed semantic similarities in the five clusters that we found. For instance, dynamic events that are inanimate and abstract and take an agentive argument will lend themselves to combinations with *kar* 'do'. Similarly, events that include the semantic property of 'transfer' will occur with *de* 'give' (although there is ostensibly an overlap here, as these events invariably also require agentive arguments). The light verb *ho* 'be' occurs often with nouns that denote mental states, resulting in an experiencer subject — but this group also includes nouns that alternate with *kar*. Less frequently occurring light verbs, especially *rakh* 'keep', *lag* 'to attach' and *a* 'come' show fewer alternations, as they do not occur in combination with all nouns and are grouped together in this result. These light verbs often form N-V CPs with a more idiosyncratic meaning, in fact Davison (2005) has argued that some of these light verbs may form 'incorporation idioms' (rather than true N-V CPs).

The average figures of N-V CP co-occurrences for a certain cluster inform us about the likelihood of a certain group of nouns to co-occur with a certain light verb. For instance, this noun grouping shows a high likelihood of occurrence with *kar* 'do' and *le* 'take', but not very likely at all with *lag* 'attach'. We take this distribution to reflect a difference in the syntactic behavior of the nouns: While the productive patterns indicate CP formation, the less productive patterns do not represent CPs at all. This is a finding in line with Butt et al. (2012), who ended up deleting many low-frequency patterns which turned out to be non-CP combinations. Similar tendencies can be derived for the five groups of nouns derived from our clustering experiment above. This information is useful for a task like lexicon development for a computational grammar. The following section therefore explores the possibility of encoding noun group information in a computational Lexical Functional Grammar.

6 Noun Groups in Hindi/Urdu Grammar Development

Our experiments show that nouns appear with several different distributions, often with one dominant light verb, but also with the possibility of occurring with one or two other light verbs. The clusters do not represent absolute certainties about N-V CPs, but report *tendencies* of occurrences; e.g., the relative frequencies of the cluster centroid for the noun group dominated by *de* ‘give’ is shown below.

(4) *de* ‘give’ 0.75, *kar* ‘do’ 0.08, *le* ‘take’ 0.06, *ho* ‘be’ 0.06, *a* ‘come’ 0.02, *rakh* ‘keep’ 0.02, *lag* ‘attach’ 0.01

In this section, we discuss the integration of our Hindi noun groupings into the grammar via the construction of templates that can be augmented to model the relevant linguistic generalizations in terms of constraints inspired by optimality theory (OT, Prince and Smolensky (2004)). A serious evaluation of the effect on the grammar of adding in this lexical resource is planned for future work.

6.1 Templates in XLE

In XLE, grammar writers can define templates in a special section of the grammar that can be called from the lexicon. Templates allow generalizations to be captured and, if necessary, changes to be made only once, namely to the template itself (Butt et al., 1999; Dalrymple et al., 2004). Consider the template in (5), which models intransitive verbs in English; these are represented in LFG terms as predicates that apply to a single grammatical function, a subject. The lexical entry in (6) for the English intransitive verb *laugh* calls up the `INTRANS` template; the argument supplied to the template is substituted for the `P`(redicate) value inside the template definition.

(5)
```
INTRANS(P) = (^ PRED) = '_P<(^ SUBJ)>'
@NOPASS.
```

(6)
```
laugh V @(INTRANS laugh).
```

In ParGram grammars, the lexicons are generally organized so that each verb subcategorization frame corresponds to a different template. Similarly, templates can be defined to encode a given set of generalizations about how certain groups of nouns combine with different light verbs. Consider the N-V CPs in (7). The noun *ıshara* ‘signal’ forms part of the cluster dominated by *de* ‘give’ (i.e., the cluster with the frequencies shown in (4)) and thus occurs most frequently with *de* ‘give’ as well as *kar* ‘do’.

(7) a. nadya=ne bılal=ko ıshara dı-ya
Nadya.F.Sg=Erg Bilal.M.Sg=Dat signal.M.Sg give-Perf.M.Sg
‘Nadya signaled Bilal.’ (lit. ‘Nadya gave a signal to Bilal.’)

b. nadya=ne bılal=ko ıshara kı-ya
Nadya.F.Sg=Erg Bilal.M.Sg=Acc signal.M.Sg do-Perf.M.Sg
‘Nadya signaled Bilal.’ (lit. ‘Nadya made a signal towards Bilal.’)

The lexical entry of the noun *ıshara* ‘signal’ is given in (8).[8] The entry points to the template `NVGROUP2` which is defined as in (9). This version of the template constrains the verbal type of the overall predication to be a CP either with the light verb *de* ‘give’ or with the light verb *kar* ‘do’, or to not be a CP at all. Thus, only light verb options with relative frequencies equaling or above 0.08 (i.e., 8%) are accepted, an arbitrary threshold.

[8] The transliteration scheme employed in the Hindi/Urdu ParGram Grammar is described in Malik et al. (2010).

```
(8) iSArA NOUN-S XLE (^ PRED) = 'iSArA<(^ OBJ)>'
                     @NVGROUP2.
```

```
(9) NVGROUP2 = { (^ VTYPE COMPLEX-PRED-FORM) =c dE
               |(^ VTYPE COMPLEX-PRED-FORM) =c kar
               | ~ (^ VTYPE COMPLEX-PRED-FORM)}
```

6.2 Preferred CPs

The template in (9), however, misses out on the fact that for all the groups identified, there are N-V combinations that are more productive (and thus more likely to be CP constructions) than other combinations (which are more likely to be non-CP constructions, e.g., plain objects). In XLE, grammar developers can model statistical generalizations using special marks that were inspired by Optimality Theory (Prince and Smolensky, 2004). On top of the classical constraint system of existing LFG grammars, a separate projection, o-structure, determines a preference ranking on the set of analyses for a given input sentence. A relative ranking is specified for the constraints that appear in the o-projection, and this ranking serves to determine the winner among the competing candidates. The constraints are also referred to as OT marks and are overlaid on the existing grammar (Frank et al., 1998).

OT marks can be added in the appropriate place in the grammar to punish or prefer a certain analysis. For example, (10) states that `Mark1` is a member of the optimality projection. The order of preference of a sequence of OT marks can be specified in the configuration section of the grammar; an example preference ordering is given in (11). Here, the list given in `OPTIMALITYORDER` shows the relative importance of the marks. In this case `Mark5` is the most important, and `Mark1` is the least important. Marks that have a `+` in front of them are preference marks. The more preference marks that an analysis has, the better. All other marks are dispreference marks (the fewer, the better).

```
(10) ...  Mark1 $ o::* ...
```

```
(11) OPTIMALITYORDER Mark5 Mark4 Mark3 +Mark2 +Mark1.
```

Given the relative ordering of light verb tendencies in our noun groups, we can augment the templates with OT marks that represent such tendencies. The noun template in (9) is changed in two ways. First, *all* the light verbs are included; second, each disjunct is extended by two OT marks that represent the statistical likelihood of this particular combination forming a CP or not.[9] The ordering of the marks is shown in (13), where the mark `cp-dispref` is most severely punished, and the mark `+cp-pref` is most strongly preferred. With an ordering like this, a CP analysis for (7a) is preferred, while a compositional analysis is dispreferred by XLE; the inverse will apply to *ıshara lag*, which is not a CP.

```
(12) NVGROUP2 = { { (^ VTYPE COMPLEX-PRED-FORM) =c dE
                cp-pref $ ::*
                | ~ (^ VTYPE COMPLEX-PRED-FORM)
                non-cp-dispref $ ::* }
                ...
                |(^ VTYPE COMPLEX-PRED-FORM) =c lag}.
                cp-dispref $ o::*
                | ~ (^ VTYPE COMPLEX-PRED-FORM)
                non-cp-pref $ ::* } }.
```

```
(13) OPTIMALITYORDER cp-dispref non-cp-dispref +cp-pref +non-cp-pref.
```

[9] For space reasons, only the disjuncts for *de* 'give' as well as *lag* 'attach' are shown.

7 Conclusion

We have discussed a corpus study of Hindi/Urdu N-V CPs that makes use of a novel methodology in terms of a noun seed list and an evaluation based on WordNet. We found that the k-means algorithm with $k = 5$ and a frequency cutoff of 50 gave us the best result in terms of semantic coherence of the resulting clusters. We are optimistic that the resulting noun groups can be used in different NLP settings and have presented one such setting, the Hindi/Urdu ParGram Grammar, where lexical information about nouns and their combinatory possibilities in CPs are vital for grammar extension.

Acknowledgements

We would like to thank the DAAD (Deutscher Akademischer Austausch Dienst) for sponsoring Ashwini Vaidya's research stay at the University of Konstanz and Dr Miriam Butt for hosting her. We are also thankful to Dr Miriam Butt, Dr Martha Palmer and Christian Rohrdantz for their feedback on this paper. Any errors that remain are our own.

References

Tafseer Ahmed and Miriam Butt. 2011. Discovering Semantic Classes for Urdu N-V Complex Predicates. In *Proceedings of the International Conference on Computational Semantics (IWCS 2011).*

Tafseer Ahmed, Miriam Butt, Annette Hautli, and Sebastian Sulger. 2012. A Reference Dependency Bank for Analyzing Complex Predicates. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), May.

Leslie Barrett and Anthony R Davis. 2003. Diagnostics for determining compatibility in English support-verb-nominalization pairs. In *Proceedings of the 4th international conference on Computational Linguistics and Intelligent text processing (CICLing 03).*

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Sharma, and Fei Xia. 2009. A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189, Suntec, Singapore, August. Association for Computational Linguistics.

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 3785–3792.

Miriam Butt and Tracy Holloway King. 2007. Urdu in a Parallel Grammar Development Environment. *Language Resources and Evaluation: Special Issue on Asian Language Processing: State of the Art Resources and Processing*, 41.

Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook.* CSLI Publications.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Miriam Butt, Tina Bögel, Annette Hautli, Sebastian Sulger, and Tafseer Ahmed. 2012. Identifying Urdu Complex Predication via Bigram Extraction. In *In Proceedings of COLING 2012, Technical Papers*, pages 409 – 424, Mumbai, India.

Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu.* CSLI Publications.

Miriam Butt. 2003. The Light Verb Jungle. *Harvard Working Papers in Linguistics*, 9.

Miriam Butt. 2010. The Light Verb Jungle: Still Hacking Away. In Mengistu Amberber, Brett Baker, and Mark Harvey, editors, *Complex Predicates in Cross-Linguistic Perspective*. Cambridge University Press.

Dick Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula Newman, 2012. *XLE Documentation.* Palo Alto Research Center.

Mary Dalrymple, Ronald M. Kaplan, and Tracy Holloway King. 2004. Linguistic Generalizations over Descriptions. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG04 Conference*. CSLI Publications.

Alice Davison. 2005. Phrasal predicates: How N combines with V in Hindi/Urdu. In Tanmoy Bhattacharya, editor, *Yearbook of South Asian Languages and Linguistics*, pages 83–116. Mouton de Gruyter.

Anette Frank, Tracy Holloway King, Jonas Kuhn, and John T. Maxwell III. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In *Proceedings of the LFG98 Conference*. CSLI Publications.

Peter Hook. 1974. *The Compound Verb in Hindi*. Center for South and Southeast Asian Studies, University of Michigan.

Muhammad Humayoun. 2006. Urdu Morphology, Orthography and Lexicon Extraction. Master's thesis, Department of Computing Science, Chalmers University of Technology.

Yamuna Kachru. 2006. *Hindi*. John Benjamins.

Andreas Lamprecht, Annette Hautli, Christian Rohrdantz, and Tina Bögel. 2013. A Visual Analytics System for Cluster Exploration. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–114, Sofia, Bulgaria, August. Association for Computational Linguistics.

Beth Levin. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. The University of Chicago Press.

James B. MacQueen. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press.

Muhammad Kamran Malik, Tafseer Ahmed, Sebastian Sulger, Tina Bögel, Atif Gulzar, Ghulam Raza, Sarmad Hussain, and Miriam Butt. 2010. Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*.

Tara Mohanan. 1994. *Argument Structure in Hindi*. CSLI Publications.

Ryan North. 2005. *Computational Measures of the Acceptability of Light Verb Constructions*. Ph.D. thesis, University of Toronto.

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishing.

Ghulam Raza. 2011. *Subcategorization Acquisition and Classes of Predication in Urdu*. Ph.D. thesis, University of Konstanz.

Siva Reddy and Serge Sharoff. 2011. Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, pages 11–19, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Ruth Laila Schmidt. 1999. *Urdu: An Essential Grammar*. Routledge.

Shiva Taslimipoor, Afsaneh Fazly, and Ali Hamzeh. 2012. Using Noun Similarity to Adapt an Acceptability Measure for Persian Light Verb Constructions. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

S. Urooj, F. Jabeen, F. Adeeba, R. Parveen., and S. Hussain. 2012. Urdu Digest Corpus. In *Proceedings of the Conference on Language and Technology 2012*, Lahore, Pakistan.

Tim Van de Cruys. 2006. Semantic Clustering in Dutch. In *Proceedings of the Sixteenth Computational Linguistics in Netherlands (CLIN)*, pages 17–32.