Identifying Urdu Complex Predication via Bigram Extraction

Miriam Butt[1] *Tina Bögel*[1]
Annette Hautli[1] *Sebastian Sulger*[1] *Tafseer Ahmed*[2]
(1) Universität Konstanz, Germany
(2) University of Karachi, Pakistan
`firstname.lastname@uni-konstanz.de`[1], `tafseer@uok.edu.pk`[2]

Abstract

A problem that crops up repeatedly in shallow and deep syntactic parsing approaches to South Asian languages like Urdu/Hindi is the proper treatment of complex predications. Problems for the NLP of complex predications are posed by their productiveness and the ill understood nature of the range of their combinatorial possibilities. This paper presents an investigation into whether fine-grained information about the distributional properties of nouns in N+V CPS can be identified by the comparatively simple process of extracting bigrams from a large "raw" corpus of Urdu. In gathering the relevant properties, we were aided by visual analytics in that we coupled our computational data analysis with interactive visual components in the analysis of the large data sets. The visualization component proved to be an essential part of our data analysis, particular for the easy visual identification of outliers and false positives. Another essential component turned out to be our language-particular knowledge and access to existing language-particular resources. Overall, we were indeed able to identify high frequency N-V complex predications as well as pick out combinations we had not been aware of before. However, a manual inspection of our results also pointed to a problem of data sparsity, despite the use of a large corpus.

KEYWORDS: Urdu, complex predicates, visualization, bigrams, corpus.

1 Introduction

A problem that crops up repeatedly in shallow and deep syntactic parsing approaches to South Asian languages like Urdu/Hindi[1] is the proper treatment of complex predications. Whereas verbal expressions in European languages are mostly realized with a single predicate (e.g. 'to remember'), South Asian languages tend to use combinations of more than one element to express an action (e.g., memory+do = 'remember'). In Urdu, only about 700 simple verbs exist (Humayoun 2006) — the remaining verbal inventory consists of complex predications. Complex predicates (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 (cf. Bhattacharyya 2010), encounters complex predication sooner rather than later.

While the productive capability as well as constraints on V+V combinations are now comparatively well understood from a theoretical point of view (e.g, see Hook 1974; Butt 1995, 2010 and references therein), the constraints governing N+V, ADJ+V as well as P+V combinations are less well understood (the standard reference for N+V is Mohanan 1994, there is no standard reference for the other types). This is despite the fact that these latter three are very productive. Indeed, very little is known overall about P+V combinations (with Raza 2011 providing a first description), but as our results show, they do occur in newspaper corpora (section 5).

With respect to NLP applications the frequency and productivity of complex predications means that it is not possible to construct a static list of N/ADJ/P+V combinations, rather, there must be a way in which one can dynamically predict which kinds of combinations are possible and which should be impossible. In a recent paper, Ahmed and Butt (2011) propose that the combinatory possibilities of N+V combinations are in part governed by the lexical semantic compatibility of the noun with the verb. They propose an initial classification based on a small corpus study. If they are right, then lexical resources such as WordNet or lexica used in deep grammars 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.

For this paper, we took Ahmed and Butt (2011) as a point of departure and explored whether we could confirm and perhaps extend their results with respect to a larger corpus study. Given that to date no high-quality annotated large-scale corpus for Urdu exists,[2] we decided to experiment with a large-scale 7.9 million newspaper corpus of Urdu we have collected.

The idea was to take advantage of statistical methods and proceed per standard methods currently embraced in the field. That is, use an available corpus that should in principle be large enough to adequately reflect language use and to extract bigrams from this corpus in order to identify patterns in N+V combinations and to use our knowledge of the extracted patterns in further NLP applications.

In pursuing this experiment, we were indeed able to adduce new information about combinatory possibilities in CPS (section 5). However, our experiment also provides a cautionary tale with respect to diving into a corpus "blindly", i.e, assuming that mere statistical analysis will provide good enough results and any noise due to language particular considerations will simply wash out if the corpus is large enough. Some of the difficulties we encountered had to do with the

[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 orthographies differ considerably.

[2] However, some data is becoming available via the Hindi-Urdu Treebank (Bhatt et al., 2009) and a large-scale, balanced corpus for Urdu will be released soon (Urooj et al., 2012).

non-standardized nature of Urdu orthography, some with the structure of Urdu (section 3) and some with the complex nature of the data (section 4).[3]

With respect to the complex interrelationships between our data, we were able to achieve significant improvement by using novel methods coming from the field of *visual analytics* (Card et al., 1999; Thomas and Cook, 2005; Keim et al., 2008, 2010). Instead of trying to make sense of bare numbers, we used a visualization tool that maps figures to colors and therefore makes the statistical analysis immediately accessible via easy to process visual means. The visualization allowed us to assess our complex data "at-a-glance", to take corrective measures and to generate new hypotheses as to CP formation and the validation of existing hypotheses.[4]

The paper is organized as follows: section 2 presents relevant background information with respect to the linguistic phenomenon of CPs in section 2.1, followed by related work in section 2.2. Sections 3 and 4 present the details of our study. We describe the steps with respect to the corpus preparation and bigram extraction (section 3.1), the statistical basis for the collocation analysis (section 3.2), the clustering performed on N+V bigrams (section 3.3) and the visualization used (section 4). We also reflect on the depth of the language particular knowledge that was necessary and discuss our results in section 5. We were not in fact able to achieve our initial goal of confirming or extending Ahmed and Butt's original hypothesis. This turns out to be due to data sparsity, even with a 7.9 million token corpus. However, we were able to identify previously unreported information about highly productive combinations and gain further structural insights into the language, particularly due to our use of novel methods coming from the field of visual analytics.

2 Motivation and Background

2.1 Urdu and complex predicates

As already mentioned, an important means of expressing verbal concepts in Urdu is the usage of CPs. There is not one single way of forming CPs, rather there are several different kinds of V+V CPs, different kinds of ADJ+V combinations and different classes of N+V CPs (Butt, 1995; Mohanan, 1994). A discussion of the various combinatory possibilities and types of classes identified so far would lead us too far afield, but see Ahmed et al. (2012) for some examples of each type. Moreover, CPs are highly productive with respect to their combinatorial possibilities and can also be stacked on top of one another. This means that the compilation of a static list of possible combinations of different types of CPs is not feasible as it does not solve the challenges inherent in capturing the syntactic combinatory possibilities and the semantic interpretability via computational means — simply searching a corpus for all possible (and potentially infinitely many) combinations cannot do justice to the dynamic and manifold combinatory possibilities.

For the purpose of this paper, we decided to focus on N+V CPs, i.e., CPs which are formed by combining a noun and a verb. The verb is generally called a *light verb* (Mohanan, 1994), as the noun provides the main predicational content, while the verb specifies additional information

[3] In light of these difficulties, one reviewer encouraged us to explore the possibility of using treebanks. However, the existing definitive treebank for Urdu/Hindi does not do well with annotating complex predicates, merely noting that several bits of a clause are "part-ofs" another word or constituent (Bhatt et al., 2009). This part-of relation is not confined to CPs and when it does indicate a CP, it does not specify what kind. Treebanks which aim to do better in this respect are in the process of being built (Hautli et al., 2012; Ahmed et al., 2012). Indeed, the work reported on here was partly motivated by our on-going treebank effort.

[4] The visualization tool we used was designed by Christian Rohrdantz as part of an on-going cooperation. We would also like to thank him for a very useful discussion of the material in this paper.

about the action/event, like whether the action was agentive or telic. On the morphosyntactic side, the light verb determines the case marking on the subject, controls agreement patterns and provides information about tense and aspect. Some standard examples are in (1).[5]

(1) a. لڑکی نے کہانی یاد کی
lɑṛki=ne kɑhani yad k-i
girl.F.Sg=Erg story.F.Sg.Nom memory.F.Sg.Nom do-Perf.F.Sg
'The girl remembered a/the story.' (lit.: 'The girl did memory of the story.')

b. لڑکی کو کہانی یاد ہے
lɑṛki=ko kɑhani yad hɛ
girl.F.Sg=Dat story.F.Sg.Nom memory.F.Sg.Nom be.Pres.3P.Sg
'The girl remembers/knows a/the story.' (lit.: 'Memory of the story is at the girl.')

c. لڑکی کو کہانی یاد ہوی
lɑṛki=ko kɑhani yad hu-i
girl.F.Sg=Dat story.F.Sg.Nom memory.F.Sg.Nom be.Part-Perf.F.Sg
'The girl came to remember a/the story.'
(lit.: 'Memory of the story became to be at the girl.')

In all of the examples in (1), it is evident that the noun and the verb form a single predicational element. The object *kɑhani* '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, *kɑhani* 'story' functions as the syntactic object of the joint predication (see Mohanan 1994 for details on the argument structure and agreement patterns).

Mohanan (1994) already identified two subclasses of N+V CPs: one in which the light verb agrees with the noun (and, confusingly, the noun is both a syntactic object and a part of the verbal predication) and one in which the verb does not agree with the noun (and instead agrees with some other NP in the clause; (1) is of the non-agreeing type). Ahmed and Butt (2011) propose three further classes that cut across this morphosyntactic distinction. Their classification is based on the observation that while about five different verbs can function as light verbs in N+V CPs, not all nouns are necessarily compatible with each of these verbs.

The examples in (1) represent just one class of N+V CPs. This class is compatible with all of the possible light verbs and is identified as a smallish class in Ahmed and Butt (2011). (1) shows the combination with three of these. In (1a) the noun *yad* 'memory' is combined with the light verb *kɑr* 'do'. In this case the subject must be ergative and the overall reading is one of an agentive, deliberate remembering. In (1b), in contrast, *lɑṛki* 'girl' is already taken to be in the state of remembering the story. The difference between (1b) and (1c) is one of stative vs. eventive, so that in (1b), *lɑṛki* 'girl' is already taken to be in the state of remembering the story (and not actively entering a state of remembering the story). In (1c) the light verb is the participial form of *ho* 'be' and essentially means 'become'.

Table 1 summarizes the results presented in Ahmed and Butt (2011). Class A refers to examples as in (1) and this type seems to encompass what is known as *psych-predications*, i.e, actions of remembering, thinking, feeling, etc. However, not all nouns are as versatile as *yad* 'memory'.

[5] The Urdu script, which is also provided in (1), is based on the Arabic script and is written from right to left.

One type, Class B in Table 1, does not allow the subject to be non-agentive. That is, it does not allow combinations with those light verbs that require a dative subject, cf. (1b–c). This pattern is illustrated with an example in (2) and this class was identified as by far the largest class in Ahmed and Butt (2011).

	Light Verb			
N+V Type	kar 'do'	hɛ 'be'	hU- 'become'	**Analyis**
CLASS A	+	+	+	psych-predications
CLASS B	+	−	+	only agentive
CLASS C	+	+	−	do not allow subject to be an undergoer

Table 1: Classes of nouns identified by Ahmed and Butt (2011)

(2) a. بلال نے مکان تمیر کیّا
bılal=ne makan tamir ki-ya
Bilal.M.Sg=Erg house.M.Sg.Nom construction.F.Sg do-Perf.M.Sg
'Bilal built a/the house.'

b. بلال کو مکان تمیر ھ/ھوا
*bılal=ko makan tamir hɛ/hu-a
Bilal.M.Sg=Dat house.M.Sg.Nom construction.F.Sg be.Pres.3.Sg/be.Part-Perf.M.Sg
'Bilal built a/the house.'

The third class allows dative subjects in principle, but not when they are the undergoer of the action (cf. (1c)). An example with the noun *ıntızar* 'wait' is given in (3). Other nouns that pattern similarly are *taslim* 'acceptance' and *bardaʃt* 'tolerance'. This was again identified as a fairly small class in Ahmed and Butt (2011).

(3) a. بلال نے نادیا کا انتزار کیّا
bılal=ne nadya=ka ıntızar ki-ya
Bilal.M.Sg=Erg Nadya.F.Sg=Gen.M.Sg wait.M.Sg do-Perf.M.Sg
'Bilal waited for Nadya.'

b. بلال کو نادیا کا انتزار ھ/ھوا
bılal=ko nadya=ka ıntızar hɛ/*hu-a
Bilal.M.Sg=Dat Nadya.F.Sg=Gen.M.Sg wait.M.Sg be.Pres.3.Sg
'Bilal is waiting/*waited for Nadya.'

As already mentioned, while the classes identified by Ahmed and Butt (2011) seem promising, the corpus work was done manually and was limited to a total of 45 nouns. The goal of our experiment was to expand the search space by using automatized methods, to thus extract information about a significant number of nouns and to be able to confirm, expand or revise Ahmed and Butt's proposal. In particular, understanding the semantic constraints on N+V CP formation in more detail would be welcome for further NLP applications.

2.2 Related work

Most related work on South Asian languages with the focus on an automatic extraction of complex predicates has been done for Hindi (Mukerjee et al., 2006; Chakrabarti et al., 2008; Sinha, 2009) and Bengali (Das et al., 2010; Chakraborty and Bandyopadhyay, 2010), with the predominant aim of identifying and extracting CPS from corpora or treebanks. Mukerjee et al. (2006) identify Hindi CPS based on the statistical correspondence between English verbs and Hindi multiword expressions (MWES), using the parallel EMILLE corpus in a sentence-aligned version. The assumption here is that a verb in English will project onto a CP MWE in Hindi most of the times.[6] Bhattacharyya (2010), in presenting a WordNet for Hindi, lists the ~100 most frequently occurring CPS. This is welcome information to be included in a WordNet. However, this static list ultimately does not do justice to either the overall productivity of CPS, nor does it provide details as to their syntactic/semantic patterning.

This paper has different aims than previous work in that we are not interested in merely identifying CP constructions in a large corpus, but are trying to understand more about their syntactic and semantic properties. To this end, we follow the classic assumption of Levin (1993) that semantic predicational classes can be identified on the basis of a study of the syntactic contexts the predicates occur in (cf. also Schulte im Walde 2009; Raza 2011).

3 Methodology

In this section, we describe the methodology used for harvesting CP candidates from a raw Urdu text corpus and for identifying classes among the candidate CPS via a simple bigram analysis in conjunction with techniques from visual analytics. We also discuss the problems we discovered to be associated with working with an unannotated "raw" corpus for a language like Urdu. Our overall results make a strong case for the prioritization of (mainly manual) high quality resource building — it seems that significant progress with entirely shallow methods cannot be achieved unless high quality, linguistically informed resources can be drawn upon.

3.1 Corpus

A prerequisite for our experiment is access to a large corpus for Urdu. No large corpus for Urdu (annotated or not) is publicly available to date (but see Urooj et al. 2012). We therefore decided to use a 7.9 million word corpus we have been harvesting from the BBC Urdu website[7] for a number of months. The corpus consists of news articles on various different topics, e.g., entertainment, multimedia, science, sports. These articles were automatically collected and parsed into raw text using the Perl HTML module.[8] Inspection of the corpus showed that the BBC Urdu script encoding is particularly clean and systematic in comparison to other Urdu newspaper sites. We therefore did not clean or preprocess the corpus with respect to punctuation, orthography or other normalization issues.

[6] One reviewer wonders whether we could not have used comparable/parallel data from Wikipedia to help us in identifying semantic classes among nouns. This is a potential line of avenue to pursue, but one that is not taken on lightly. Take the pair English-Urdu, for example. Most of the simple verbal predications do correspond to some kind of complex predication in Urdu, but different kinds, none of which are well understood. In addition, English also contains N-V constructions such as *take a bath* whose collocational properties are not very well understood either. It is unclear whether looking at two sets of data for which the collocational constraints within each set of data is not well understood will be able to yield useful results (or results that could not have been achieved manually with less effort).

[7] http://www.bbc.co.uk/urdu/

[8] http://search.cpan.org/dist/HTML-Parser/Parser.pm

3.2 Bigram Collocation Extraction

As a first pass, we went through the corpus and extracted all the bigrams containing the four verbs *kɑr* 'do', *ho* 'be', *hu-* 'become' and *rɑk^h* 'put'. We decided to extend our set of light verbs beyond the ones used in Ahmed and Butt (2011) by including *rɑk^h* 'put' in the hopes of arriving at a finer-grained picture of the distribution of the nouns in N+V CPS. We thus looked for all instances of these verbs (in all of their conjugated forms) and extracted them plus any word immediately preceding them. These bigrams were stored and their frequency was recorded. This procedure yielded four initial lists of bigrams, one for each of the four verbs *kɑr* 'to do', *ho* 'to be', *hu* 'to become' and *rɑk^h* 'to put'.

A manual inspection of these lists revealed that while we were finding N+V CPS of the type we were looking for, most of the highly frequent bigrams were either junk or not the kinds of combinations we were trying to find. We also had an issue with low frequency in that many bigrams were recorded just once or twice. Since Urdu allows quite a bit of scrambling and also allows the nouns to be scrambled away from the light verbs, it was clear from the outset that we would not necessarily net all of the instances of N+V CPS that occur in the corpus. However, we were not prepared for the amount of false hits we did get.

Closer inspection of the bigram lists revealed that many of the false hits were due to certain case markers, conjunctions and pronouns, which all occur frequently before our set of verbs. This is actually to be expected, given the structure of the language (see section 3.4 for discussion). Since our set of verbs is very versatile in that they can not only act as light verbs in N+V CPS, but also function as main verbs and as auxiliaries, many of our top bigrams turned out to be verb-verb combinations of one type or another. Although these are all valid bigrams, they are unwanted noise in the context of our investigation. Further errors were introduced by punctuation and tokenization or white-space issues (see section 3.4).

As a consequence, the initial bigram lists, which consisted of 16033 possible combinations, were pruned. For one, all bigrams which appeared less than 6 times in the corpus were removed. We assumed that most of them were tokenization errors and other machine processing lapses. We also removed all those bigrams which had a negative X^2 value (see below). This reduction left us with only around 4500 bigrams. In a second step, we constructed a list of stop words from various sources. For one, we removed any bigrams containing problematic closed class items such as case markers, conjunctions and pronouns. For another, we used a verb list containing all the conjugated forms of about 700 simple verbs in Urdu (15285 verb forms in total, from the verb conjugator in Raza 2011) in order to remove all those bigrams which we could identify as verb+verb combinations via this verb list.

After applying these steps, we were left with just 2154 candidate bigrams from the original 16033 bigram possibilities drawn from the raw corpus. These candidate bigrams still contained problematic items (see section 4), but identifying and removing them at this stage would have involved intense manual inspection and labor. We therefore decided to work with this list of bigrams for further analysis by different methods.

As a first analysis step, the association strength between the bigram members was computed using the Chi-Square (X^2) measure. This ensures that the more often one of our light verbs occurs with a certain word compared to all other words, the stronger the association is and the higher the bigram is ranked among the group of bigrams. The statistics were computed by means of the UCS toolkit.[9] We decided to use the X^2 association measure to determine the positive or negative association between the words in the bigram for two reasons. First, papers using comparatively sized corpora have reported encouraging results for similar experiments (Ramisch et al., 2008; Kizito et al., 2009; Hautli and Sulger, 2011). Second, initial manual comparison between bigram lists ranked according to all measures implemented in the UCS toolkit revealed the most convincing results for the X^2 test. Based on the ranking, we reduced the list of bigrams and discarded all bigram instances with either a negative X^2 or a frequency below 6 (see above). A negative X^2 value indicates a negative word association, i.e. the bigram members do not occur together very often in comparison to their frequency in other bigrams.

At this point, we still had our four separate lists of bigrams for the verbs *kɑr* 'do', *ho* 'be', *hu-* 'become' and *rɑk*^{h} 'put', but the lists are now of ranked bigrams. Manual inspection revealed that the top items in these lists now did contain N+V combinations of the type that Ahmed and Butt (2011) were looking at, among them also some of the nouns that were discussed in that paper. Our various steps of filtering and ranking thus did allow us to extract a list of strongly associated lexical items of the right type. The next step was to proceed to an analysis of our extracted data. Since our interest lay in the determination of possible classes of N+V CPS, we decided to run a clustering algorithm on our data.

3.3 Automatic clustering

Based on the filtered and ranked lists of bigrams from section 3.2, we investigated the occurrences of words with light verbs across the different types of bigram combinations, i.e. we combined the information of the four lists into one list recording the light verb behavior of every single word. That is, we had different classes of words: words occurring with all four light verbs or words with only *kɑr* 'do' and *ho* 'be', *hu-* 'become' and *rɑk*^{h} 'put' or words occurring with various combinations of these verbs. Table 2 shows an exemplary matrix of four nouns (*hasıl* 'achievement', *alan* 'announcement', *bat* 'talk' and *ʃuru* 'beginning' in that order) and their relative frequency of co-occurrence with the four light verbs, i.e. out of all occurrences of noun 1 (*hasıl* 'achievement') with one of the four light verbs, the relative frequency of it occurring with *kɑr* 'do' is 0.771.

ID	Noun	Rel. freq. with *kɑr*	Rel. freq. with *ho*	Rel. freq. with *hu*	Rel. freq. with *rɑk*^{h}
1	حاصل	0.771	0.222	0.007	0.000
2	اعلان	0.982	0.011	0.007	0.000
3	بات	0.853	0.147	0.000	0.000
4	شروع	0.530	0.384	0.086	0.000

Table 2: Relative frequencies of co-occurrence of nouns with light verbs

Note that although the motivation for our experiment stemmed from an interest in N+V CPS, our bigrams in fact contain all kinds of POS in combination with our set of four verbs. This is a direct and unavoidable consequence of using an untagged "raw" corpus.

Based on the pattern of relative co-occurrence with the four light verbs, the results were clustered automatically. This was done using a data mining platform developed at the University

[9] http://www.collocations.de; see Evert (2004) for documentation.

of Konstanz: KNIME.[10] We used a k-means clustering algorithm in order to assign each noun to a cluster. We found that a number of five clusters minimized the average distance between the nouns and the cluster centers. However, both the table of numbers as illustrated in Table 2 as well as the clusters were difficult to evaluate in this form. We therefore decided to experiment with visualization techniques as they are currently being pioneered in computational linguistics (e.g., Dörk et al. 2008; Collins 2010; Mayer et al. 2010). Before moving to a discussion of these techniques in section 4, we take a step back and consider the language particular knowledge we had to rely on so far.

3.4 Discussion: Language Particular Issues

As mentioned in section 3.2, our first pass at bigram extraction resulted in a large number of false hits. Upon some reflection on the structure of Urdu, this was only to be expected.

Urdu is a language which is not particularly morphologically complex (in comparison with Native American or Australian Aboriginal languages, for example), but it does use a significant amount of morphology. One unfortunate feature from the perspective of NLP is that the same material is used for several different purposes. For example, *-a*, *-i* and *-e* are morphemes used to mark gender and number on nouns, adjectives, verbs, participles as well as the genitive case. Additionally, there is significant homonymy with respect to frequently used words. For example, one that had a significant impact is the perfect masculine form of *kɑr* ‘to do’, *kɪya*, which is written the same way as the interrogative pronoun *kya* ‘what’. Another example is the perfect feminine form *ki* of *kɑr* ‘do’, which is written the same way as the complementizer ‘that’ as well as the feminine singular form of the genitive case marker.

The genitive case marker in general posed a problem. It is structurally a clitic (Butt and King, 2004) and is written as a separate word in Urdu. Given that it is a genitive case marker, it is naturally found adjacent to nouns. For us this meant that we extracted many bigrams which turned out to be collocations of the feminine singular genitive marker and a noun. We therefore decided to remove all instances of bigrams with *ki*. This meant that we probably lost many “good” bigram candidates, but we did not see a way of filtering out the “bad” instances of the genitive while keeping the good instances of the perfect feminine singular of *kɑr* ‘do’.

As already mentioned, our set of four verbs *kɑr* ‘do’, *ho* ‘be’, *hu-* ‘become’ and *rɑk*^{h} ‘put’ can be employed as simple verbs in Urdu as well and ‘do’, ‘be’, and ‘become’ can also function as auxiliaries. This meant that our initial bigram extraction netted many V+V sequences in which an item of our set of four verbs occurred as an auxiliary after a main verb. We dealt with this by employing a list of verbs along with all of their inflections that was constructed as part of the work done by Raza (2011). Since this encompassed a total of 15285 verb forms, having access to this already existing resource was invaluable.

With respect to verb conjugation, we also naturally normalized over the bigrams we looked for. That is, we looked for a total of 238 different forms of our set of four verbs, but normalized the different inflected versions to just the stem form for purposes of bigram storage. This is necessary for the task at hand as the inflectional variability would cause a futile co-occurrence analysis. However, it also means that we lose some information with respect to being able to understand whether a given bigram was really associated with a genitive marker or the feminine singular form of ‘do’, for example.

[10] http://www.knime.org

Finally, our initial list of bigrams contained instances of words with punctuation attached to them. We naturally removed these; however, there are other problems arising with respect to the Urdu script that are not dealt with as easily. For one, there are several different ways of spelling certain words in Urdu. One preprocessing step that we could have done is to run a normalization module across the corpus. However, this also requires specialized knowledge about the language/orthography and this source of errors was not large enough for us to take this step. Similarly, our bigram counts contain instances of words which have not been spaced correctly. The Urdu script is such that each letter has joined and non-joined versions. The conditions governing when to use a joined vs. non-joined version of a letter are fairly complex.

For example, take بات (*bat*) from Table 2. The first "letter" is a combination of a 'b' and an 'a' (the joined forms), the second letter is the non-joined version of 't'. In order to differentiate between spaces within words and spaces between two words, two different types of spaces have been defined. One is a normal space, the other is zero-width non-joiner (HTML code: &zwnj;). However, authors are not always consistent in their use, thus giving rise to errors in the corpus, which we again cannot deal with without adding time-consuming manual inspection coupled with deep language-particular knowledge to the process. These errors thus remain in the bigram list we use for the analysis detailed in the next section.

4 Analysis via Visualization

Visual Analytics is based on the tight coupling of algorithms for automatic data analysis and interactive visual components (Thomas and Cook, 2005; Keim et al., 2010). The idea is to exploit human perceptive abilities to support pattern detection (Card et al., 1999). This involves the mapping of data dimensions to eight *visual variables* (Bertin, 1983), namely *position* (two variables x and y), *size*, *value*, *texture*, *color*, *orientation* and *shape*. While some numerical data dimensions can be mapped directly to one visual variable, other data features may require complex layout algorithms that project a combination of multiple data dimensions to a combination of visual variables, e.g., to the combination of the two powerful positional variables x and y. Finally, a data analyst should be able to manipulate the visual display interactively for different perspectives on the data, following Shneiderman's Visual Information-Seeking mantra "Overview first, zoom and filter, then details on demand" (Shneiderman, 1996).

The purposes of visualizing data are manifold. On the one hand, visualizations can be used to achieve an overview of complex data sets. On the other hand, the visualization approach can serve as a starting point for interactively exploring data, ideally detecting hidden patterns. In addition, new hypotheses can be generated and existing hypotheses verified.

The visualization employed in this paper uses the visual variable *color* and encodes the relative frequency of occurrences with different light verbs. This relative frequency is mapped onto a linear saturation scale, i.e. the higher the relative frequency, the more saturated the color.

Figure 1 shows a reference visualization on the top, where the saturation is exemplified with the relative frequencies of 0.25, 0.5, 0.75 and 1.0 in the columns from left to right. Below the reference visualization, the relative frequencies of the data in Table 2 are encoded visually. The left-most color column shows the relative frequencies with *kɑr* 'do', the next columns show the relative frequencies with the light verbs *ho* 'be', *hu* 'become' and *rɑk*^{h} 'put', respectively. The lexical item on the left hand side is the transliterated version of the Urdu input in Table 2, using the transliterator of Bögel (2012), following the transliteration scheme by Malik et al. (2010).[11]

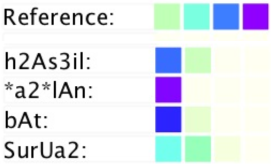

Figure 1: Visualization of the relative frequencies in Table 2

The visualization tool allows for an interactive exploration of the data in that the human investigator can scroll through the list of visually encoded N+V combinations, with the possibility of zooming in to get a detailed view on a restricted set of nouns, as well as zooming out to see a greater number of combinations and their overall behavior. Mousing over a colored box reveals the relative frequency of the N+V CP.

4.1 Visual Analysis: Round 1

The best clustering result for the clustering method described in section 3.3 was obtained with a specification of five clusters. These are numbered 0–4. Cluster 4 is a large cluster of about 1100 words. This cluster contained just those sets of words which occured only or almost only with *ho* 'be'. A manual inspection of this cluster showed that all of these words were either errors, false hits containing inflected verbs or words like 'what' and 'how many' or items like 'small', 'teacher', 'market', 'tomorrow' (the vast majority). The last category are items that occur in run-of-the-mill copula contexts like 'He is a teacher.'. This class is of no interest to us as there are no useful combinatory constraints to be discovered. We therefore decided to discard this cluster from our bigram list in its entirety.

Cluster 3 contained a small set of words that occurred with *hu* 'become' 100% of the time. All of these were false hits and were discarded. Similarly, Cluster 1 was a comparatively small cluster that consisted of words occurring mainly with *rɑk*^{h} 'put'. Manual inspection showed that all the words ocurring with *rɑk*^{h} 'put' a 100% of the time were false hits: the words were all objects of the main verb 'put' and not CPs of any kind. Again, we culled our bigram list to remove this set.

The other two clusters are more of a mixed bag. Cluster 2 has very large number of items that occur only with *kɑr* 'do'. Figure 2 shows the top part of Cluster 2 on the left. This contains nouns already identified as belonging to the ones combining in the type of N-V CP we are interested in. However, this class contains desired results as well as false hits which cannot be separated from

[11] Short vowels are not encoded in the Urdu script, the transliterator puts a default "*" in places where a short vowel is expected. The ambiguous characters *vao* and *ye* (consonant or vowel) are represented with ⟨vao⟩ and ⟨ye⟩, respectively. In case of entries with '???', the automatic transliteration could not find an adequate transliteration — this can be due to typing errors or unusual vowel/consonant combinations of English loan words. We needed to use a transliterated version of our bigram lists as the visualization tool we used could not deal with UTF-8 input.

one another on a visual (or purely numerical basis). Culling down this cluster would involve intense manual labor, which we decided to forego.

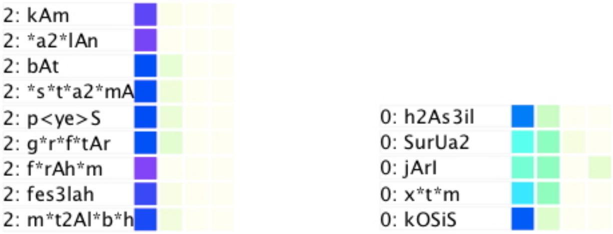

Figure 2: Visualization of the top of Cluster 2 (left) and Cluster 0 (right)

Similarly, Cluster 0 contains desired results as well as false hits. Again, these are difficult to discern visually, though there are clearer subclusters because this cluster contains those words which occur with all the four light verbs (at different levels of frequency). In Figure 2 *hasil* ‘achievement’, *ʃuru* ‘beginning’, and *koʃɪʃ* ‘struggle’ are the kinds of nouns we are interested in. On the other hand, *xɑtɑm* ‘finished’ and *jari* ‘continued’ are adjectives.

As part of our quick visual inspection, we also noted several types of unusual patterns found in the clusters, such as the ones illustrated in Figure 3 for the verb *ʊṭ*^{h}*a* ‘raise’. These often turned out to be false hits. The visualization thus allowed for a quick and easy identification of further errors. All unusual patterns that were identified as errors via this method of analysis were also removed from our bigram list. Detecting these types of errors would have been neigh impossible without an easily accessible visual representation.

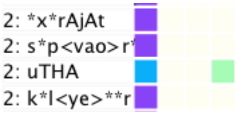

Figure 3: Visually prominent pattern of a false hit

All in all, after the manual selection process, we were left with only 1090 instances of bigrams. This list still contains false hits, but removing these would take intensive manual labor. We therefore decided to rerun our automatic clustering algorithm on the (partially) cleaned data.

4.2 Visual Analysis: Round 2

The new clustering again yielded the best results with a specification of five clusters. The visualization shows that the cleaning and culling described in the previous section has had a positive effect, but that even the remaining 1090 bigrams still contain a good amount of false hits which would need to be culled on the basis of intense manual inspection. Nevertheless, the overall results are encouraging as different types of nouns are indeed being distributed over the different clusters.

As shown in Figure 4, Cluster 0 still contains *hasil* ‘achievement’ and *koʃɪʃ* ‘struggle’. Items from the previous Cluster 2 also now appear here: *bat* ‘talk’, *ɪstemal* ‘use’ and *peʃ* ‘presentation/happening’. On the other hand, *ʃuru* ‘beginning’, *xɑtɑm* ‘finished’ and *jari* ‘continued’ are now in Cluster 3, which mainly consists of adjectives, but also a few of the nouns we are interested in. While ADJ-V CPS were not the target of this paper, this cluster contains potentially useful information with respect to ADJ+V CPS.[12]

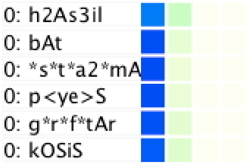

Figure 4: Visualization of the top of Cluster 0

The same is true of Cluster 1: it consists of mainly adjectives, but there is a smattering of some of the nouns we are interested in as well as a number of spatial terms including postpositions.

Cluster 4 contains words which occur almost exclusively with *rɑk*^{h} ‘stay’ and this set still does not contain much that is of interest within the scope of this paper. On the other hand, Cluster 2 contains many nouns which occur in N-V CPS and indeed, contains many of the nouns that Ahmed and Butt (2011) identified as belonging to Class 2. A comparison of our results with that of Ahmed and Butt (2011) shows that while we did not find all of the nouns used in Ahmed and Butt, of the nouns that we did find most are Class B nouns (no dative subjects allowed). These are distributed over Clusters 0 and 2, while the Class A nouns (full range of light verb use) are found in Clusters 0, 1 and 3 (no Class C nouns were found in our study).

5 Results and Discussion

In this paper, we set out to see if we could find fine-grained information about the distributional properties of nouns in N+V CPS by extracting bigrams from a large “raw” corpus of Urdu. In addition to finding target instances of N+V combinations, our work has also resulted in lists of possible ADJ+V combinations and, in particular, the realization that P+V combinations are highly frequent. The prepositions involved are, for instance, ‘front’, ‘back’, ‘on’ and ‘beside’ and in combination with our set of verbs mean something like ‘place on/beside/front/back’ (with ‘do’ and ‘put’) or ‘be/become on/beside/front/back’ (with ‘be’ and ‘become’). Given that these combinations are highly frequent, it is imperative that more be understood about P+V.

Unfortunately, this cannot be accomplished via our data because we actually have a problem with *data sparsity*. Manual inspection of some of the CPS we found showed that not all possible combinations of N/ADJ/P+V were in fact attested in the corpus. That is, our bigram lists will need to be complemented by manual native speaker work (traditional lexicography).[13]

We also find it remarkable that out of a corpus of 7.9 million words, we are at this point left with only 1090 bigrams (and are aware that a portion of these bigrams would still need to be culled as they represent errors or false hits). This problem of data sparsity could perhaps be ameliorated by using a different type of corpus, but we suspect that for our type of enterprise, data sparsity might be a problem regardless of how large or different a corpus is chosen — one cannot guarantee that all possible combinations will indeed be attested in any given corpus. On

[12] But note that Cluster 0 also contains adjectives, for instance *gırıftar* ‘arrested’ in Figure 4.

[13] One reviewer would like an indication of how many valid CPS we might have missed. Since we have no idea how many CPS were indeed contained in our corpus, we cannot say. And as N-V CPS are combinatorily dynamic, there are potentially infinitely many of these. All we can say is how many of the potentially infinitely many combinations we did find.

the other hand, one result that is very clear is that any word that occurred only in conjunction with *ho* 'be', *hu-* 'become' or *rak*^{h} 'put' was in fact not one that would occur in CPS. This means that one can exclude those words from candidate lists of nouns for N+V CPS in future work.

Our paper also makes contributions with respect to two methodological points. In order to be analyze our data more perspicuously, we experimented with new methods from *visual analytics*. This experimentation was successful as it allows for quick visual analysis of our data sets, which in turn also enables a non-labor-intensive way of further cleaning and culling our data. In particular, the visual analysis allowed us to be able to quickly assimilate information about complex interrelationships in our data: which types of verbs does the word in question occur with with what level of frequency and how does this compare to other words in the list?

The visualization component proved to be an essential part of our data analysis. Another essential component turned out to be our language-particular knowledge and access to language-particular resources. The idea that we could work more or less "blindly" with a large corpus did not pan out. Rather, at every step we needed to be aware of language particular issues with respect to orthography, morphology and syntax. We could not have done our work without the use of a list of all the conjugated forms of 700 simple verbs (Raza, 2011) or a transliterator (Bögel, 2012) (needed to massage the input for the visualization component). We thus conclude that while we were partly successful in achieving what we set out to do, our work would have been able to be yield more precise results if we had access to standardized language-particular resources. On the other hand, our "blind" extraction of patterns yielded new insights into what kinds of CPS are highly frequent in newspaper corpora in addition to N+V CPS and which therefore need to be paid attention to with respect to NLP applications and the creation of language particular resources. This pertains in particular to P+V and ADJ+V combinations, about which not much is known either from a computational or from a theoretical perspective.

References

Ahmed, T. and Butt, M. (2011). Discovering Semantic Classes for Urdu N-V Complex Predicates. In *Proceedings of the international Conference on Computational Semantics (IWCS 2011)*, pages 305–309.

Ahmed, T., Butt, M., Hautli, A., and Sulger, S. (2012). A Reference Dependency Bank for Analyzing Complex Predicates. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3145–3151, Istanbul, Turkey. European Language Resources Association (ELRA).

Bertin, J. (1983). *Semiology of Graphics*. University of Wisconsin Press.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D., and Xia, F. (2009). A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189, Suntec, Singapore. Association for Computational Linguistics.

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

Bögel, T. (2012). Urdu – Roman Transliteration via Finite State Transducers. In *Proceedings of FSMNLP'12*.

Butt, M. (1995). *The Structure of Complex Predicates in Urdu*. Stanford: CSLI Publications.

Butt, M. (2010). The Light Verb Jungle: Still Hacking Away. In Amberber, M., Harvey, M., and Baker, B., editors, *Complex Predicates in Cross-Linguistic Perspective*, pages 48–78. Cambridge University Press, Cambridge.

Butt, M. and King, T. H. (2004). The Status of Case. In Dayal, V. and Mahajan, A., editors, *Clause Structure in South Asian Languages*, pages 153–198. Kluwer Academic Publishers, Berlin.

Card, S. K., Mackinlay, J. D., and Shneiderman, B., editors (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Chakrabarti, D., Sarma, V. M., and Bhattacharyya, P. (2008). Hindi Compound Verbs and their Automatic Extraction. In *Proceedings of COLING 2008*, pages 27–30.

Chakraborty, T. and Bandyopadhyay, S. (2010). Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule-Based Approach. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 72–75.

Collins, C. (2010). *Interactive Visualizations of Natural Language*. PhD thesis, University of Toronto.

Das, D., Pal, S., Mondal, T., Chakraborty, T., and Bandyopadhyay, S. (2010). Automatic Extraction of Complex Predicates in Bengali. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 37–45.

Dörk, M., Carpendale, S., Collins, C., and Williamson, C. (2008). VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of the IEEE Conference on Information Visualization (InfoVis '08))*, 14(6):1205–1213.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, IMS, University of Stuttgart.

Hautli, A. and Sulger, S. (2011). Extracting and Classifying Urdu Multiword Expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Langauge Technologies (ACL-HLT '11): Student Session*, pages 24–29.

Hautli, A., Sulger, S., and Butt, M. (2012). Adding an Annotation Layer to the Hindi/Urdu Treebank. *Linguistic Issues in Language Technology*, 7(3):1–18s.

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

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

Keim, D. A., Kohlhammer, J., Ellis, G., and Mansmann, F., editors (2010). *Mastering The Information Age - Solving Problems with Visual Analytics*. Goslar: Eurographics.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). Visual Analytics: Scope and Challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Lecture Notes in Computer Science, pages 76–91. Springer.

Kizito, J., Fahmi, I., Sang, E. T. K., Bouma, G., and Nerbonne, J. (2009). Computational Linguistics and the History of Science. In Dibattista, L., editor, *Storia della Scienza e Linguistica Computazionale*. FrancoAngeli.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation*. The University of Chicago Press.

Malik, M. K., Ahmed, T., Sulger, S., Bögel, T., Gulzar, A., Raza, G., Hussain, S., and Butt, M. (2010). Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*.

Mayer, T., Rohrdantz, C., Butt, M., Plank, F., and Keim, D. (2010). Visualizing Vowel Harmony. *Journal of Linguistic Issues in Language Technology (LiLT)*, 4(2).

Mohanan, T. (1994). *Argument Structure in Hindi*. CSLI Publications.

Mukerjee, A., Soni, A., and Raina, A. M. (2006). Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties (MWE '06)*, pages 28–35.

Ramisch, C., Schreiner, P., Idiart, M., and Villavicencio, A. (2008). An Evaluation of Methods for the Extraction of Multiword Expressions. In *Proceedings of the Workshop on Multiword Expressions: Towards a Shared Task for Multiword Expressions (MWE '08)*, pages 50–53.

Raza, G. (2011). *Subcategorization Acquisition and Classes of Predication in Urdu*. PhD thesis, University of Konstanz.

Schulte im Walde, S. (2009). The Induction of Verb Frames and Verb Classes from Corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–, Washington, DC, USA. IEEE Computer Society.

Sinha, R. M. K. (2009). Mining Complex Predicates in Hindi Using a Parallel Hindi-English Corpus. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pages 40–46.

Thomas, J. J. and Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center.

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