Inferring Subcat Frames of Verbs in Urdu

Ghulam Raza

Universität Konstanz
78457 Konstanz, Germany
ghulam.raza@uni-konstanz.de

Abstract

This paper describes an approach for inferring syntactic frames of verbs in Urdu from an untagged corpus. Urdu, like many other South Asian languages, is a free word order and case-rich language. Separable lexical units mark different constituents for case in phrases and clauses and are called case clitics. There is not always a one to one correspondence between case clitic form and case, and case and grammatical function in Urdu. Case clitics, therefore, can not serve as direct clues for extracting the syntactic frames of verbs. So a two-step approach has been implemented. In a first step, all case clitic combinations for a verb are extracted and the unreliable ones are filtered out by applying the inferential statistics. In a second step, the information of occurrences of case clitic forms in different combinations as a whole and on individual level is processed to infer all possible syntactic frames of the verb.

1. Introduction

The set of arguments a verb takes is called its subcategorization frame (SCF). It is possible that a single verb takes a variable set of arguments in different situations. In that case the verb is said to have more than one subcat frame. Knowing about all possible frames of a verb is very important in natural language processing tasks such as computational grammar development, compilation of comprehensive dictionaries, machine translation, information retrieval and parsing. For example, Briscoe and Carrol (1993) parsed unseen test data on a parsing system utilizing a lexicalist grammatical framework and noted that half of the parse failures were due to inaccurate subcategorization information. To project accurate syntactic structure of any language, most of the grammar formalisms today require comprehensive lexicons having accurate information about the predicate subcategorization.

For English, many subcategorization lexicons have been developed manually. Among them the largest one is VerbNet (Kipper-Schuler, 2005) which has been made on the basis of Levin (1993)'s verb classification. Many efforts have also been made for auto building of such lexicons for English. These efforts comprise the acquisition from raw English corpus (Brent, 1993) and from an annotated part of speech (POS) tagged corpus (Manning, 1993; Ushioda et al., 1993) and from a corpus parsed partially or fully (Briscoe and Carrol, 1997; Kinyon and Prolo, 2002; O'Donovan et al., 2005). No complete subcat lexicon for Urdu has been developed manually or automatically till now. There exist many a dictionaries of Urdu and recently, the Urdu Lughat Board in Pakistan has published twenty one volumes of a large Urdu dictionary. However, all these dictionaries lack subcategorization information. Since we do not have more refined resources at hand for Urdu, we have developed a system that can infer subcat frames of verbs from an unannotated corpus. Our system differs from the previous work on other languages in that we recognize verbs in the corpus by matching corpus words to members of conjugation set of a given verb and acquire frame types indirectly from clitic combinations.

In this system, we test for any verb how many of the possible 64 case clitic plus complementizer combinations (CLCs) are valid for it. These combinations are based on the presence or absence of five case clitics and one complementizer (2^6=64) in the candidate sentence of the target verb. Before testing a verb for these combinations, necessary screening is made for candidate sentences of the target verb. The screened sentences are delimited to the scope of the target verb and possible spurious case phrases are ignored. Using the hypothesis testing technique, different case clitic combinations are validated. By further processing the information of the valid case clitic combinations for a given verb, its subcategorization frames are inferred. The paper is organized as follows: Section 2 introduces some morpho-syntactic particularities of Urdu, Section 3 provides a description of the system, in Section 4 results and evaluations are given and Section 5 concludes the paper with directions for future work.

2. Morpho-syntactic Particularities of Urdu

Urdu is a verb-final language and the verb's arguments might fall in any order, so it is generally called a free word order language. The core arguments or complements of the verb in Urdu are case marked (Butt and King, 2005). Adjuncts are either case marked phrases or postpositional phrases. For distinguishing between arguments and adjuncts see (Pollard and Sag, 1987) and (Meyers et al., 1994). In this work, however, we have not made a distinction between arguments and adjuncts.

For Urdu, being a case-rich language, it apparently seems trivial to recognize arguments of predicates based on case clitics clues but the complexity of the case system and free word order nature of the language make the task difficult. The following five issues are challenging for automatic acquisition of subcategorization frames of verbs in Urdu from a raw corpus.

2.1. Absence of Unique Case Clitic Forms

There are several different case clitic forms: NULL, *ne*, *ko*, *se*, *meñ*, *par* in Urdu. However, there is not always a one to one correspondence between case clitic form and the case finction. For example, the same clitic form *ko* can

mark nouns for different case functions: accusative, dative, locative and temporal. Consider the example sentences in (1)-(4).[1]

(1) alii=ne nidaa=**ko** bulaa-yaa
Ali=Erg Nida=Acc call-Perf
'Ali called Nida.' (Accusative)

(2) alii=ne nidaa=**ko** xat lik^h-aa
Ali=Erg Nida=Dat letter.M.3Sg write-Perf.M.3Sg
'Ali wrote a letter to Nida.' (Dative)

(3) alii raat=**ko** aa-yaa
Ali.Nom night=Temp come-Perf.M.Sg
'Ali came at night.' (Temporal)

(4) alii g^{h}ar=**ko** ga-yaa
Ali.Nom home=Loc go-Perf.M.Sg
'Ali went home.' (Locative)

For a detailed description of *ko* in Urdu, see Ahmed (2006). Most of other case clitics also show a similar multifunctional distribution.

2.2. Multicase of Grammatical Functions

There is not always a unique case for grammatical functions in Urdu. The direct object of a verb can be a nominative or accusative (Butt and King, 2005), for example, consider the following sentences:

(5) alii=ne seb k^{h}aa-yaa
Ali=Erg apple.Nom.M.3Sg eat-Perf.M.3Sg
'Ali ate an apple.'

(6) alii=ne seb=ko k^{h}aa-yaa
Ali=Erg apple=Acc.M.3Sg eat-Perf
'Ali ate the apple.'

The direct object *seb* 'apple' in (5) is null-marked for nominative case and it is *ko*-marked for accusative case in (6). Specific direct objects in Urdu are marked accusative (Butt, 1993). The subject in Urdu also is not associated with a single case. In Table 1 clitics forms, cases and their associations with grammatical functions are listed.

2.3. Free Word Order

Fixed order in a language like English is useful in that the order itself provides clues for recognizing the arguments of a verb in the sentence. However, being a case-rich language, Urdu is a free word order language. The verb in a sentence usually comes last and its arguments are put in any order before it. For example, the arguments of the verb

[1]In the transcription scheme consider 'a' , 'i', 'u' as short vowels and 'aa' , 'ii', 'uu' as long vowels. The equal symbol '=' marks a clitic boundary. Glosses used in this paper are as follows: 1, 2, 3 stand for 1st, 2nd and 3rd person, respectively; Nom=Nominative; Acc=Accusative; Dat=Dative; Erg=Ergative; Temp=Temporal; Loc=Locative; F=Feminine; and M=Masculine; Comp=Complementizer; Relp=Relative Pronoun; Perf=Perfective; Imp=Imperative; Subjn=Subjunctive; SConj=Subordinating Conjunction.

Clitic Form	Case	Gramaatical Function
NULL	Nominative Locative Temporal	Subject, Direct Object Location Time
ne	Ergative	Subject
ko	Accusative Dative Temporal Locative	Direct Object Subject, Indirect Object Time Location
se	Instrumental Comitative Ablative Locative Ability	Instrument Object Indirect Object Location Subject
meñ	Locative Temporal	Location Time
par	Locative Temporal	Location Time

Table 1: Case and Grammatical Functions

likh 'to write' in (2) can be arranged in different orders as in (7)-(11). The truth-conditional meaning of the sentence remains the same, regardless of the order of the verb's arguments.

(7) alii=ne xat nidaa=ko lik^h-aa
Ali=Erg letter.M.3Sg Nida=Dat write-Perf.M.3Sg
'Ali wrote a letter to Nida.'

(8) nidaa=ko alii=ne xat lik^h-aa
Nida=Dat Ali=Erg letter.M.3Sg write-Perf.M.3Sg
'Ali wrote a letter to Nida.'

(9) nidaa=ko xat alii=ne lik^h-aa
Nida=Dat letter.M.3Sg Ali=Erg write-Perf.M.3Sg
'Ali wrote a letter to Nida.'

(10) xat alii=ne nidaa=ko lik^h-aa
letter.M.3Sg Ali=Erg Nida=Dat write-Perf.M.3Sg
'Ali wrote a letter to Nida.'

(11) xat nidaa=ko alii=ne lik^h-aa
letter.M.3Sg Nida=Dat Ali=Erg write-Perf.M.3Sg
'Ali wrote a letter to Nida.'

The case and animacy features of nouns predict their grammatical functions in a sentence in Urdu. The verb *likh* 'to write' is a transitive verb and takes an ergative subject in sentences (7)-(11). Urdu is a split ergative language (Anderson, 1977). The ergative subject is taken by verbs with some conditions of transitivity, tense, aspect, choice of auxiliary and volitionality (Butt, 2006). The subject in (7)-(11) will be marked nominative in case the tense is present. In that case the two arguments of the predicate that is *alii* 'Ali' and *xat* 'letter' will be nominative and the subject in such a case could be recognized due to its animacy feature. The direct auto recognition of subject in sentences in an Urdu

corpus is possible if in the corpus, nouns are annotated for such type of features like animacy and abstractness and the verbs are annotated by their types unlike the fixed order languages where only the position of a noun tells about its grammatical function.

2.4. Multifunction of Complementizer Form

To recognize complementizer clause argument of a verb in Urdu by identifying complementizer form *kih* in some sentence is not straightforward as the same complementizer form has some other functions of relative pronoun and conjuncts also.

(12) ham=ne maan-aa **kih** ...
We=Erg.1P agree-Perf that.Comp
'We agreed that ...' (Complementizer)

(13) kitaab jo/**kih** achii ho, laa-o
book.F.3Sg that.Relp good be.Subjn bring-Imp
'Bring the book that is good.' (Relative)

(14) alii cal-aa hii thaa **kih**
Ali.Nom walk-Perf.M.Sg just be.Past.Sg that
baarish barasne lag-ii
rain.Nom.F.3Sg shower.Inf start-Perf.F.3Sg
'It started raining when Ali just started walking.'
(Temporal)

(15) alii avval aa-yaa yaa/**kih** nidaa?
Ali.Nom first come-Perf.M.Sg or Nidaa
'Ali stood first or Nida?' (Conjunction)

The form *kih* in sentence (12) is used as a canonical complementizer, in (13) it is used as a relative pronoun, in (14) as a temporal marker and in (15) as a conjunction. The same form in Urdu poetry is sometimes used inplace of *balkih* 'but' and *kiyoñkih* 'because'. The important point here is that with use of some adverbs *kih* can appear with any verb in Urdu. It is very hard to filter the canonical complementizer clause for verbs.

2.5. Attachment Ambiguities

Syntactically not only the verbs subcategorize for arguments in Urdu but the nouns and adjectives do too, as in many other languages. These nouns and adjectives in Urdu usually are derived from verbal stems. To automatically determine whether some case phrases actually are part of the noun or the adjective modifying the noun or the verb in the sentences of an unannotated corpus is not an easy task. This problem is like the PP-attachment problem in English, German and other languages.

(16) nidaa=ne [zukaam=se bacaao]=kii
Nida=Erg.F.3Sg flu=Abl.3Sg saving=Gen.F.3Sg
davaaii xariid-ii
medicine.F.3Sg purchase-Perf.F.3Sg
'Nida purchased medicine for saving from flu.'

(17) nidaa=ne baazaar=se zukaam=kii
Nida=Erg.F.3Sg market=Abl.3Sg flu=Gen.F.3Sg
davaaii xariid-ii
medicine.F.3Sg purchase-Perf.F.3Sg
'Nida purchased medicine for flu from the market.'

In (16) *davaaii* 'medicine' is the nominative argument of the verb *xariid* 'buy'. This noun has a genitive specifier *bachaao* 'saving' that itself takes a *se*-marked argument. So the case marked arguments found in the sentence are not always the arguments of the main verb in the sentence. In (17), however, the *se*-marked noun is not part of any noun. Here it should be considered as an adjunct of the verb *xariid* 'to buy'.

3. Description of the System

The corpus which we have experimented with was collected from Urdu websites. The major data of the corpus was from Urdu newspaper Roznama Jang website and BBC Urdu website. In a first step the corpus was cleaned. There are some Urdu characters that were composed by two symbols in Unicode before. But later all Urdu alphabets were assigned their own codes. Such differences were found in the corpus and were normalized. Different spellings of some key words used in the corpus were also unified. The corpus had many segmentation errors too. The words at the end of most sentences where a verb is found were not properly separated. To make it possible to have maximum possible candidate sentences for the target verb, the corpus was segmented. In the segmenting module, an Urdu lexicon of more than 60, 000 words was used to identify correct words.

The corpus that was used in our system after preprocessing has 276825 sentences. The five components were used in sequence in our system to infer frames of a specific verb from a preprocessed unannotatd Urdu corpus. These are discussed in the following subsections.

3.1. Verb Conjugator

Any verbal root in Urdu can have up to four stems: for intransitive, transitive, causative and indirect causative. For example, the root dik^h in Urdu has four stems dik^h 'to appear', dek^h 'to see', dik^haa 'to show' and dik^hvaa 'to make some one show'. Most of the stems are conjugated in a regular pattern for tense, aspect, number and gender. Any regular stem can have upto 16 conjugations, among them three are for infinitives.

There are about 700 verbal roots and about 1200 verbal stems in Urdu that represent basic verbs in Urdu. It should be noted here that not all roots are inflected for four stems, for example, the root lik^h has only three stems: the transitive stem, lik^h 'to write', the causative stem, lik^haa 'to cause some one write' and the indirect causative stem, lik^hvaa 'to cause some one make some one else write'. Sometimes direct causative stems and indirect causative stems are alternately used. Verbs other than basic verbs are multi word expressions formed by combining some other words with the basic verbs and are called complex predicates. In Urdu noun-verb, verb-verb, and adjective-verb complex predicates are found (Butt, 1995). In this system Verb Conjugator module can conjugate all regular and irregular basic verbs. The conjugator takes the stem form of a verb and generates all of its possible inflected forms. The

conjugator has been implemented by just incorporating the simple rules for inflections of the verb stems of different morphological classes in Urdu.

3.2. Candidate Finder and Scope Delimiter

The good candidate sentences for a target verb are extracted form the corpus in many steps. First, all the sentences where any of the conjugations of the target verb are found are extracted. Sometimes one conjugation of the target verb is either form-identical to some noun or some conjugation of another verb or some functional word. In such a case those conjugations are not considered. After this initial step the screening and delimiting is made in three phases.

3.2.1. Initial Screening Phase

In this phase the candidate sentences are examined for the position of the target verb in them. The following three tests are performed.

Screen 1: *Exclude those sentences where some other verb is found just before the target verb.*

In that case the target verb would have been used as a light or an auxiliary verb in the sentence rather as a main verb. Consider, for example, the verb *ḍaal* 'to put' in (18) and (19).

(18) alii=ne saañp maar *ḍ*aal-aa
Ali=Erg.M.Sg snake.M.3Sg kill put-Perf.M.3Sg
'Ali killed a snake.'

(19) alii=ne paanii=meñ kankar *ḍ*aal-aa
Ali=Erg.M water=Loc pebble.M.3Sg put-Perf.3Sg
'Ali put a pebble in water.'

In (18) another verb *maar* 'to kill' precedes the target verb *ḍaal* and the target verb is functioning as the light verb of a complex predicate, rather than as a main verb. In (19) no other verb precedes the target verb and hence, the locative marked argument of the target verb is there.

Screen 2: *The target verb should not be found in the subordinating clause of the sentence.*

The problem is that in this case some of the arguments of the verb could be controlled externally. So remove all the sentences where the target verb is followed by subordinating conjunctions *kar*, *key*, *huaa/hue/huii*. Consider (20) and (21) with the target verb *bʰaag* 'to run'.

(20) alii=ne bʰaag kar patañg pakṙ-ii
Ali=Erg.3Sg run SConj kite.F.3Sg catch-F.3Sg
'Ali caught a kite by running.'

(21) alii bʰaag-aa
Ali.Nom.M.3Sg run-Perf.M.3Sg
'Ali ran.'

The verb *bʰaag* 'to run' is an intransitive verb that does not take an ergative subject but a nominative subject as in (21). In (20) the ergative subject is due to the main clause verb *pakṙ* 'to catch'. If the sentence like (20) is not blocked to be a candidate sentence of the target verb *bʰaag* 'to run', then there is a chance that the system incorrectly infers an ergative subject for this verb.

Screen 3: *The target verb should not precede the* jaanaa *auxiliaries*.

The passive clause of a verb in Urdu is formed by the perfect participle of the verb followed by *jaanaa* 'to go'. The verb *cal* 'to walk' is an exception to this rule. In passive construction agent of the verb usually is demoted. Therefore such sentences should not be included in the set of candidate sentences of the target verb.

(22) kʰaanaa kʰaa-yaa ga-yaa
Meal.M.3Sg eat-Perf.M.3Sg go-Perf.M.3Sg
'The meal was eaten.'

(23) alii=ne kʰaanaa kʰaa-yaa
Ali=Erg.M.3Sg Meal.Nom.M.3Sg eat.M.3Sg
'Ali ate the meal.'

In the passive construction (22) the subject of the verb *kʰaa* 'to eat' is demoted and left out, while (23) gives the information that this verb takes ergative subject.

Some verbal stems with *jaanaa* verb form multiword verbs and this phenomenon in Urdu affects the argument structure of the main verb. The verb *kʰaa* 'to eat' that usually takes an ergative subject when it forms a multiword verb *kʰaa jaana* 'to eat' then it takes always a nominative subject as in (24). Here again the target verb is forming a complex predicate but this time it is acting as main verb rather than as a light verb and the case of its subject argument has changed.

(24) alii kʰaanaa kʰaa ga-yaa
Ali.Nom.3Sg Meal.M.3Sg eat.Stem go-Perf.3Sg
'Ali ate the meal.'

As the argument structure of the main verb changes in both cases: when *jaanaa* comes as a passive auxiliary or as an aspectual auxiliary after the main verb, therefore, we exclude all such sentences from the candidates list. There are other aspectual auxiliaries also that change the canonical argument structure of main verbs in Urdu. But in this work only *jaanaa* aspectual auxiliary has been considered. But in future an extensive list of aspectual auxiliaries, that change and that do not change canonical argument structure of predicates, will be worked out.

3.2.2. Delimiting the Scope Phase

A single sentential clause can be composed of more than one clause, for example: coordinating clause, relative clause etc. Each clause can have a different main verb. Delimiting of the candidate sentences to the scope of the target verb is made by identifying the following three patterns.

Pattern 1:

Verb + *aor* // Verb + *yaa* // Verb + *nah*

The words *aor* 'and', *yaa* 'or' and *nah* 'not' are conjunctions that can conjoin two nouns or two clauses. If any of the above patterns is found in the forward direction after the target verb index, then the sentence should be delimited to that point because in such a case another sentential clause might have been conjoined. This pattern is not tested before the target verb in this phase. If we discard the sentence

up to that pattern before the target verb then some of arguments of the target verb clause could be deleted due to the deleted first conjunction clause. If such a pattern is found before the target verb then we exclude such a sentence in the final screening phase.

Pattern 2:

Verb + any other conjunct or complementizer

When such a pattern is found before or after the target verb then the sentence is delimited to those points because here again there is a sentential conjunction. If the complementizer is found after the target verb then it is retained because it is a signal for the complementizer clause argument of the target verb else it should be deleted as in such a case it is not a signal that the target verb is taking complementizer clause argumetn.

Pattern 3:

Verb + any relative pronoun

The sentence, this time too, is delimited on both sides of the target verb. If the relative pronoun is found after the target verb then it should be deleted as in such case it is the argument of some verb in the coming sentential clause, else it is retained as it is in the domain of the target verb. The objective relative pronoun is an alternate signal of the presence of *ko* clitic in the sentence.

3.2.3. Final Screening Phase

Although we have delimited the sentences to the scope of the target verb in the previous phase, still there might be other participle adjectives with their arguments within the scope of the target verb. So to avoid counting a non-argument of the verb as an argument of the verb, we ignore such sentences.

Screen 4: *Ignore the sentences where any other verb before the target verb is found or after it after the light and/or auxiliary verbs*.

Consider the target verb *uṭʰaa* 'to pick' in (25)-(27).

(25) alii=ne mez=par rakʰe
Ali=Erg.M.Sg table=Loc.3Sg place-Perf.Obl
qalam=ko uṭʰaa-yaa
pen=Acc.M.3Sg pick-Perf
'Ali picked the pen placed on the table.'

(26) alii=ne mez=par qalam=ko
Ali=Erg.M.Sg table=Loc.3Sg pen=Acc.M.3Sg
rakʰ-aa aor uṭʰaa-yaa
place-Perf and pick-Perf
'Ali placed the pen on the table and picked.'

(27) alii=ne qalam=ko uṭʰaa-yaa
Ali=Erg.M.3Sg pen=Acc.M.3Sg pick-Perf.M.3Sg
'Ali picked the pen.'

In (25) *rakʰe* 'placed' is a participle adjective and takes a locative marked argument. In (26) the verb *rakʰ* 'to place' is a main verb and takes a locative marked argument. Such sentences are excluded so that the locative marked argument should not be inferred for the target verb uṭʰaa 'to pick', that does not take such argument as in (27).

A limited number of adjectives are used in Urdu that are derived from Arabic verbal stems and do take case marked arguments. A few examples of such adjectives are *laahiq* 'attached', *mustasnaa* 'excepted', *shaamil* 'included', *mushtamil* 'comprised' taking *ko*, *se*, *meñ* and *par* marked arguments respectively. If such adjectives are found in candidate sentences, then the clitics associated with their arguments are ignored while acquiring CLCs of the target verb.

3.3. CLCs Acquisition

Once the candidate sentences have been found, screened and delimited to the scope of the target verb, the counts for the different types of case clitics and complementizer combinations (CLCs) are computed. The type of a CLC is distinguished by the value of a six bit vector. These bits are for the five clitic forms *ne*, *ko*, *se*, *meñ*, *par* and one complementizer form *kih* from the most significant bit to the least significant bit, respectively. For example, the vector value 110000 represents that CLC of the target verb for a sentence where only *ne* and *ko* clitics are found and others are absent. Depending upon the absence or presence of five clitics and one complementizer in the sentence, 64 CLC types are possible.

The counts of different CLCs types for the verb *uṭʰaa* 'to pick' from its 248 final candidate sentences are given in Table 2. CLCs of zero count are not mentioned.

CLC Type (*ne+ko+se+meñ+par+kih*)	CLC Frequency
000000	72
100000	75
010000	08
001000	25
000100	19
000001	02
110000	04
101000	07
100100	14
100001	05
011000	01
010100	04
001100	04
111000	02
101100	02
100101	03
111100	01

Table 2: Types of CLCs and their counts recognized for the verb *uṭʰaa* 'to pick' with 248 candidate sentences

3.4. CLCs Filtering

The CLC types recognized in the last step may have some noise due to some wrong hits for the clitics. For filtering usually a null hypothesis (H_0) is formulated, which is assumed true unless there is some evidence to the contrary. An alternate hypothesis (H_1) is accepted in case some evidence proves H_0 false. Four methods reported in the literature can potentially be used for filtering out the unreliable

CLCs.

3.4.1. Relative Frequencies

One simple method is that relative frequencies of different CLC types be computed and if relative frequency of a CLC is higher than some threshold value then CLC is accepted else it is rejected. Lapata (1999) used this method to filter subcat frames for diathesis alternation detection. She used the COMLEX Syntax dictionary (Grishman et al., 1994) to compute threshold frequency for each SCF from the the frequencies of SCFs in the dictionary. She reported that this method produced slightly better results than binomial filter discussed in Section 3.4.4. Korhonen et al. (2000) also showed that this method performed better than binomial filter.

As we had no reference resource for setting a cut-off on the relative frequencies of CLCs, this method was not selected for filtering in our system.

3.4.2. Log Likelihood Ratio

The Log likelihood Ratio (LLR) reflects the difference between the observed and the expected distribution. As a null hypothesis, it could be assumed that the distribution of CLC is independent of the distribution of a verb that is *p(CLC|verb) = p(CLC)*. The LLR statistic verifies or rejects this hypothesis. The greater the LLR value, more likely it is that the CLC is associated with the verb. If LLR is greater than some threshold value, then the null hypothesis is rejected. To calculate the LLR for each verb and CLC combinations four counts are required.

k_1 = Number of times CLC occurs with the verb
n_1 = Number of occurrences of the verb
k_2 = Number of times CLC occurs with any other verb
n_2 = Number of occurrences of other verbs

Using these numbers, the following three probabilities are computed:

$$p_1 = \frac{k_1}{n_1}, p_2 = \frac{k_2}{n_2}, p = \frac{k_1 + k_2}{n_1 + n_2}$$

Assuming that these probabilities are binomially distributed, the LLR statistic as given by Dunning (1993) is computed as:

$$-2\log\lambda = 2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)]$$

where,

$$\log L(p, n, k) = k \log p + (n-k)\log(1-p)$$

Sarkar and Zeman (2000) compared the LLR method with T-scores and binomial filter in their SCF acquisition system for Czech and showed that the F-measure was better with a binomial filter. And Korhonen et al. (2000) also showed that F-measure with binomial filter was better than the LLR method.

3.4.3. T-scores

T-scores can also be used to measure association between the CLC and the verb. Using the definitions from Section 3.4.2, T-score is computed as follows:

$$T = \frac{p_1 - p_2}{\sqrt{\sigma^2(n_1, p_1) + \sigma^2(n_2, p_2)}}$$

where,

$$\sigma(n, p) = np(1 - p)$$

The CLC will be assumed valid for the verb if T is greater than some threshold value.

3.4.4. Binomial Filter

In a binomial filter (Brent, 1993) H_0 is formulated as the SCF (CLC in our case) is not associated with the verb and that the error probability (p_e) of occurring CLC with the verb is binomially distributed. Then probability of appearing m counts of CLCs in the total n occurrences of the verb can be computed by the following formula.

$$P(m, n, p_e) = \frac{n!}{m!(n-m)!} p_e^m (1 - p_e)^{n-m}$$

The following summation equation gives the probability of a CLC occurring m or more times.

$$P(m+, n, p_e) = \sum_{i=m}^{n} P(i, n, p_e)$$

If this probability $P(m+, n, p_e)$ is less than some threshold value, then H_0 is rejected and the CLC is assumed valid for the verb. A lesser threshold value gives a higher confidence level.

Briscoe and Carrol (1997) estimated error probabilities using reference resources. They defined verb classes based on frames and class membership probabilities of verbs were computed by dividing the number of verbs occurred with the class in Alvey NL Tools (ANLT) dictionary by total number of verbs in the dictionary. The probability of a pattern for class i was computed by dividing the number of patterns for class i extracted from the Susanne corpus by the total number of the patterns. The probability (p_e) of the verb not of class i occurring with a pattern of class i was computed by multiplying the complement of class member ship probability to the pattern probability as follows.

$$p_e = (1 - \frac{|verbs\ in\ class\ i|}{|verbs|}) \frac{|patterns\ for\ i|}{|patterns|}$$

Brent, however, computed error probabilities experimentally from the corpus. In his method (see (Brent, 1993) for detail) a fixed number of first occurrences of verbs in the corpus are examined for different SCFs. For a specific SCF, verbs distributions over different bins is established, that tells how many verbs occurred with how many occurrences of the SCF. The verbs that do not associate with the SCF are clustered towards the lower bins. Starting form the fist bin to higher ones, the error probability is estimated that nearly fits to the binomial distribution.

In our experimentation, we used Brent's method to compute error probabilities by extracting CLCs of first 100 occurrences of 60 verbs in the corpus.

3.5. SCFs Induction

Once the unreliable CLCs for a verb are filtered out, the SCFs for the verb are induced in following three stages.

3.5.1. Application of Metarules

The information about the presence or absence of *ne* and *ko* bits in different CLCs is sufficient to infer about the subject of the verb. The following three metarules are applied to infer the subject of the verb.

1. If the *ne* bit in some of the CLCs is 1 then the subject of the verb is ergative and the verb is potentially transitive.[2]
2. If the *ne* bit is almost always zero and the *ko* bit is almost always 1 in CLCs of the verb then the subject of the verb is dative.
3. If both the *ne* and the *ko* bits are almost always zero then the subject of the verb is nominative and the verb is potentially intransitive.

3.5.2. CLCs Collapsion

After the application of metarules we can collapse CLCs of the verb ignoring the bits in CLCs whose information has already been exploited. If the subject of the verb is inferred by applying Metarule 1 then ignoring the *ne* bit, 64 CLCs could be collapsed into 32 CLCs. Application of Metarule 2 leads to the collapsion of 64 CLCs to 16, as here two bits *ne* and *ko* are ignored. After applying Metarule 3 also, CLCs are collpased into 16. So, for three types of verbs, at the most 32, 16 and 16 number of SCFs respectively can be induced with the system.

3.5.3. SCF Information Collection

Different bits of CLCs after collapsion give the exhaustive information about the SCFs of the verb. The number of these CLCs are actually the number of valid SCFs for the verb. In case of the transitive verbs, the value 1 of the *ko* bit is the signal that verb takes accusative object and the 0 value of the *ko* bit tells that the verb takes nominative object. The other bits fill the rest of information about SCFs. For the intransitive verbs with nominative subject, other arguments in SCFs recognized are usually adjuncts of those verbs.

4. Results and Evaluation

Sixty basic verbs of Urdu were examined for CLCs and SCFs with our system. In the first stage, we tested our CLC acquisition system. The reliable CLCs were filtered out by binomial hypothesis testing. The error probabilities were computed by the method as proposed by Brent (1993). The threshold value was set to 0.01 to achieve a 99% confidence level. The valid CLCs for different verbs were compared with hand judgments. The results for 22 CLCs have been displayed in Table 3. As was already mentioned that the *kih* form is multifunctional in Urdu, the CLCs results show that false positives are more when the *kih* bit is on in the CLCs.

[2] A very few intransitive verbs take volition-based ergative subject. These are handled separately. We are more sure that ergative subject is of transitive verbs if in some of CLCs, *ko* bit is also 1 alongside *ne* bit.

CLC	p_e(CLC)	TP	FP	TN	FN	MC	%MC
000000	0.0000	59	1	0	0	1	1.6
100000	0.0023	38	6	13	3	9	15.0
010000	0.0137	32	2	17	9	11	18.3
001000	0.0093	36	5	13	6	11	18.3
000100	0.0025	44	11	3	2	13	21.6
000010	0.0000	0	0	54	6	6	10.0
000001	0.0077	13	19	28	0	19	31.6
110000	0.0021	19	4	17	20	24	40.0
101000	0.0044	12	3	39	6	9	15.0
100100	0.0009	37	7	16	0	7	11.6
100010	0.0000	0	0	57	3	3	5.0
100001	0.0021	8	8	39	5	13	21.6
011000	0.0050	10	2	47	1	3	5.0
010100	0.0024	19	7	34	0	7	11.6
010010	0.0000	0	0	56	4	4	6.6
010001	0.0018	11	9	38	2	11	18.3
001100	0.0058	15	7	36	2	9	15.0
001010	0.0000	0	0	58	2	2	3.3
001001	0.0059	5	3	50	2	5	8.3
000110	0.0000	0	0	60	0	0	0
000101	0.0022	11	6	43	0	6	10.0
000011	0.0000	0	0	57	3	3	5.0

Table 3: Results of 22 CLCs for 60 verbs compared with hand judgments

In the second stage, the SCFs information was induced from the CLCs extracted in the first stage. In our system, SCF information includes both grammatical function and case of the argument of the verb. The subject has also been included in SCF information, as in Urdu subjects of different types of verbs are marked for different cases. Table 4 shows evaluation for 9 SCFs.[3]

Subcat Frame	TP	FP	TN	FN	MC
$\text{Subj}_{erg}\text{Obj}_{acc}$	25	4	15	16	20
$\text{Subj}_{erg}\text{Obj}_{nom}$	38	5	14	3	8
$\text{Subj}_{erg}\text{Comp}$	8	8	43	1	9
$\text{Subj}_{nom}\text{Comp}$	0	7	53	0	7
Subj_{nom}	14	5	37	4	9
Subj_{dat}	0	0	59	1	1
$\text{Subj}_{erg}\text{Obj}_{nom}\text{Loc}_{from}$	12	3	39	6	9
$\text{Subj}_{erg}\text{Obj}_{nom}\text{Loc}_{in}$	37	7	16	0	7
$\text{Subj}_{erg}\text{Obj}_{nom}\text{Loc}_{on}$	0	0	57	3	3

Table 4: SCFs induced for 60 verbs

Theoretically accusative direct object is specific and nominative object is underspecified for specificity. However, in newspaper data accusative specific object are rarely found. That is which we see in the table that for many transitive verbs accusative object is not detected.

[3] In Tables 3 and 4, TP=true positives; FP=false positives; TN=true negatives, FN=false negatives; MC=misclassified verbs; %MC=percentage of misclassified verbs

5. Conclusion and Future Directions

The scheme presented in this paper is very promising for inferring subcat frames of verbs in Urdu. As other South Asian languages like Saraiki, Sindhi, Balochi, Nepali etc., are very near to Urdu structurally, this scheme could also be applied to these languages. In principle, the approach is applicable to any case-rich language, whether of fixed word order or of free word order, if the nouns in the language are marked for case by separate lexical elements. The system does an extensive screening for the candidate sentences of the target verb, therefore if the experimentation is done on a sufficient large Urdu corpus, the results are expected to improve further.

In our work we tested verbs for frames of elementary arguments. In the future, work on compound postpositional arguments/adjuncts could be done by enriching the scheme with more rules. The system developed could also be helpful in a more refined classification of Urdu verbs. The syntactic information of verbs can also be used to define the semantic verb classes of Urdu as has been done for German (Schulte im Walde and Brew, 2002; Schulte im Walde, 2006). Further the data for analyzing complex predication phenomenon (Butt, 1995) in Urdu can be collected by using this system. So far we have tested our system only for simple predicates in Urdu. Acquiring complex predicates from the corpus automatically will be a value-added task in the system

6. References

T. Ahmed. 2006. Spatial, Temporal and Structural Uses of Urdu KO. In *Proceedings of the LFG06*.

S. R. Anderson. 1977. On mechanisms by which languages become ergative. In C. Li, editor, *Mechanisms of Language Change*, pages 317–363. University of Texas Press, Austin, Texas.

M. R. Brent. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics (19)2*, pages 243–262.

T. Briscoe and J. Carrol. 1993. Generalized probabilistic LR parsing for unification-based grammars. *Computational Linguistics (19)1*, pages 25–60.

T. Briscoe and J. Carrol. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–263, Washington,DC.

M. Butt and T. H. King. 2005. The status of case. In Dayal and Mahajan, editors, *Clause Structure in South Asian Languages*, pages 153–198. Springer Verlag, Berlin.

M. Butt. 1993. Object Specificity and Agreement in Hindi/Urdu. In *Papers from the 29th Regional Meeting of the Chicago Linguistic Society*, pages 80–103.

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

M. Butt. 2006. The dative-ergative connection. In *Proceedings of the Colloque Syntax-Semantique Paris (CSSP) 2005*.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics (19)1*, pages 61–47.

R. Grishman, C. Macleod, and A. Meyers. 1994. Comex syntax: building a computational lexicon. In *Proceedings of the International Conference on Computational Linguistics, COLING-94*, pages 268–272, Kyoto, Japan.

A. Kinyon and C. A. Prolo. 2002. Identifying Verb Arguments and their Syntactic Function in the Penn Treebank. In *Proceedings of the 3rd International Conference on Languages Resources and Evaluation*, pages 1982–1987, Las Palmas de Gran Canaria, Spain.

K. Kipper-Schuler. 2005. *VerbNet, A broad-coverage, comprehensive verb lexicon*. Ph.d. diss., University of Pennsylvania.

A. Korhonen, G. Gorrel, and D. McCarthy. 2000. Statistical Filtering and Subcategorization Frame Acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Lanaguage Processing and Very Large Corpora*, pages 199–205, Hong Kong, China.

M. Lapata. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 397–404, Maryland.

B. Levin. 1993. *English Verb Classes and Alternation: A Preliminary Investigation*. The University of Chicago Press, Chicago.

C. D. Manning. 1993. Automatic Acquisition of a Large Subcategorization Dictionary From Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Columbus, OH.

A. Meyers, C. Macleod, and R. Grishman. 1994. Standardization of the Complement Adjunct Distinction. In *Proceedings of the 7th EURALEX International Congress.*, Goteborg Sweden.

R. O'Donovan, M. Burke, A. Cahil, J. V. Genabith, and A. Way. 2005. Large-scale induction and evaluation of lexical resources from the penn-ii and penn-iii treebanks. *Computational Linguistics (31)3*, pages 329–365.

C. Pollard and I. A. Sag. 1987. *Information-Based Syntax and Semantics*. CA: CSLI Publications.

A. Sarkar and D. Zeman. 2000. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 691–697, Saarbrücken, Germany.

S. Schulte im Walde and C. Brew. 2002. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Philadelphia, PA.

S. Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics (32)2*, pages 159–194.

A. Ushioda, D. A. Evans, T. Gibson, and A. Waibel. 1993. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In *Proceedings of the Workshop on the Acquisition of Lexical Knowledge from Text*, pages 95–106, Columbus, OH.