Syntactic and Semantic Analysis of Urdu Modal Verbs using XLE Parser

Qaiser Abbas
Department of Computer Science & IT
University of Sargodha, 40100, Sargodha, Pakistan

Tehseen Zia
Department of Computer Science & IT
University of Sargodha, 40100, Sargodha, Pakistan

Ahsan Nabi Khan
Data Scientist @Teradata
Teradata Alumni of UET and NUCES, Lahore, Pakistan

ABSTRACT
In this extended work, a syntactic and semantic analysis of Urdu Modal verbs is presented using lexical functional grammar encoded in an environment of the XLE parser. This grammar is developed to understand the syntactic and semantic structure of Urdu modal verbs taken from the sentences of a corpus. Based on this analysis, Urdu modal verbs are finally classified. The corpus is analyzed and sentences including modal verbs like کرنا پا را (kar pAnA) ‘can do’, کرنا ستا کیا (kar saknA) ‘can do’, کرنا پرا (karnA pAra) ‘must/obliged to do’, etc., are extracted from the corpus. Rules are studied, analyzed and encoded for these particular sentences in the form of lexical functional grammar which provides us finally the syntactic and semantic structure of different classifications.

General Terms
Computational Linguistics, Natural Language Processing, Language Engineering, Linguistics

Keywords
Lexical Functional Grammar; Urdu Modal Verbs; XLE Parser; Classification; C-Structure; F-Structure;

1. INTRODUCTION
The types of modal verbs include adative and venitive, anticausative, captive, catenative, ditransitive, frequentative, germanic, intensive, intransitive, light, momentane, predicative, transitive, copula, etc. A detail discussion on these types can be seen in [9, 16] or in the initial version of this work [15]. Modal verbs provide additional information about the main verb (predicate) of the sentence. It has a scale ranging from possibility (can) to necessity (must) which is further divided into two functional divisions. One is freedom to act which includes ability, permission and duty. Other includes likelihood or certainty e.g. must is used in sentences where absolute obligation, order, requirement and necessity, etc., are demanded, can/could represents physical or mental ability, may/might represents permission, option or choice. Similarly, will demonstrates intention in 1st person, volition in 2nd and 3rd person, shall/should is objective in 1st person though not moral obligation. In 2nd and 3rd person, shall imply an obligation imposed by the speaker. However, shall/will misleads us in future tense auxiliary verbs where shall is used for 1st person and will is used for 2nd and 3rd person or shall and will are interchangeable in modern English. Modal auxiliary verbs are classified in two distinct interpretations. One is epistemic claiming how certain the factual status of the embedded proposition is while the other is deontic included notions of permission and obligation [3, 7].

Modal verbs have been categorized into many different forms. Their simple forms are self-explanatory and their discussion can be seen in [4]. For example, I could swim at the beach (Modal Simple), I could be swimming at the beach right now (Modal Continuous), I could have swum at the beach yesterday (Modal Perfect). Similarly, had better can be used to represents recommendation, desperate, or warn in many cases. For example, You had better take your luggage with you is the recommendation, That girl had better get here soon! is the case of desperate hope, while You had better watch the way you talk to me in the future! is a warning to some person. Similarly, have to is used to express certainty, necessity and obligation while have got to is used to express necessity and obligation. Ought to, is used when advice or recommendation is required e.g. you ought to stop smoking (Recommendation) etc. It also represents assumption or expectation as strong probability, Ought not without to is also used to advise something. Many other modal verbs with examples are discussed in [4]. A similar study was performed by Abbas et. al, but on Urdu copula verbs with a little different approach which can be seen in [10].

In Urdu, the verb has four divisions in major e.g. root, imperfective participle, perfective participle and infinitive. Detailed divisions of verbs can be seen in [11, 12]. Auxiliaries and suffixes transfer it into a complex system of verb tense and aspect described in [13]. The information of verbs and auxiliaries can also be encoded at part of speech (POS) level discussed in [14]. The verb root can be used with modal verbs e.g. (saknA) ‘to be able’ to show ability and پانا (cuknA) ‘to be finished’ to show completion. The preliminary work on Urdu modals was discussed in [15]. As an example, modal verbs form a complex system of verbs with a little different approach which can be seen in [10].

In this grammar, we have included the forms that are in the environment of the XLE parser. This grammar is used for the recommendation, that girl had better get here soon! is the case of desperate hope, while You had better watch the way you talk to me in the future! is a warning to some person. Similarly, have to is used to express certainty, necessity and obligation while have got to is used to express necessity and obligation. Ought to, is used when advice or recommendation is required e.g. you ought to stop smoking (Recommendation) etc. It also represents assumption or expectation as strong probability, Ought not without to is also used to advise something. Many other modal verbs with examples are discussed in [4].

The tense of (saknA) ‘to be able’ to show ability and پانا (cuknA) ‘to be finished’ to show completion. The preliminary work on Urdu modals was discussed in [15]. As an example, modal verbs form a complex system of verbs with a little different approach which can be seen in [10].

In this grammar, we have included the forms that are in the environment of the XLE parser. This grammar is used for the recommendation, that girl had better get here soon! is the case of desperate hope, while You had better watch the way you talk to me in the future! is a warning to some person. Similarly, have to is used to express certainty, necessity and obligation while have got to is used to express necessity and obligation. Ought to, is used when advice or recommendation is required e.g. you ought to stop smoking (Recommendation) etc. It also represents assumption or expectation as strong probability, Ought not without to is also used to advise something. Many other modal verbs with examples are discussed in [4].

Modal verbs have been categorized into many different forms. Their simple forms are self-explanatory and their discussion can be seen in [4]. For example, I could swim at the beach (Modal Simple), I could be swimming at the beach right now (Modal Continuous), I could have swum at the beach yesterday (Modal Perfect). Similarly, had better can be used to represents recommendation, desperate, or warn in many cases. For example, You had better take your luggage with you is the recommendation, That girl had better get here soon! is the case of desperate hope, while You had better watch the way you talk to me in the future! is a warning to some person. Similarly, have to is used to express certainty, necessity and obligation while have got to is used to express necessity and obligation. Ought to, is used when advice or recommendation is required e.g. you ought to stop smoking (Recommendation) etc. It also represents assumption or expectation as strong probability, Ought not without to is also used to advise something. Many other modal verbs with examples are discussed in [4].

Similarly, the word پانا (pAnA) showing the possibility of an action dependent on other circumstances is used mostly in
negative aspect. The verb phrase has the structure as VERB ROOT + inflected form of ِنا (na) while the verb phrase structure of ِنا (na) is VERB ROOT + inflected form of ِنا (na). The example of phrase structure with ِنا (na) is given in (III).

(III). ِنا (na) -e: ِنا (na) -e

vah 21d mas2ruf hE kal kI davat mEN nahi A pAyE gA

‘Wahid is busy; he cannot manage to come at tomorrow’s party’.

In this sentence ِنا (na) is used as modal and is intransitive. It agrees with the subject as advocated in [8]. In contrast, it can also be used as a non-modal verb with transitve valency when used as a sentence predicate like in (IV).

(IV). ِنا (na) -e: ِنا (na) -e

انجم نے امتحان میں اچھے نمبر پاےٗ

‘Anjum took a good mark in the examination’.

The modal verb ِنا (na) (cuknA) shows the completion of an action or event prior to a second action or event mostly used in imperfective tenses. Its phrase structure is given as VERB ROOT + inflected form of ِنا (na). It is confusing because it remains intransitive even in the case of perfect tense and ِنا (na) cannot be used when the verb phrase form belongs to transitive verb like in (V).

(V). ِنا (na) -e: ِنا (na) -e

ham gARI bEc cukE tHE jab malik s2Ahib nE daryafT kiyA

‘We had already sold the car when Malik Sahib inquired about it’.

Modal infinitives are divided into personal and impersonal constructions. Its phrase structure is NOMINATIVE INFINITIVE + an inflected form of ِنا (na) (cAhnA) ‘to want (to)’. The subject is nominative, however if an infinitive form has an object, then it can also agree optionally with it like adjectives.

The impersonal infinitive has the structure NOMINATIVE INFINITIVE + IMPERSONAL VERB and the necessity ranging from advisability to lack of choice. Verbs agree with the object if present unless the object is followed by the postposition. The infinitive also agrees with it, but this agreement is weakened in longer sentences and it remains masculine singular. In place of ‘should to’ or ‘ought to’ of English modal verbs, the Urdu has NOMINATIVE INFINITIVE + ِنا (na) (cAhnA) ‘to want (to)’ and ِنا (na) (cAhnA) represents advisability or wish or necessity. It agrees with the object of infinitive in number, unless the past auxiliary is added to the phrase as in (VI).

(VI). ِنا (na) -e: ِنا (na) -e

mujhE sE rjIm cAHnE

‘I should go for a walk’.

Where I-OBL-DAT (subject) should do-INF-f (~INF-m.sg) wished (sg) and a sentence in (VII).

(VII). ِنا (na) -e: ِنا (na) -e

as eD eDhE wErdiNE cAHnE

‘He should buy eggs’.

Where he-obl-DAT egg-pl. buy-INF-m.pl (~INF-m.sg) wished-pl. In (VI) and (VII), the modal verb ‘have to’ has a structure in Urdu as NOMINATIVE INFINITIVE + AUXILIARY shows necessity. Moreover, if the subject is inanimate then the dative case marking is not required as in (VIII).

(VIII). ِنا (na) -e: ِنا (na) -e

mujhE voh kitAb paRHnI hE
gI

‘I have to read that book’.

Where I-OBL-DAT that book read-INF-F (~INF-M-SG) AUX-3SG. Some other auxiliaries are also used like ِنا (paRNhI) and ِنا (hi) and much more like in (IX).

(IX). ِنا (na) -e: ِنا (na) -e

mujhE voh kitAb paRHnI hO gl

‘I will have to read that book’.

Where I-OBL-DAT that book read-INF-F-SG (~INF-M-SG) AUX-3SG. FUT-F-SG. There are issues in modal verbs like ‘telic aspect’ and ‘marker of sources’, which are clarified in [1] and [2]. Agha represents ِنا (paRNhI) and ای (AnI) as an aspectual marker called as TELIC aspect [5]. He argues that ِنا (paRNhI) is homophones with the infinitive marker ِنا (~ paRNhI hE) but has a different morphological composition as in the reproduced version of (IX) as I-OBL-DAT that book read-TEL-F AUX-3SG. He also argues that split in case infinitive (dative with [+animate] and nominative with [-animate]) is aspectually driven and thus comparable to the split ergative of the perfective aspect. The مصیبت (mus2bat) being inanimate takes the nominative as in (X).

(X). ِنا (na) -e: ِنا (na) -e

mus2bat AnI hE

‘Misfortune is bound to strike’.

In misfortune come-TEL-F AUX-3SG, [1] does not call it the ‘telic participle’ but the agreement of infinitive with its object is a pattern usually associated with participles and not with infinitives, which are basically verbal nouns. Similarly, [2] gives NOMINATIVE INFINITIVE + AUXILIARY with ergative case marking by ِنا (ne) of the subject in Urdu of Pakistan. He declared it as a marker of source in a generalized sense. In the sentence in (XI).

(XI). ِنا (na) -e: ِنا (na) -e

Aj mEN nE z2arUr jA nA hE

‘Today I am definitely going [+conscious choice]’.

Where today I-ERG certainly go-INF AUX-3SG Volutition is involved. By contrast, dative subject is usually unmarked for SOURCE, so the issue of volution does not arise in (XII).

(XII). ِنا (na) -e: ِنا (na) -e

mujhE z2arUr jA hE

‘Today I definitely have to go [+conscious choice]’.

Where today I-OBL-DAT certainly go-INF AUX-3SG is the glossing of (XII). This gives an idea of the form as an unexplored dimension of the modal system at least in Pakistani Urdu. The English word ‘must’ has structure in Urdu as NOMINATIVE INFINITIVE + inflected form of ِنا (paRNhI) ‘to fall, befall’. It represents external compulsion or lack of choice as in (XIII).
2. METHODOLOGY

Methodology involves three steps. At first, we searched through the corpus keenly and identified those sentences which contained modal verbs. POS tagging was done using a tagger developed by Hassan [6]. Underlying structure was observed then by using the tags assigned to the verb phrases. Secondly, we determined the syntactic and semantic structure governing the varieties of modals and developed a LFG (lexical functional grammar) framework. Thirdly, we implemented the LFG using XLE Parser and analyzed the results.

A corpus of more than 114,000 words was used to find the modal verbs which contained text mostly from newspapers and 210 sentences containing modal verbs were tagged which were identified in the first step. The tag set and the technology used has been mentioned earlier. Variations in the modal verbs were recorded and shown in results. In order to streamline the steps solely for modal verb analysis, LFG vocabulary is limited to the words dependent on modals only. The rest of the words were substituted for the ease of comparison, thus parsing and testing was rendered easily. Simple past tense for implementing the LFG for most of the modal verbs was taken. Since our analysis was not on the tense, so we took these variations as constant in our analysis. We converted the verbs that differ in tense to the above form. The two exceptions are in the modal categories of 'advised' and 'binding' as can be seen in the results. They are always used in infinitive sense.

We chose different values of gender and case in order to observe the agreement in constituent structures. The rules discovered are mentioned in results. Two LFG files (grammar rules and lexicon) are created with a common lexicon. Grammar rules recognize Nominative Case sentences and the lexicon recognizes Ergative and Dative cases. These files of grammar rules and lexicon are given in Figure 1. Where the POS are m: modal, v: main verb, n: noun, cm: case marker.

File1: Grammar Rules

S --> n: (^SUBJ) =! (NUM) = (!NUM); (GEND) = (!GEND); (OBJ) = (!OBJ); (PRED) = ! (SUBJ) (^OBJ); (CASE) = ! (CASE) = NOMINATIVE.
NP --> {n cm | n }.
VP --> v m.

3. RESULTS

After analyzing the corpus, we found the following variations of modals with the verb ہے (kar) 'do'. These modals have been observed to come often with the other verbs and showing the same semantic properties. The modal phrases take the subcategorization frame <SUBJ, OBJ> in all the phrases used because of the main verb ہے. In phrases with other main verbs,
Ahmad Aslam ko kitab de AyA

‘Ahmed has given the book to Aslam’

The modal verb here does not determine the subcategorization frame. This is not an exception but a rule.

The modal verbs do follow agreement rules. They inflect in gender and number with either subject or object. In case of verbs that have <SUBJ, OBJ, OBJ1> subcategorization frame, if the modal verb has to agree with direct object, it agrees with OBJ. For example in the sentence in (XV) as follows:

Ahmad ne Aslam ko kitab de di

‘Ahmed has given the book to Aslam [and returned]’

The subcategorization frame becomes <Ahmad, kitab, Aslam>, then the modal verb in di agrees with the case of the noun phrases. Hence they determine the case of the verb phrase. The case required by the modals must be provided by the noun phrase structure.

Table 1 lists the case requirement with all the modal verbs mentioned. These rules have been taken care of in the implementation of LFG. The screen shots of the run are depicted in the Appendix.

### Table 1: Modal variations with the verb

<table>
<thead>
<tr>
<th>Semantic Properties</th>
<th>Agrees with</th>
<th>Case Required</th>
<th>Modal Verb types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive</td>
<td>OBJ</td>
<td>Ergative</td>
<td>ker dia (did something)</td>
</tr>
<tr>
<td>Personal</td>
<td>OBJ</td>
<td>Ergative</td>
<td>ker lia (finish doing something)</td>
</tr>
<tr>
<td>Complete</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker cha (did completely)</td>
</tr>
<tr>
<td>Starting</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker may lag (began to do)</td>
</tr>
<tr>
<td>Lapsed</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker bai (did impulsively)</td>
</tr>
<tr>
<td>Performed</td>
<td>OBJ</td>
<td>Ergative</td>
<td>ker dikhaya (performed something)</td>
</tr>
<tr>
<td>Successful</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker gia (someone performed/did)</td>
</tr>
<tr>
<td>Able</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker saka (could do)</td>
</tr>
<tr>
<td>Marginally Able</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker paya (could do)</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker para (suddenly happened)</td>
</tr>
<tr>
<td>Unavoidable</td>
<td>OBJ</td>
<td>Dative</td>
<td>kerna para (must obligated to do)</td>
</tr>
<tr>
<td>Binding</td>
<td>OBJ</td>
<td>Dative</td>
<td>kerna ho ga (will have/bound to do)</td>
</tr>
<tr>
<td>Needed</td>
<td>OBJ</td>
<td>Dative</td>
<td>kerna tha (supposed/needed to do)</td>
</tr>
<tr>
<td>Advised</td>
<td>not agree</td>
<td>Dative</td>
<td>kerna chahyee (should do)</td>
</tr>
<tr>
<td>Approaching</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker aya (arrival after doing)</td>
</tr>
<tr>
<td>Leaving</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker chala (departure after doing)</td>
</tr>
<tr>
<td>Leaving for Good</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker guara (left after doing)</td>
</tr>
<tr>
<td>Indifferent</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>karta ho ga (must doing/use to do)</td>
</tr>
<tr>
<td>Suggestive</td>
<td>SUBJ</td>
<td>Nominative</td>
<td>ker saka (can do)</td>
</tr>
</tbody>
</table>

### 4. DISCUSSION

Section 2 shows six different modal verbs with variations in meaning and case required. We have expanded the types to 19 different ones classified into five major categories of relationships, processes, ability, mobility and likelihood.

In English the modals have the major categories of epistemic and deontic. Here epistemic maps to our category of likelihood and deontic maps to the category of ability. However, the finer changes in the meaning identified in the 19 types, have not been identified before. Hence, we see more information encoded in the modal verb structure of Urdu.
During this work, it was seen that modals can behave as verb and auxiliary in Urdu like میں چاہوں گا (mEN cAhUN gA) ‘I will want’. Here میں چاہوں گا (mEN cAHUN gA) ‘I want’ is behaving as modal verb with subjunctive morphological form. On the other hand، میں چاہوں گا (mEN jAhAn cAhUN gA) ‘I had wanted to go’. In this sentence میں چاہوں گا (mEN jAhUN gA) ‘want’ is behaving as an aspectual modal auxiliary with imperfective morphology. This difference noted is quite interesting and plan to cover up in future work.

5. CONCLUSION

Except the future work mentioned in Section 4, Modal verb expressions have been shown are divided into five major categories of relationships, processes, ability, mobility and likelihood, which are further broken down into 19 different types. Modals govern the sentence case and agree with either subject or object which is demonstrated in this paper. In this future, this work of Urdu modals will be extended into further categories according to their behavior as a verb or as an auxiliary.

6. REFERENCES


APPENDIX

Sentences of Urdu are encoded in English alphabets. Only five and half out of nineteen screen shots of results mentioned in Table 1 are displayed in this paper to provide an idea about classification and their analysis through XLE parser. The C and F structure of a sentence in (XVI) is depicted in the following screen shots.

(XVI). میں کا کام کر دیا

Sara ne kam ker diya

‘Sara has done (shows interactive modality) the work’
(XVII). Sara ne kam kar liya
‘Sara has done (personal modality) the work’

The sentence given in (XVII) has the following C and F structure produced in the analysis during parsing, which shows the personal ability of ‘Sara’ to finish the work.

(XVIII). Sara kam ker chuki
‘Sara has done (completion modality) the work’

The sentence given in (XVIII) shows the completion of work, hence concluding completion modality.
The sentence in (XIX) shows the starting modality of an event by ‘Sara’, The screen shots are as follows.

(XIX). سارہ کام کرنے لگی

‘Sara is going to start (starting modality) the work’

The screen shots as follows depicts the analysis of lapsed modality resides in the sentence in (XX).

(XX). سارہ کام کر بیٹھی

‘Sara has finished (lapsed modality) the work’
Sara shows the ability of perform the work

Similarly, the ability to perform analyzed through parsing the sentence (XXI) is depicted in the following screen shot.