Clustering Urdu verbs on the basis of related aspectual auxiliaries and light verb

Tafseer Ahmed
Department of Linguistics
Universität Konstanz
Germany
tafseer@gmail.com

Abstract

The paper describes work done on Urdu aspectual auxiliaries/ light verbs. The frequencies of main_verb-auxiliary and main_verb-light_verb sequences are obtained by shallow processing of an Urdu corpus. The normalized data corresponding to each of the main verbs is used for clustering (unsupervised learning) of Urdu verbs. It gives four major clusters of verbs that are hierarchally arranged. The inspection of the verbs in these clusters show that most of the verbs in a cluster share some common semantic property and these clusters correspond to semantic classes of the verbs. Some other properties about auxiliary/light verbs are also found in this analysis.

1 Introduction

Urdu is an Indo-Aryan language spoken in Pakistan and India. It is closely related to Hindi with same grammatical structure but differences in script and vocabulary. In Urdu, we find sequences of verbs in which the main verb is followed by another verb (Schmidt 1999). The following-verb can be an auxiliary, a modal or a light verb. Consider these examples.

(1) TrEn A-I (thI)
train come-Perf.F.SG Past.F.SG
'The train came.'

(2) TrEn A rah-I thI
train come Prog-F.SG Past.F.SG
'The train was coming.'

(3) TrEn A ga-I (thI)
train come go-F.SG Past.F.SG
'The train had come.'

All of the above examples have tense auxiliaries at the last position. The tense auxiliary can follow the main verb or the verb-verb sequence. However, we are not interested in tense auxiliaries. In this paper, we are interested in the verbs like *rah* 'stay' (used for progressive) and *jA* 'go' (use for completion).

Siddiqui (1971) and Hook (1974) provided a list of verbs that follow main verbs. The verbs are: *dE* 'give', *lE* 'take', *A* 'come', *jA* 'go', *DAl* 'insert', *paR* 'fall', *rah* 'stay', *beTH* 'sit', *cuk* (for completion), *sak* (for ability), *pA* 'get', *kar* 'do', *hO* 'be', *uTH* 'rise', *cAH* 'want', *dE* 'give', *rakH* 'put', *ban* 'get make', *lag* 'touch/hit', *nikal* 'come out', *Tahar* 'stop' and *cal* 'move'.

As mentioned earlier, this list does not contain a single type of verbs. The list includes auxiliaries e.g. *rah* for progressive, modals e.g. *cAH* for want and light verbs e.g. *jA* for completion. We present all of these in a single list, because in the latter part of this paper we argue that some of the auxiliaries especially the progressive auxiliary *rah* 'stay' are correlated to certain semantic classes of the verb. Hence, we need to study the behavior of all of these verbs irrespective of their syntactic properties. We use the term V2 for all of these verbs throughout this paper. Many writers e.g. Butt (1995) distinguish between aspectual auxiliaries and light verbs, but we use the same term V2 for all these verbs.

There is no significant work on the semantic verb classes of Urdu. There are few references to some verb classes such as ingestives (Saksena 1982) and the intransitive verbs that allow optional *ne* (Butt 1995).

The modals and aspectual auxiliaries can follow any verb, but it is not the case for light verbs. Consider the following example.

(4) a. gARI cal dI
vehicle.F.SG move give.Perf.F.SG
'The vehicle started moving.'
b. *gARI ruk dI
vehicle.F.SG stop give.Perf.F.SG
'The vehicle stopped.'

The light verb *dE* 'give' is not used with the verb *ruk* 'stop'. Hence, each light verb is compatible with some, and does not occur with the verbs of other semantic classes.

Our experiment tests the hypothesis and investigates whether there is a correlation between the progressive marker and certain verb class(es).

There is an important syntactic issue in the processing of V2 verbs. Each V2 verb governs the morphological form of the main verb preceding it. Different morphological forms of the previous verb correspond to different syntactic/semantic interpretation of the V2 following it.

Consider the example of *jA* 'go'. It can be interpreted as passive marker, completion marker or continuity marker on the basis of the form of the main verb preceding it. If the (preceding) main verb is in perfective form, *jA* is considered as the passive marker. If the main verb is in root form, *jA* is considered as the completion marker and if the imperfective form of the main verb is used, it is considered as a continuity marker. See the following examples.

(5) sEb kHA-yA gayA
apple.M.Sg eat-Perf.M.Sf go.Perf.M.Sg
'Apple was eaten.' (Passive)

(6) sEb pak gayA
apple.M.Sg ripe go.Perf.M.Sg
'Apple had ripen.' (Completion)

(7) vuh sEb kHA-tA gayA
3SG apple eat-Impf.M.Sg go.Perf.M.Sg
'He kept on eating the apples.' (Continuity)

The aim of our experiment is to analyze the corpus to get some empirical results about V2 verbs and the related issues introduced above. What is the behavior of these V2 verbs related to different verbs and different classes of verbs? Can we find verb classes based on distribution of V2 verbs with the main verb?[1]

2 Experiment

In the above section, we presented a hypothesis that there is a semantic relation among many of V2 verbs and the main verbs. In this experiment, we try to find empirical evidence for this hypothesis. The experiment has two parts. The first part provides the frequency of each V2 verb corresponding to each main verb. These frequencies can tell us about the syntactic/semantic properties of V2 verbs. The second part of the experiment tries to cluster the (main) verbs on the basis of frequencies of V2 verbs associated with them. Can we find semantic classes on the basis of V2 verb frequencies?

It is not easy (and possible with limited resources) to employ deep parsing methods to perform this experiment. For Urdu, there is no tree bank available. Similarly, no sizable POS tagged corpus is available. Moreover, no morphological analyzer is publicly available that is easily integratable with other applications. Hence, it is necessary to use shallow methods to perform this experiment.

We plan to count the occurrence of the main verb followed by the V2 verb. The main verb can be in one of the different morphological forms as Urdu verb is inflected on the basis of number, gender and/or person agreement. As we are not able to use a morphological analyzer, we planned to obtain data only for those V2 verbs that are followed by the (uninflected) root form of the main verb[2]. There are 12 such V2 verbs that are preceded by the root form of the main verb. We use these verbs in our experiment. The list of these verbs is present in table 1.

A list of Urdu (main) verbs is obtained from Humayoun's (2006) online resources. As most of the Urdu verbs are in form of noun/adjective + verb complex predicate e.g. *intizAr* 'wait' *kar* 'do' (for 'wait'), there are less than thousand simple verbs e.g. *gir* 'fall' in Urdu. The used verb list contains these simple verbs only.

There is a potential problem in using the root form of main verb without deep processing. The masculine singular perfective form of a verb is form identical to its root causative form. For example, the verb *gir* '(get) fall' has perfective form *girA* used for masculine singular agreement. The root form of the corresponding causative verb is *girA* '(make) fall'. (The perfective form of this causative verb *girA* is *girAyA* for masculine singular agreement.) Hence, we remove all such verb-causative pairs that introduce this form ambiguity.

[1] There are certainly other features like subcategorization frame and alternations that can be used in verb clustering, however we tried to find out how much can be done solely on the basis of these (V2) verbs.

[2] The native speaker knowledge tells that these V2 verbs are most frequently used in Urdu. So, it can be assumed that we do not lose much data.

As we use the V2 verbs that are preceded by the root form of the main verb, we do not need to search the other morphological forms of the main verb. However, we do need to find different morphological forms of V2 verbs immediately following the root form of the main verbs. For this purpose, we manually generated all the morphological forms of these twelve V2 verbs.

As a corpus, we processed 7337 documents having 14,196,045 tokens. The documents are obtained from CRULP's (www.crulp.org) Urdu corpus and websites www.urduweb.org and www.kitaabghar.com.

The documents of the corpus are processed one by one. The text of each document is divided into sentences on the basis of sentence breakers. Each word of these sentences is matched with the list of the main verbs. If the word is found in Urdu verb list, the next word is matched with the (inflected) words in the V2 verb list. If it is also found, we increase the count of that verb-V2 combination. To make the data better for normalization, the count of each main verb in imperfective form is also calculated.

After the processing of all the documents of the corpus, we got a table having counts of verb-V2 combinations. We selected 183 verbs for further processing. These are the verbs for which the sum of all the counts is greater than 20.

These data is to be normalized for further processing. The count of each verb-V2 combination is divided by the sum of counts of all combinations for that verb (plus counts of imperfective forms). This gives normalized frequencies of the combination that can be compared in further processing. The normalized frequency table for some verbs is given in table 1. As the sum at denominator includes the count of imperfective forms, the frequencies in each column (that use V2 counts only) do not add up to 1.

The normalized frequencies for the combinations corresponding to each main verb constitute a vector. These vectors are used for clustering that is the unsupervised learning of classes. The software tool Cluster 3.0 is used for hierarchal cluster of these vectors using centroid method. The tool is available at
http://bonsai.ims.utkyo.ac.jp/~mdehoon/software/cluster/software.htm.

V2/main	*gir* 'fall'	*hans* 'laugh'	*tOR* 'break'
rah 'stay'	0.0937	0.1064	0.0771
dE 'give'	0.0032	0.0292	0.5176
lE 'take'	0	0.0133	0.0193
A 'come'	0.0032	0	0
jA 'go'	0.4345	0	0.1350
DAl 'insert'	0	0	0.0354
paR 'fall'	0.1260	0.1064	0
bETH 'sit'	0.0016	0	0
cuk complete	0.0339	0	0.0354
sak able	0.0129	0.0026	0.0482
pA 'find'	0.0016	0	0
uTH 'rise'	0	0	0

Table 1: A sample from the Frequency table corresponding to main_verb-V2 sequences

3 Results

The hierarchal clusters obtained by this exercise are shown in figure 1.

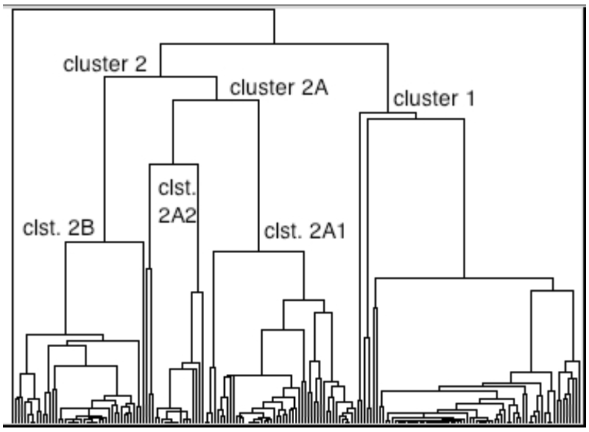

Figure 1: The dendogram showing hierarchal clustering of Urdu verbs on the basis of V2 frequencies.

Visually we can see two/four major clusters. There are two major clusters: cluster 1 and cluster 2. Cluster 2 is subdivided into two major clusters: 2A and 2B. One of these two clusters i.e. 2A is subdivided into two more clusters 2A1 and 2A2. Hence, visually we find four major clusters: 1, 2A1, 2A2 and 2B that are hierarchically arranged. Some of the verbs from these clusters are:

Cluster 1: (approx. 70 verbs)

nikal 'emerge/come out', *jI* 'live', *A* 'come', *jal* '(get) burn', *guzar* 'pass', pak '(get) bake', ban '(get) make', *tHak* 'get tired', *kUd* 'jump (from one point to other)', ucHal 'jump (up and down motion)

kHA 'eat', samajH 'understand', jIt 'win', lUT 'rob', *nahA* 'bath', *pI* 'drink', *nigal* 'swallow'

Cluster 2A1: (approx. 45 verbs)

cal 'move', *hans* 'laugh', *gA* 'sing', *muskurA* 'smile', *bOl* 'speak', *kHEl* 'play, *cIx* 'scream', *nAc* 'dance', *laTak* '(get) hang', *jAg* 'wake up'

paRH 'read', *dEkH* 'see', *sun* 'hear', *mAng* 'ask', *cUs* 'suck', *pIT* 'hit/beat', *kamA* 'earn', *lA* 'bring'

pUcH 'ask', *DHUND* 'search', *cHU* 'touch', *kANp* 'shiver', *baj* '(get) ring'

Cluster 2A2: (approx. 20 verbs)

jAn 'know', *pahcAn* 'recognize', *apnA* 'adopt', *pahan* 'wear', *mAn* 'accept', *tHAm* 'hold', *kHENc* 'pull', *cun* 'pick/pluck'

gin 'count', *dHO* 'wash', *pIs* 'grind', *cHAn* 'filter', *gHEr* 'cover', *tal* 'fry',

Cluster 2B: (approx. 40 verbs)

kah 'say', *dE* 'give', *likH* 'write', *bEc* 'sell', *sukHA* '(make) dry', *navAz* 'give', *batA* 'tell', *jagA* '(make) wake up', *tOR* 'break', *kHOl* 'open', *rakH* 'put', *rOK* '(make) stop' *bAnT* 'divide/distribute', *kas* 'tighten'

4 Discussion

These clusters correspond to semantic classes discussed earlier in the literature. Most of the verbs in the cluster 1 are unaccusative verbs whose subject is a patient/theme. These unaccusative verbs e.g. *nikal* 'emege/come-out' etc. are listed in first paragraph of cluster 1 verb list given in the above section. The second paragraph in this list has another class of verbs i.e. ingestives. The verbs like *kHA* 'eat' etc. are transitive. However, the subject of these verbs is considered as a theme that traverses a path (the object) (Ramchand 2008). Hence ingestives are semantically closer to the unaccusatives. The subjects of both are patient/theme or undergoer in Ramchand's framework.

Cluster 2 corresponds to the (transitive and intransitive) verbs that have agentive subjects. The cluster 2B corresponds to the transitive verbs whose subject bring some change to the object. Most of the verbs are either causing change of state verbs like *toR* 'break' or *bHigO* '(make) wet' or ditransitives like *kah* 'say' or *rakH* 'put'. The subject does not get affected in these types of verbs.

The analysis of verb list of cluster 2A1 shows that it has three kinds of verbs (listed in three paragraphs in above section). The verbs in first paragraph e.g. *cal* 'move' and *hans* 'laugh' etc. correspond to the (intransitive) unergative verbs that have agentive subject. Most of the verbs in second and third paragraphs e.g. *dEkH* 'see'/*kamA* 'earn' and *pUcH* 'ask'/*DHUND* 'search' are transitive verbs whose subject is agentive.

Most of the verbs in class 2A2 are those whose subject gets something physically or logically. The verbs e.g. *cun* 'pick' in the first paragraph easily fit this description, however the verbs in second paragraph form a pragmatic class of those actions e.g. *pIs* 'grind' in which the subject often gets benefit logically.

It must be noted that the "verbs in nth paragraphs" described in above text are not the subcluster given by the clustering tool. We subjectively (and manually) made these subdivisions among the verbs of each clusters to adequately explain the semantics of the verbs in each cluster.

When we sort the frequency table (having cluster labels) with respect to frequencies of V2 verbs, we find interesting observations. The frequencies of progressive auxiliary *rah* 'stay' is correlated to the cluster 2A1. It means that this auxiliary occurs with all kinds of verbs, but it is more frequent with verbs of certain semantic properties. However, the high frequency occurrences of the verb *sak* used as ability marker do not correlate to any verb class.

The frequency analysis gives the productivity/compatibility of each of the V2 verbs. The progressive marker *rah* is found to occur with 161 (out of 183) verbs. The light verb *jA* is found to occur with 121 verbs. On the other hand, light verbs *DAl* 'put' and *uTH* 'rise' are found to appear only with 22 and 23 (main) verbs respectively.

5 Conclusion

Urdu corpus is processed to find frequencies of main_verb-V2(auxiliary/light_verb) combinations. The clustering of this data gives four major clusters that have verbs with common semantic properties. This experiment is an effort to find how much we can comprehend about semantics of Urdu verbs solely on the basis of light verbs/auxiliary frequencies.

References

Abul Lais Siddiqui. 1971. *Jamaul Qawaid (Comprehensive Grammar)*. Karachi.

Anuradaha Saksena. 1982. *Topics in the Analysis of Causatives with an Account of Hindi Paradigms*. Berkeley: University of California Press.

Gillian Ramchand. 2008. *Verb Meaning and the Lexicon: A First Phase Syntax*, Cambridge: Cambridge University Press.

Peter Edwin Hook. 1974. *The compound verb in Hindi*. Ann Arbor: University of Michigan.

Ruth Laila Schimdt. 1999. *Urdu: An Essential Grammar*, London: Routledge.

Muhammad Humayoun. 2006. *Urdu Morphology, Orthography and Lexicon Extraction*. Master's Thesis, Chambéry: Chalmers University of Technology.