Discovering Semantic Classes for Urdu N-V Complex Predicates

Tafseer Ahmed
Universität Konstanz
tafseer.khan@uni-konstanz.de

Miriam Butt
Universität Konstanz
miriam.butt@uni-konstanz.de

Abstract

This paper reports on an exploratory investigation as to whether classes of Urdu N-V complex predicates can be identified on the basis syntactic patterns and lexical choices associated with the N-V complex predicates. Working with data from a POS annotated corpus, we show that choices with respect to the number of arguments, case marking on subjects and which light verbs are felicitous with which nouns depend heavily on the semantics of the noun in the N-V complex predicate. This initial work represents an important step towards identifying semantic criteria relevant for complex predicate formation. Identifying the semantic criteria and being able to systematically code them in turn represents a first step towards building up a lexical resource for nouns as part of developing natural language processing tools for the underresourced South Asian language Urdu.

1 Introduction

Urdu is an Indo-Aryan South Asian language spoken primarily in Pakistan and India. It is structurally almost identical to Hindi and together Urdu and Hindi consitute the third-most spoken language (Graddol, 2004). At the same time, Urdu/Hindi is a severely underresourced language. We are currently engaged in building a broad-coverage, robust computational ParGram grammar for Urdu (Butt and King, 2007; Butt et al., 2009) and one of the major bottlenecks for development is the lack of lexical resources, which are needed, for example, for the development of a verb lexicon with subcategorization frames or lists of argument taking nouns and verbs.

Urdu actually has only about 700 simple verbs (Humayoun, 2006), so the task of finding the range of possible subcategorization frames could be done mostly manually in a reasonable amount of time. However, as is characteristic of South Asian languages in general, Urdu employs wide variety of different types of complex predicates (Butt, 1995; Mohanan, 1994) to express its full range of verbal predication. The complex predicates can be V-V, Adj-V, PP-V or N-V combinations. In this paper, we focus on the highly productive N-V complex predicates in order to try to identify: 1) possible constraints on the range of combinatory possibilities; 2) possible systematic semantic groupings/classes of the nouns involved.

The paper is organized as follows. In section 2 we first describe the basic phenomenon. In section 3 we describe the corpus-based study we performed to see if we can identify systematic semantic classes for nouns. The results are presented in section 4 and the paper is concluded by section 5.

2 Combinatory Possibilities for N-V Complex Predicates

Urdu makes use of only about 700 simple verbs. The bulk of verbal predication in Urdu is effected by complex predicates of various types. The complex predicates are highly productive and different types can be stacked on top of one another (Butt and Ramchand, 2005), so capturing their use computationally in a systematic, generalizable and efficient manner is a challenge. One cannot just trawl a corpus to extract and then list various possibilities as there are potentially infinitely many combinations (though one can choose to list the 100 or so most frequently occurring ones, as done in the Hindi WordNet, for example; Bhattacharyya 2010).

In this paper, we focus on the combinatorial possibilities in N-V complex predicates. In N-V complex predicates the noun contains the main predicational content. The verb, usually referred to as the *light verb*, dictates the case marking of the subject, determines agreement patterns, carries information about tense/aspect and adds information about agentivity vs. experiencer subjects and makes some further subtle semantic contributions. We illustrate the basics of the construction with respect to the noun *yad* 'memory' and the light verbs *kar* 'do' and *ho* 'be'. Other light verbs may be used as well, but these are two of the most basic ones.

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

b. nadya=ko kahani yad hɛ
Nadya.F.Sg=Dat story.F.Sg.Nom memory be.Pres.3.Sg
'Nadya remembers/knows a/the story.' (lit.: 'Memory of the story is at Nadya.')

c. nadya=ko kahani yad hu-i
Nadya.F.Sg=Dat story.F.Sg.Nom memory be.Part-Perf.F.Sg
'Nadya came to remember a/the story.' (lit.: 'Memory of the story became to be at Nadya.')

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

In (1a) the noun *yad* 'memory' is combined with the light verb *kar* 'do'. In this case the subject must be ergative and overall reading is one of an agentive, deliberate remembering. In (1b), in contrast, Nadya is already taken to be in the state of remembering the story. The difference between (1b) and (1c) is one of eventive vs. stative, so that in (1b), Nadya 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'.

A superficial look at Urdu patterns shows that not all nouns are as versatile as *yad* 'memory'. That is, certain nouns are only compatible with a subset of the potentially available light verbs. What has not so far been explored, however, is what the semantic constraints on N-V complex predicate formation are. In order to achieve a first understanding of the relevant patterns, we follow Levin (1993)'s classic assumption 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). Our main aim is therefore to identify semantic classes of nouns on the basis of their syntactic patterns with respect to complex predicates.

3 Corpus Study

According to the best of our knowledge there is no systematic inventory of which types of nouns are allowed to combine with which types of light verbs in Urdu, though the basic problem has been recognized for Hindi by Hwang et al. (2010), who are developing annotation guidelines for complex predicate constructions. We used a small Part-of-Speech (POS) tagged corpus to extract a number of N-V complex predicates and then used native speaker judgements to further manually explore their ability to appear with each of the light verbs *kar* ‘do’, *ho* ‘be’, *hu-* ‘become’.[1] The manual exploration was necessary due to a data sparseness problem, since the available tagged corpora for Urdu are of a limited size.

3.1 Corpus

We used an Urdu POS tagged corpus compiled by the Center for Research in Urdu Language Processing (CRULP) in Lahore, Pakistan (available at http://www.crulp.org/software/ling_resources/UrduNepali-EnglishParallelCorpus.htm). The corpus consists of 100 000 words from the English Penn Treebank that have been (manually) translated into Urdu. The corpus consists of three files and the tag-set contains a specialized POS tag called VBL for the light verbs that are used in N-V complex predicates.

3.2 Method

We manually collected N-V complex predicates starting from the beginning of each of the corpus files. Given that we were interested in conducting an initial feasibility study, we stopped going through the files once we had collected 45 distinct nouns that appeared in N-V complex predicates containing the light verbs *kar* ‘do’, *ho* ‘be’ *hu-* ‘become’. We compiled a full set of combinatorial (im)possibilities of these 45 nouns with the three light verbs by taking the instances identified in the corpora and supplementing the “missing cells”, so to speak, via native speaker judgements as to whether the combination is possible.

An analysis of the resulting patterns did allow an identification of several distinct semantically coherent classes. Pertinent semantic factors appear to be stative vs. eventive nouns, agentivity vs. experiencer verbs (psych predications) and the licensing of a dative recipient.

4 Results

4.1 Class A: Full Range

4 out of 45 nouns allowed the full range of patterns shown in (1). The complex predicates these nouns appear in are psych verbs and include the nouns *yad* ‘memory’ and *yaqin* ‘belief’.

4.2 Class B: Exclusion of Dative Subjects

The bulk of the nouns, namely 38 out of the 45, allow an agentive (ergative) subject, but this subject does not alternate with a dative subject, as shown in (2).

(2) a. bılal=ne mɑkan tɑmir ki-ya
Bilal.M.Sg=Erg house.M.Sg.Nom construction.F.Sg do-Perf.M.Sg
‘Bilal built a/the house.’

[1] Further common light verbs are *de* ‘give’ and *a* ‘come’. These light verbs have a more complex distribution and so we chose to concentrate initially on just three basic and very common light verbs. Further light verbs could be investigated in an extension of this work.

b. *bɪlal=ko mɑkan tɑmir 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 nouns here are eventive nouns which presuppose an agent. As such, a non-agentive dative subject N-V complex predicate cannot be formed with this version of the noun. As shown in (3), grammatical combinations of these nouns with the light verb *hu-* ‘become’ do exist — this has an intransitivizing effect. Semantically, these are resultative state readings that are straightforwardly related to (2).

(3) mɑkan tɑmir hu-a/*hɛ
house.M.Sg.Nom construction.F.Sg be.Part-Perf.M.Sg
‘A/The house was/*is built.’

One noun in our set patterns essentially as shown in (2) and (3) with the difference that the noun licenses a dative recipient rather than a direct object (which can be marked as nominative or accusative, depending on the definiteness of the object in a well-known pattern of object alternation). In (3) the nominative object of (2a) is realized as a nominative subject. Similarly, as shown in (4), a dative object in a complex predicate with *kar* ‘do’ is realized as a dative subject when the light verb is *hu-* ‘become’. Other nouns in Urdu which display this pattern are: *ɪʃara* ‘signal’, *xabar* ‘news’ and *ɪnkar* ‘refusal’.

(4) a. nadya=ne bɪlal=ko ɪʃara ki-ya
Nadya.F.Sg=Erg Bilal.M.Sg=Dat signal.M.Sg do-Perf.M.Sg
‘Nadya signaled Bilal.’

b. bɪlal=ko ɪʃara hu-a
Bilal.M.Sg=Dat signal.M.Sg be.Part-Perf.M.Sg
‘Bilal was signaled.’ (lit.: A signal came to be at Bilal.’)

4.3 Class C: Exclusion of Light Verb *hu-* ‘become’

Another class (2 nouns in our set) allows for combinations with the light verbs *kar* ‘do’ and *ho* ‘be’, but not with *hu-* ‘become’, as illustrated in (5) for the noun *ɪntɪzar* ‘wait’. Other nouns like this are *taslim* ‘acceptance’ and *bardaʃt* ’tolerance’. Presumably the *hu-* ‘become’ does not work with these nouns because the subject is too agentive to be felicitous as the undergoer of a ‘become’ predication.

(5) a. bɪlal=ne nadya=ka ɪntɪzar ki-ya
Bilal.M.Sg=Erg Nadya.F.Sg=Gen.M.Sg wait.M.Sg do-Perf.M.Sg
‘Bilal waited for Nadya.’

b. bɪlal=ko nadya=ka ɪntɪzar 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.’

5 Discussion and Conclusions

Our corpus study showed that one can identify at least 3 different classes of nouns with one class consisting of at least two subclasses (Class B). The identification of classes was based on an investigation of their syntactic distribution in N-V complex predicates with respect to the light verbs *kar* ‘do’, *hu-* ‘become’ and *hɛ* ‘be’. A follow up study could include an extension of the set of light verbs. Another follow up study could look at the N-V complex predicates in relation to another set of light verbs which occur with V-V complex predicates. The N-V complex predicate is predicationally equivalent to a simple verb and as such can further combine with light verbs. Initial investigations have shown that the semantics of the noun governs the choice of this further light verb, so that the phenomenon of complex predicate stacking could provide further clues as to a semantic basis for the classification of Urdu nouns.[2]

The semantic factors identified so far include the eventive vs. statitivity of the nouns, the agentivity vs. experience of the action and whether the noun licenses a dative recipient. The first identification of noun classes in terms of systematic syntactic and semantic differences achieved in this paper represents a step towards overcoming the lack of lexical resources for natural language processing of Urdu.

References

Bhattacharyya, P. (2010). IndoWordNet. In *Proceedings of LREC2010*. Malta, May.

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

Butt, M., T. Bögel, A. Hautli, and S. Sulger (2009). Urdu and the modular architecture of ParGram. In *Proceedings of the Conference on Language and Technology 2009 (CLT09)*, pp. 1–7.

Butt, M. and T. H. King (2007). Urdu in a parallel grammar development environment. *Language Resources and Evaluation 41*, 191–207.

Butt, M. and G. Ramchand (2005). Complex aspectual structure in Hindi/Urdu. In N. Ertischik-Shir and T. Rapoport (Eds.), *The Syntax of Aspect*, pp. 117–153. Oxford: Oxford University Press.

Graddol, D. (2004). The future of language. *Science 303*, 1329–1331.

Humayoun, M. (2006). Urdu morphology, orthography and lexicon extraction. MSc Thesis, Department of Computing Science, Chalmers University of Technology.

Hwang, J. D., A. Bhatia, C. Bonial, A. Mansouri, A. Vaidya, N. Xue, and M. Palmer (2010). Propbank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW), ACL 2010*, Uppsala, Sweden, pp. 82–90.

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

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

Schulte im Walde, S. (2009). The induction of verb frames and verb classes from corpora. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

[2] A reviewer asks whether our method could scale up for larger corpora and whether resources such as WordNet could be used to assist the investigation. We here are faced with a lack of resources. We would first need a larger POS tagged corpora with a more differentiated POS tag set. However, in order to achieve this larger and more differentiated tagging, more information about the language is needed. We see our paper as contributing to this effort. With respect to WordNet, we face the problem that the classes provided for the English WordNet do not always match what we find in Urdu. In our Class B, nouns of communication form an identifiable subclass in Urdu and are also found to be related in English. However, the members of our Class C do not form a related net in English WordNet. With respect to Hindi WordNet, the ontology provided is not deep enough as yet to be able to provide useful information for investigations of this type.