Productive Encoding of Urdu Complex Predicates in the ParGram Project

Miriam Butt
Centre for Comp. Linguistics
UMIST
PO Box 88
Manchester M60 1QD UK
mutt@ccl.umist.ac.uk

Tracy Holloway King
NLTT/ISTL
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304 USA
thking@parc.com

John T. Maxwell III
NLTT/ISTL
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304 USA
maxwell@parc.com

Abstract

Complex Predicates are a crosslinguistically general phenomenon, but are more pervasive in South Asian than in European languages. This paper describes an LFG solution for Urdu/Hindi complex predication in terms of a RESTRICTION OPERATOR. The solution is theoretically well motivated and can be extended straightforwardly to related phenomena in European languages such as German, Norwegian, and French.

1 The ParGram Project

In this paper, we report on the implementation of complex predicates (CP) for Urdu in the Parallel Grammar (ParGram) project (Butt et al., 1999; Butt et al., 2002). The ParGram project originally focused on three European languages: English, French, and German. Three other languages were added later: Japanese, Norwegian, and Urdu. The ParGram project uses the XLE parser and grammar development platform (Maxwell and Kaplan, 1993) to develop deep grammars, i.e., grammars which provide an in-depth analysis of a given sentence (as opposed to shallow parsing or chunk parsing, where a relatively rough analysis of a given sentence is returned).

All of the grammars in the ParGram project use the Lexical-Functional Grammar (LFG) formalism, which produces c(onstituent)-structures (trees) and f(unctional)-structures (attribute-value matrices) as syntactic analyses. LFG assumes a version of Chomsky's Universal Grammar hypothesis, namely that all languages are governed by similar underlying structures. Within LFG, f-structures encode a language universal level of analysis, allowing for crosslinguistic parallelism. ParGram aims to see how far parallelism can be maintained across languages. In the project, analyses for similar constructions across languages are held as similar as possible. This parallelism requires the formulation of a rigid standard for linguistic analysis. This standardization has the computational advantage that the grammars can be used in similar applications, and it can simplify cross-language applications such as machine translation (Frank, 1999).

The conventions developed within the ParGram grammars are extensive. The ParGram project dictates not only the form of the features used in the grammars, but also the types of analyses chosen for constructions. The integration of new languages into the project has so far proven successful, including the adoption of the standards that were originally designed for the European languages (Butt and King, 2002b). As the new languages also contain constructions not necessarily found in the original European languages, the integration of new languages has contributed to the formulation of new standards of analysis. One such example is furnished by complex predicates in Urdu.

2 South Asian Complex Predicates

South Asian languages are known for the extensive and productive use of CPs. CPs combine a light verb with a verb, noun or adjective to produce a new verb. For example, Urdu has a large class of "aspectual" CPs which combine with verbs to change the aktionsart properties of the event. Examples are shown in (1b,c), cf. (1a).

(1) a. nAdyA AyI
Nadya-NOM came
'Nadya came.'

b. nAdyA A gayI
Nadya-NOM come went
'Nadya arrived.'

c. nAdyA A paRI
Nadya-NOM come fell
'Nadya came (suddenly, unexpectedly).'

The addition of a light verb modulates the event predication in subtle ways: beyond expressing defeasible meanings such as benefaction, suddenness, inception, or responsibility, the CP expresses a different aktionsart in comparison to the simple main verb. For example, in (1b) Nadya is in the result state of having arrived. The aktionsart effects of the light verbs on the event predication are quite complex and continue to be the subject

of on-going theoretical research (Butt and Ramchand, 2003). The general effect is the encoding of a result state (a song is in the state of having been sung, a person is in the state of having arrived). However, a result state can be interpreted in two differing ways depending on whether one wants to consider the event to come (inception), or the event that has passed (completion). The precise interpretation is lexically determined by the light verbs. For the purposes of the Urdu grammar, we mark light verbs like 'go' as signifying completion of an action, whereas light verbs like 'fall' signify inception.

Although these aspectual CPs do not alter the subcategorization frame of the verb, they change the resulting functional structure of the sentence, providing new information about the kind of event/action that is being described. The light verb also determines case marking on the subject: light verbs based on intransitive main verbs like *paR* 'fall' require a nominative subject. Light verbs like *lE* 'take' or *dE* 'give', which are based on (di)transitives main verbs, require an ergative subject. For example, transitive main verbs in the perfect tense usually require an ergative subject, as in (2a). When combined with a light verb like *paR* 'fall', the subject must be nominative as in (2b). Case marking in Urdu is governed by a combination of structural and semantic factors which we do not go into here (Butt and King, 2001). The light verb facts present an extension of the basic pattern.

(2) a. nAdyA nE gAnA gayA
Nadya-ERG song sang
'Nadya sang a song.'

b. nAdyA gAnA gA paRI
Nadya-NOM song sing fell
'Nadya burst into song.

c. nAdyA nE gAnA gA lIyA
Nadya-ERG song sing took
'Nadya sang a song (completely).'

As already mentioned, these CPs are extraordinarily productive in Urdu: most verbal predication involves complex predicate formation of the kind in (1) and (2). A light verb is in principle compatible with any main verb; however, (mostly semantic) selectional restrictions do apply so that some combinations are ruled out completely, whereas others are subject to considerable dialectal variation. Furthermore, the CPs are not formed within the lexicon, but are the result of the *syntactic composition* of two predicational elements (Alsina, 1996; Butt, 1995). Within LFG (as well as other syntactic frameworks), predicational elements play a special role: it is over these that argument saturation is checked. The difficulties involved with CP formation are better illustrated by means of another type of CP, the Urdu permissive, which alters the argument structure of the verb (Butt, 1995). The permissive light verb adds a new subject and "demotes" the other verb's subject to a dative-marked indirect object, as in (3b), cf. (3a).

(3) a. nAdyA sOyI
Nadya-NOM slept
'Nadya slept.'

b. yassin nE nAdyA kO sOnE dIA
Yassin-ERG Nadya-DAT sleep-INF gave
'Yassin let Nadya sleep.'

Since CPs are productive and occur frequently, an implementation that is both scalable and efficient is necessary. Most verbs can occur with several light verbs, and a given light verb can in principle occur with any verb of a given class (e.g., agentive verbs). So, it is not feasible to have multiple lexical entries for each verb depending on which light verb they occur with. This is especially true since the CPs combine with auxiliaries and other light verbs in predictable ways.

3 Implementation

The XLE implementation in use when Urdu joined ParGram allowed for basic modifications of predicates. In particular, it had an implementation of lexical rules that was sufficient to handle the English passive: argument grammatical functions could be renamed or deleted. An example of this is shown in (4) for the Urdu passive; the template is practically identical to that of English. In this template, _SCHEMATA indicates the predicate with grammatical functions of the verb (e.g., for transitive 'open': 'kHOl<(SUBJ)(OBJ)>'). In the active, nothing happens (left disjunct); in the passive, the object becomes the subject and the original subject is deleted (right disjunct).

(4) PASS(_SCHEMATA) =
{ _SCHEMATA | _SCHEMATA
(^ PASSIVE) = − (^ OBJ) −> (^ SUBJ)
(^ SUBJ) −> NULL
(^ PASSIVE) = + }.

However, this operation over lexical items is not sufficient to cover Urdu CPs. In the permissive, a subject is added and the predicate of the original verb is treated as an argument of the light verb, while at the same time assigning its arguments to the light verb. The problem of Urdu CPs is somewhat reminiscent of the head-switching type of structural mismatch discussed in the context of machine translation. The RESTRICTION operator has been proposed as a possible solution to the general problem of structural mismatches, with the Urdu permissive cited as a particular instance

(Kaplan and Wedekind, 1993). However, as first formulated, the solution only allowed the application of the restriction operator within the lexicon and thus did not take into account the powerfully recursive nature of complex predication in Urdu, which allows the different types of CPs to be stacked (Butt, 1994).

The need to treat a special type of Norwegian passive and the CPs in Urdu brought the issue of complex predication into the forefront of the discussions within the ParGram project. As part of these, a solution was found in the recent implementation of restriction within XLE (summer 2001) in which the restriction applies as part of the *syntactic composition* of two predicates.

Restriction allows f-structures and predicates to be manipulated in a controlled and detailed fashion. Given an f-structure like (5a), it might be necessary to restrict out the case information (e.g., in order to assign some other case to the f-structure, as with subject of the CP in (2b)). In this situation, the restriction operator '/' can be applied to the current f-structure (ˆ/CASE) in order to arrive at the restricted f-structure in (5b). A restricted f-structure is thus identical to the original f-structure except that it does not contain the restricted attribute.

(5) a.

$$\begin{bmatrix} \text{PRED} & \text{'nAdyA'} \\ \text{PERS} & 3 \\ \text{NUM} & \text{sg} \\ \text{CASE} & \text{erg} \end{bmatrix}$$

b.

$$\begin{bmatrix} \text{PRED} & \text{'nAdyA'} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix}$$

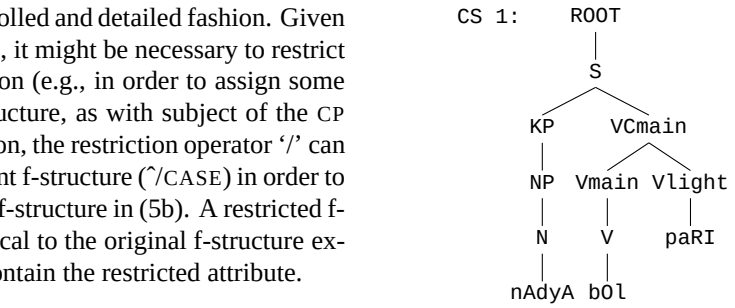
The Urdu grammar has pioneered the use of restriction. Since the implementation is recent (December 2002), the exact details of the CP analysis within the Urdu grammar are subject to change. One issue which remains to be fully resolved is the interaction of different types of light verbs and the modeling of the verbal complex as a whole. Since the verbal complex includes different kinds of auxiliaries (passive, progressive), modals, and light verbs which combine with main verbs, adjectives, and nouns, the collection of interacting phenomena is complex.

3.1 Aspectual Complex Predicates

An example of the current analysis of the aspectual CP in (6) is shown in (7) and (8). As mentioned above, in LFG, the syntactic analysis comprises two parts: a constituent-structure (a tree) and a functional-structure (an attribute-value matrix). The c-structure in (7) allows for a verbal complex which expands into a main verb followed by a light verb. There is no compelling evidence that Urdu has a VP (i.e., that a verb and its object are contained under one constituent), hence we do not assume one. Urdu is furthermore a language with fairly free word order, so the trees are quite flat: noun phrases are represented as sisters to one another under S (see the c-structures in (10) and (13)). We do assume KPs (Kase Phrases). Case markers in Urdu act as clitics to NPs (Butt and King, 2002a), and as such have their own phrase structure node. In (7) the subject is nominative, which is phonologically null, so the KP has an empty head. A full KP can be seen in the c-structure analysis for the permissive in (10).

(6) nAdyA bOl paRI
Nadya-NOM speak fell
'Nadya spoke up (suddenly, unexpectedly).'

(7) C-structure tree for aspectual CP



(8) F-structure AVM for aspectual CP

"nAdyA bOl paRI"

$$\begin{bmatrix} \text{PRED} & \text{'speak}\langle[0{:}\text{Nadya}]\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'Nadya'} \\ \text{NTYPE} & \begin{bmatrix} \text{PROPER name} \end{bmatrix} \\ \text{SEM-PROP} & \begin{bmatrix} \text{SPECIFIC} + \end{bmatrix} \end{bmatrix} \\ 0 & \text{CASE nom, GEND fem, NUM sg, PERS 3} \\ \text{TNS-ASP} & \begin{bmatrix} \text{ASPECT perf, INCEPTIVE} + \end{bmatrix} \\ \text{VMORPH} & \begin{bmatrix} \text{MTYPE infl} \end{bmatrix} \\ 19 & \text{PASSIVE} -, \text{ STMT-TYPE decl, VTYPE complex-pred} \end{bmatrix}$$

$$\begin{bmatrix} \text{PRED} & \text{'speak}\langle[0{:}\text{Nadya}]\rangle\text{'} \\ \text{SUBJ} & [0{:}\text{Nadya}] \\ \text{CHECK} & [\text{\_RESTRICTED} +] \\ \text{VMORPH} & [19\text{-VMORPH}] \\ 14 & \text{PASSIVE} -, \text{ STMT-TYPE decl, VTYPE unerg} \end{bmatrix}$$

The top f-structure in (8) represents the final analysis of the CP. The bottom f-structure shows the f-structure of the main verb *bOl* 'speak'. The features which have been restricted from the main verb's f-structure are VTYPE and TNS-ASP because these are the features which the light verb can "overwrite". In the case of (8), the TNS-ASP features are provided entirely by the light verb.

Within the ParGram project, the feature X-TYPE is used to encode distinctions within a given category X which are useful at the f-structure level of analysis. The English grammar, for example, encodes different kinds of adverbs (sentential, degree modifiers, etc.) via the feature ADV-TYPE. The feature VTYPE is used in the French grammar for auxiliary selection with unaccusative and unergative verbs. In the Urdu grammar, we use the feature VTYPE to register the type of the verbal predication. So, in (8), the final top structure has

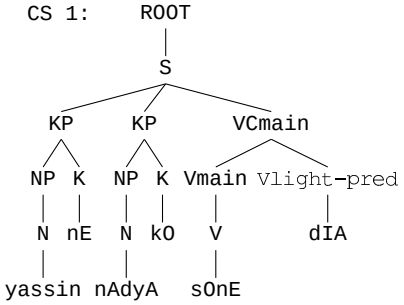
VTYPE `complex-pred`, while the lower structure for the main verb has VTYPE `unerg` because *bOl* 'speak' by itself is an unergative verb.

3.2 Permissive Complex Predicates

The restriction operation for permissive CPs is more interesting, as shown in the resulting f-structures in (11) for an intransitive main verb and in (14) for a transitive main verb.

(9) yassin nE nAdyA kO sOnE dIA
Yassin-ERG Nadya-DAT sleep-INF gave
'Yassin let Nadya sleep.'

(10) C-structure tree for permissive CP



(11) F-structure AVM for permissive CP

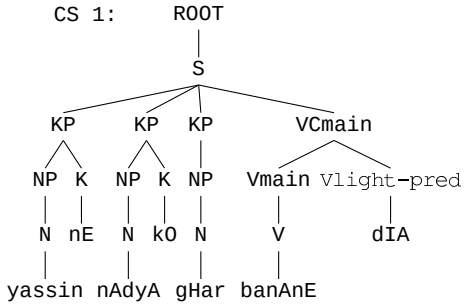
```
"yassin nE nAdyA kO sOnE dIA"
 PRED     'give<[0:Yassin] 'sleep<[16:Nadya]>'>'
          PRED     'Yassin'
          NTYPE    PROPER name
 SUBJ
          SEM-PROP SPECIFIC +
        0 CASE erg, GEND masc, NUM sg, PERS 3
          PRED     'Nadya'
          NTYPE    PROPER name
 OBJ-TH
          SEM-PROP SPECIFIC +
       16 CASE dat, GEND fem, NUM sg, PERS 3
 TNS-ASP  ASPECT perf, COMPLETIVE +
 VMORPH   MTYPE infl
51 PASSIVE -, STMT-TYPE decl, VTYPE complex-pred

   PRED   'sleep<[16:Nadya]>'
   SUBJ   [16:Nadya]
   CHECK  [_NMORPH obl, _RESTRICTED+]
   VMORPH [51-VMORPH]
32 PASSIVE -, STMT-TYPE decl, VFORM inf, VTYPE unerg
```

Recall that the light verb *dE* 'give' adds a subject argument and demotes the subject of the main verb to an indirect object. In addition to the VTYPE and TNS-ASP features, the PRED and SUBJ of the main verb's f-structure are thus also restricted. This allows the final f-structure to assign new grammatical functions when necessary, i.e., to demote the SUBJ Nadya to an OBJ-TH and to inherit any remaining arguments of the main verb. The light verb *dE* 'give' subcategorizes for a subject (the permitter) and a predicate. In (11), the PRED feature has the value of a *composite* argument structure, namely a combination of the subcategorization frame of *dE* 'give' (subject and another predicate) and the subcategorization frame of *sO* 'sleep' modulo the operations licensed via the restriction operator.

In (9) the main verb is the intransitive *sO* 'sleep' and so there are no arguments for the CP to inherit other than the demoted subject. The analysis in (14) shows what happens with a transitive main verb like *banA* 'make'.

(12) yassin nE nAdyA kO gHar banAnE dIA
Yassin-ERG Nadya-DAT house-NOM make-INF gave
'Yassin let Nadya build a house.'

(13) C-structure tree for permissive CP



(14) F-structure AVM for permissive CP

```
"yassin nE  nAdyA kO gHar banAnE dIA"
 PRED     'give<[0:Yassin] 'make<[16:Nadya] [32:gHar]>'>'
          PRED     'Yassin'
          NTYPE    PROPER name
 SUBJ
          SEM-PROP SPECIFIC +
        0 CASE erg, GEND masc, NUM sg, PERS 3
          PRED     'Nadya'
          NTYPE    PROPER name
 OBJ-TH
          SEM-PROP SPECIFIC +
       16 CASE dat, GEND fem, NUM sg, PERS 3
          PRED  'gHar'
          NTYPE GRAIN mass
 OBJ
       32 CASE nom, GEND masc, NUM sg, PERS 3
 TNS-ASP  ASPECT perf, COMPLETIVE +
 VMORPH   MTYPE infl
 72 PASSIVE -, STMT-TYPE decl, VTYPE complex-pred

   PRED   'make<[16:Nadya] [32:gHar]>'
   SUBJ   [16:Nadya]
   OBJ    [32:gHar]
   CHECK  [_NMORPH obl, _RESTRICTED+]
   VMORPH [72-VMORPH]
47 PASSIVE -, STMT-TYPE decl, VFORM inf, VTYPE agentive
```

The main verb *banA* 'make' has two arguments: a subject and an object. This is indicated in the bottom f-structure in (14). The top f-structure represents the final analysis. Here the SUBJ, PRED, and VTYPE features of the main verb's f-structure have been restricted. The VTYPE feature now states that this is a `complex-pred`. As in the previous example, the PRED feature has the value of a composite argument structure. This results in an overall three-place CP which subcategorizes for a subject via the subcategorization frame of *dE* 'give', an indirect object (OBJ-TH) which is the demoted subject of *banA* 'make', and finally an object which is inherited from the subcategorization frame of *banA* 'make'. Despite the fact that the arguments come from different sources and that the predication is complex (as evidenced by the nesting inside the PRED value in the top f-structure), at the level of f-structure, the arguments function like those of a simplex predicate (cf. Butt 1995).

4 Project Impact and Conclusions

The solution described above in terms of syntactic composition of arguments via the restriction operator allows the manipulation of subcategorization frames outside of the lexicon. This is particularly important as CPs in Urdu/Hindi and other languages are productive and separable in the syntax: they do not present instances of compounding or any other form of lexicalization. A sophisticated manipulation of subcategorization frames outside of the lexicon has not been previously possible, but finds clear applications for CPs crosslinguistically. A possible immediate application in the ParGram project would be to the well known problem of *suru* 'do' and other CPs found in Japanese. With respect to the European languages, the restriction operator opens up an innovative treatment of a subtype of the Norwegian passive, as in (15a), and allows for a potentially more satisfactory treatment of the German *lassen* 'let' construction, as in (15b), or the French causative *faire* 'make'.

(15) a. Kniven blir skåret kjøtt med.
the-knife is cut meat with
'The knife cut the meat.'

b. Der Fahrer hat den Traktor
the-NOM driver has the-ACC tractor
reparieren lassen.
repair let
'The driver had the tractor repaired.'

The current ParGram analyses treat these phenomena as instances of basic complement taking verbs, a solution which is not supported by the linguistic evidence and discussions amassed within theoretical linguistics.

The need to implement a productive analysis of CPs for Urdu resulted in the establishment of a new standard for analysis within the ParGram project: a scalable and efficient solution for the general phenomenon of complex predication is now available to the grammar writers for all of the project languages. In addition, passive and causative, which are currently treated via lexical rules in the grammars, could be reimplemented using restriction, simplifying the verbal lexical entries. Thus, we see that a change required for one language, in this case the South Asian language Urdu, can benefit the implementations of many languages.

References

Alex Alsina. 1996. *The Role of Argument Structure in Grammar*. CSLI Publications.

Miriam Butt and Tracy Holloway King. 2001. Nonnominative subjects in Urdu: A computational analysis. In *Proceedings of the International Symposium on Non-nominative Subjects*, pages 525–548, Tokyo. ILCAA.

Miriam Butt and Tracy Holloway King. 2002a. The status of case. In Veneeta Dayal and Anoop Mahajan, editors, *Clause Structure in South Asian Languages*. Kluwer Academic Publishers, Dordrecht. To Appear.

Miriam Butt and Tracy Holloway King. 2002b. Urdu and the Parallel Grammar project. In *Proceedings of COLING 2002*. Workshop on Asian Language Resources and International Standardization.

Miriam Butt and Gillian Ramchand. 2003. Building complex events in Hindi/Urdu. In Nomi Ertischik-Shir and Tova Rapoport, editors, *The Syntax of Aspect*. Oxford University Press, Oxford. Submitted.

Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. CSLI Publications.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar project. In *Proceedings of COLING 2002*. Workshop on Grammar Engineering and Evaluation.

Miriam Butt. 1994. Machine translation and complex predicates. In *Proceedings of KONVENS 94*, pages 62–71.

Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu*. CSLI Publications.

Anette Frank. 1999. From parallel grammar development towards machine translation. In *Proceedings of MT Summit VII*, pages 134–142.

Ron Kaplan and Jürgen Wedekind. 1993. Restriction and correspondence-based translation. In *Proceedings of the Sixth European Conference of the Association for Computational Linguistics*, pages 193–202.

John T. Maxwell, III and Ron Kaplan. 1993. The interface between phrasal and functional constraints. *Computational Lingusitics*, 19:571–589.