A Reference Dependency Bank for Analyzing Complex Predicates

Tafseer Ahmed, Miriam Butt, Annette Hautli, Sebastian Sulger

Universität Konstanz
firstname.lastname@uni-konstanz.de

Abstract

When dealing with languages of South Asia from an NLP perspective, a problem that repeatedly crops up is the treatment of complex predicates. This paper presents a first approach to the analysis of complex predicates (CPs) in the context of dependency bank development. The effort originates in theoretical work on CPs done within Lexical-Functional Grammar (LFG), but is intended to provide a guideline for analyzing different types of CPs in an independent framework. Despite the fact that we focus on CPs in Hindi and Urdu, the design of the dependencies is kept general enough to account for CP constructions across languages.

Keywords: Complex Predicates, Dependency Bank, Lexical-Functional Grammar, ParGram, South Asian languages, Urdu/Hindi

1. Introduction and motivation

In building NLP applications for the less well-resourced languages of the world, such as the languages of South Asia, a problem that crops up repeatedly is how to deal with complex predications.[1] Many languages of the world use a combination of more than one predicate to express concepts that in languages like English are expressed with a single verb, e.g., 'memory do' = 'remember', 'fear come' = 'fear', 'clean do' = 'clean'. In addition, verb+verb combinations are used to express permissive, causative or aspectual relations. The issue of complex predication also comes up with respect to languages like English and German, e.g., English constructions like *take a bath, do a Chomsky, give a stir* or German *eine Rede halten* 'a speech hold=give a speech'. However, in languages like German and English, these constructions do not constitute the majority of predicational instances and so are generally discussed in terms of further complex phenomena that need to be dealt with appropriately eventually, once basic lexical resources are in place. In contrast, the verbal system of South Asian languages and many other languages of the world is constructed quite differently. For example, there are only about 700 basic verbs in Urdu and the vast majority of verbal predication is achieved via complex predicational structures. As such, the need to understand how to represent complex predicates arises almost immediately in the course of building an NLP application.

In this paper, we seek to ameliorate the problem by providing a reference dependency bank for complex predicates, based on our in-depth analysis of the South Asian language Urdu (Butt, 1995; Butt and King, 2007; Ahmed and Butt, 2011). Urdu/Hindi[2] allows for the following types of complex predicates: aspectual and permissive V-V CPs, morphological causatives and different types of N-V, ADJ-V and P-V CPs. South Asian languages form an areally defined *Sprachbund* (Masica, 1993) and most other South Asian languages exhibit these or different subtypes of complex predicates. As such, we seek to provide a reference dependency bank for complex predicates that can be used as a guide by teams working on NLP applications primarily for other South Asian languages, but also for other relevant languages across the world. In order to make the reference dependency bank as useful as possible, we also provide reference analyses for related constructions, namely: modals and auxiliaries, which look like complex predicates on the surface but are not. Complex predicates are often analyzed on a par with these constructions, but their syntactic distribution in fact differs markedly from that of complex predicates (see, e.g., Butt (2010) for an overview of the structure of complex predication crosslinguistically). NLP efforts which do not take the difference into account tend to yield unsatisfactory analyses and POS tag sets that are too vague to be truly useful.

All complex predicate (CP) analyses described in this paper are implemented in the Urdu ParGram grammar (Butt et al., 1999; Butt and King, 2007; Bögel et al., 2009); although the grammar is couched within Lexical-Functional Grammar (LFG) (Dalrymple, 2001), we keep the design of the dependency bank as general as possible, following the example set by King et al. (2003) with PARC700; thereby providing a crosslinguistic and theoretically independent analysis of CPs. We also see our effort as a seed resource to which reference representations of further CP types crosslinguistically can be added, thus eventually resulting in a complete typology of CPs that can be used as a resource for a variety of NLP applications. The dependency bank may be downloaded freely from the web.[3]

We provide a sketch of four types of CPs: the aspectual V-V, the permissive V-V, N-V CPs, morphological causatives and ADJ-V CPs. These types are presented in section 2. We then provide examples of verbal complexes that are often mistaken for complex predicates, but in fact are not (section 3.). In section 4. we describe the format and the design of the dependency bank. Section 5 concludes the paper.

[1] We would like to gratefully acknowledge the DFG (Deutsche Forschungsgemeinschaft), for the funding that enabled this work via our project on Urdu grammar development.

[2] Hindi and Urdu are mostly parallel with respect to syntax and semantics, some differences exist with respect to vocabulary and derivational morphology.

[3] `http://ling.uni-konstanz.de/pages/home/pargram_urdu/main/Resources.html`

2. Types of Complex Predicates

This section goes through different types of possible complex predicates and describes each of them briefly, including some combinatory possibilities. In addition to the two types of V-V complex predicates discussed below, the standard literature acknowledges two different types of N-V CPs (agreeing vs. non-agreeing) as well as an ADJ-V CP (Mohanan, 1994). We have additionally identified a P-V CP. Since the P-V CP so far appears to be quite rare, we do not discuss it here.

2.1. An aspectual V-V complex predicate

Aspectual CPs are formed by combining two verbs as coheads: a main verb in its bare form and an inflected light verb. The light verb contributes aspectual meaning to the predicate, providing information about the kind of action/event that is described; it does not, however, add an argument or arguments to the overall subcategorization frame of the predication. Therefore, the number of arguments of the resulting CP is the same as that of the main verb. However, the finite light verb does determine the case marking of the subject and semantic selectional constraints governing the combinatory possibilities of main and light verb must be observed. See Butt (1995) and Butt et al. (2003) for detailed discussion and further references. (1a) is an example of a simple intransitive verb; (1b) is an example of a verb-verb aspectual CP.[4]

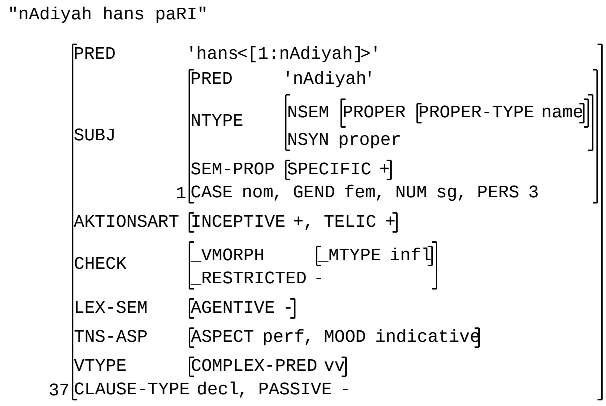
(1) a. nAdiyah hans-I
Nadya.F.Sg.Nom laugh-Perf.F.Sg
'Nadya laughed.'

b. nAdiyah hans paR-I
Nadya.F.Sg.Nom laugh fall-Perf.F.Sg
'Nadya burst out laughing.'

Figure 1 provides an LFG f(unctional)-structure analysis of (1b). The f-structure is essentially a dependency structure which expresses not only basic predicate-argument relations, but also encodes detailed information about tense/aspect, Aktionsart, mood, case, clause type, verb type, etc.[5]

Looking at the top-level PRED in Figure 1, we can see that the main verb's argument structure has not been altered. The aspectual light verb *paR* 'fall' contributes only aspectual information about the event; this information is encoded in the f-structure under the AKTIONSART feature.

This type of complex predicate is extremely common in Urdu/Hindi as well as in many of the other South Asian languages. The light verb not only provides information about Aktionsart, but also contains contextually dependent information about whether the action was to somebody's benefit, responsibility for the action, forcefulness of the action, etc. In Urdu/Hindi, a set of about 24 verbs can act as light verbs. Common ones are 'take, give, fall, rise, sit, hit'. Since the type of information contributed by these light verbs in South Asian languages is very contextually dependent and defeasible, it is not represented in our syntactic analysis.



```
"nAdiyah hans paRI"

    PRED        'hans<[1:nAdiyah]>'
                PRED     'nAdiyah'
                NTYPE    [NSEM [PROPER [PROPER-TYPE name]]
                         [NSYN proper
    SUBJ        SEM-PROP [SPECIFIC +]
              1 CASE nom, GEND fem, NUM sg, PERS 3
    AKTIONSART  [INCEPTIVE +, TELIC +]
    CHECK       [_VMORPH     [_MTYPE infl]
                [_RESTRICTED -
    LEX-SEM     [AGENTIVE -]
    TNS-ASP     [ASPECT perf, MOOD indicative]
    VTYPE       [COMPLEX-PRED vv]
  37CLAUSE-TYPE decl, PASSIVE -
```

Figure 1: LFG dependency (f-structure) analysis for (1b)

2.2. A permissive V-V complex predicate

The Urdu permissive is a combination of an infinitive verbal predicate with a finite verb. The two verbs contribute to the overall argument structure of the clause. For example, in (2) there is a single CP *dEkH dE* 'let see', composed of *dEkH* 'see' and *dE* 'give'. Two arguments come from the verb *dEkH* 'see', the seer and the seen item, and two arguments are provided by the verb *dE* 'give': the person giving permission and the action that is permitted. The total of four arguments combine into just three arguments in the CP: the person giving permission is the subject in (2), the seer and the person granted permission fall together as the indirect object (OBJ-TH or OBJ-GO in LFG) and the thing seen is the overall object of the clause. The action that is permitted is analyzed as an argument of *dE* 'give', but is also part of the complex predication, as shown in the top-level PRED value in Figure 2.

(2) nAdiyah nE yAsIn kO
Nadya.F.Sg Erg Yassin.M.Sg Dat
kitAb dEkH-nE d-I
book.F.Sg.Nom see-Inf.M.Sg.Obl give-Perf.F.Sg
'Nadya let Yassin look at the book.'

Note in particular that such constructions differ syntactically from biclausal embedding constructions such as the one in (3), where the agreement facts show that *ciTTHI* 'note' is an embedded object, unlike *kitAb* 'book' in (2).

(3) nAdiyah nE yAsIn kO
Nadya.F.Sg Erg Yassin.M.Sg Dat
[ciTTHI likH-nE kO] kah-A
note.F.Sg.Nom write-Inf.M.Sg.Obl Acc say-Perf.M.Sg
'Nadya told Yassin to write the note.'

For a full discussion of the syntactic properties of the permissive construction, see Butt (1995), who provides some tests for monoclausality vs. biclausality that may also be useful in determing complex predicate status for languages other than Urdu/Hindi.

[4] The transliteration scheme employed for representing the Arabic script of Urdu is described in Malik et al. (2010).

[5] All the f-structures shown in this paper have been generated by the Urdu ParGram grammar. Note that the CHECK features in the f-structures are used to collect features that are only used for well-formedness checking — these provide no independent functional information.

```
"nAdiyah nE yAsIn kO kitAb dEkHnE dI"

   PRED       'dE<[1:nAdiyah], 'dEkH<[21:yAsIn], [41:kitAb]>'>'
              PRED      'nAdiyah'
              NTYPE     [NSEM [PROPER [PROPER-TYPE name]]]
   SUBJ                 [NSYN proper                    ]
              SEM-PROP  [SPECIFIC +]
             1CASE erg, GEND fem, NUM sg, PERS 3
              PRED      'yAsIn'
              NTYPE     [NSEM [PROPER [PROPER-TYPE name]]]
   OBJ-GO               [NSYN proper                    ]
              SEM-PROP  [SPECIFIC +]
            21CASE dat, GEND masc, NUM sg, PERS 3
              PRED  'kitAb'
   OBJ        NTYPE [NSEM [COMMON count]]
                    [NSYN common        ]
            41CASE nom, GEND fem, NUM sg, PERS 3
   LEX-SEM    [AGENTIVE +, GOAL +]
   TNS-ASP    [ASPECT perf, MOOD indicative]
   VTYPE      [COMPLEX-PRED vv-perm]
 83CLAUSE-TYPE decl, PASSIVE -, PERS 3
```

Figure 2: LFG dependency (f-structure) analysis for (2)

2.3. A N-V complex predicate

For purposes of illustration, we provide a well-known example of a Hindi/Urdu N-V CP in (4).

(4) nAdiyah nE kahAnI
Nadya.F.Sg Erg story.F.Sg.Nom
yAd k-I
memory.F.Sg.Nom do-Perf.F.Sg
'Nadya remembered a/the story.'

In (4), the verb *kar* 'do' provides two arguments to the complex predication, namely, the doer and the action done. The noun *yAd* 'memory' contributes one further argument: the thing remembered. These three arguments combine into two in the syntax, namely the doer/rememberer as the subject, the thing remembered as the object. The performed action is again encoded as an argument of the verb as part of the complex predication, namely the top-level PRED in the analysis in Figure 3.

```
"nAdiyah nE kahAnI yAd kI"

   PRED     'kar<[1:nAdiyah], 'yAd<[21:kahAnI]>'>'
            PRED      'nAdiyah'
            NTYPE     [NSEM [PROPER [PROPER-TYPE name]]]
   SUBJ               [NSYN proper                    ]
            SEM-PROP  [SPECIFIC +]
           1CASE erg, GEND fem, NUM sg, PERS 3
            PRED  'kahAnI'
   OBJ      NTYPE [NSEM [COMMON count]]
                  [NSYN common        ]
          21CASE nom, GEND fem, NUM sg, PERS 3
   LEX-SEM  [AGENTIVE +]
   TNS-ASP  [ASPECT perf, MOOD indicative]
   VTYPE    [COMPLEX-PRED nv]
 57CLAUSE-TYPE decl, PASSIVE -
```

Figure 3: LFG dependency (f-structure) analysis for (4)

Note that the above construction differs from simple transitive clauses such as the one in (5). Here, the noun *kahAnI* 'story' does not contribute any additional arguments — the simple transitive verb *paRH* 'read' selects its two arguments *nAdiyah* and *kahAnI*, the reader and the thing read.

(5) nAdiyah nE kahAnI paRH-I
Nadya.F.Sg Erg story.F.Sg.Nom read-Perf.F.Sg
'Nadya read a/the story.'

The CP *yAd+ kar* is an example of a non-agreeing N-V CP where the noun *yAd* 'memory' is feminine and seems at first sight to be agreeing with the verb. However, the verb is in fact agreeing with *kahAnI*. This can be seen clearly when one prevents agreement from occuring by adding an overt accusative case marker to *kahAnI*, as in (6).[6] In (6) the verb has default masculine agreement as it is not agreeing with any noun in the sentence.

(6) nAdiyah nE kahAnI kO
Nadya.F.Sg Erg story.F.Sg Acc
yAd ki-yA
memory.F.Sg.Nom do-Perf.M.Sg
'Nadya remembered the story.'

In contrast, in (7), the verb agrees with the noun that is part of the CP, namely the feminine noun *bah2as2*. This is an example of an agreeing N-V CP.

(7) meNDak nE biccHU sE
frog.M.Sg Erg scorpion.M.Sg Inst
bah2as2 k-I
debate.F.Sg.Nom do-Perf.F.Sg
'The frog argued with the scorpion.'

[6] The presence of overt case markers in Urdu generally blocks agreement — agreement can in principle be with either an unmarked subject or an unmarked object in that order of preference.

We follow Mohanan (1994), who analyzes this type of noun as being the grammatical object of the clause (which accounts for the agreement pattern) but also, simultaneously, as part of the N-V CP. That is, it plays two roles in the clause. Additionally, Ahmed (2011) shows that some of the nouns within the agreeing N-V CPs class allow modifiers, thus identifying a further subclass of N-V CPs.

The light verbs commonly used with N-V CPs are 'do, be, become, stay and keep'. It should be noted that there are many instances of N-V sequences which look like CPs but are actually simple transitive sentences as in (5). Example (8) looks like a CP on the surface because of the verb 'do', which also allows for a light verb use as in (4), (6) and (7).

(8) nAdiyah nE kAm ki-yA
Nadya.F.Sg Erg work.M.Nom do-Perf.M.Sg
'Nadya did some/the work.'

However, a crucial difference is that in the examples we have identified as CPs, we can point to an extra argument in the clause that is being contributed by the noun. This is not the case in (8) and we therefore do not analyze this as a CP.

2.4. Adj-V sequences

Similarly, with ADJ-V sequences we have examples where the adjective contributes an argument of its own and ones where this situation does not obtain. We consider the former as CPs, but not the latter. In (9), we have an example of an ADJ-V CP. Here, the verb 'do' combines with the adjective *alag* 'separate', which introduces its own argument (*kHiRkI* 'window'). Evidence from scrambling, etc. speaks in favor of an analysis in which this extra argument should be treated as a clause-level argument rather than being embedded as an argument of the adjective. The f-structure analysis is provided in Figure 4.

(9) yAsIn nE SISE kO
Yasin.M.Sg Erg glass.M.Sg.Obl Acc
kHiRkI sE alag ki-yA
window.F.Sg Inst separate.Nom do-Perf.M.Sg
'Yassin removed the glass from the window.'

On the other hand, in examples as in (10) and (11), the adjective does not contribute any extra arguments of its own. We therefore do not consider these to be CPs.

(10) mEz s3Af he
table.F.Sg.Nom clean.Nom be.Pres.3.Sg
'The table is clean.'

(11) yAsIn nE mEz s3Af
Yasin.M.Sg Erg table.F.Sg.Nom clean.Nom
k-I
do-Perf.F.Sg
'Yasin cleaned the table.'/'Yasin made the table clean.'

However, all of these examples are special in a different sense – they are *resultatives*. The result of doing something to the table or to the glass in the examples above is that they become clean or separate. Resultatives are known as instances of *secondary predication*. In our representations, we have accounted for this by introducing the grammatical relation of a PREDLINK (cf. Butt et al. (1999)). The PREDLINK is what is predicated of a certain entity, e.g., the objects in the examples above. That is, the table is predicated to be clean, the glass separate from the window. However, this relationship is not shown overtly in the f-structure, as the interpretation of resultatives as well as depictives crosslinguistically tends to be context dependent. The relationship must be inferred through the presence of the PREDLINK, which must be related to one of the other grammatical relations in the clause, in our examples the OBJ.

2.5. Causatives

Languages may have morphological or periphrastic causatives. Some periphrastic causatives may qualify as complex predicates (i.e., be functionally monoclausal), while others may not. The status of periphrastic causatives thus needs to be investigated carefully on a language-by-language basis. Morphological causatives, on the other hand, tend to be monoclausal. In both morphological and periphrastic causatives the arguments of the clause are contributed by different elements.

Urdu is a language with morphologically formed causatives, where transitive verbs have an intransitive root form (12a) that is related to the transitive via vowel lengthening or the suffixation of an *-A* morpheme (12b). The indirect causative is formed by the morphme *-vA*, as shown in (12c).

(12) a. makAn ban-A
house.M.Sg.Nom be made-Perf.M.Sg
'The house was built.'

b. nAdiyah nE makAn
Nadya.F.Sg Erg house.M.Sg.Nom
ban-**A**-yA
be made-Caus-Perf.M.Sg
'Nadya built a house.'

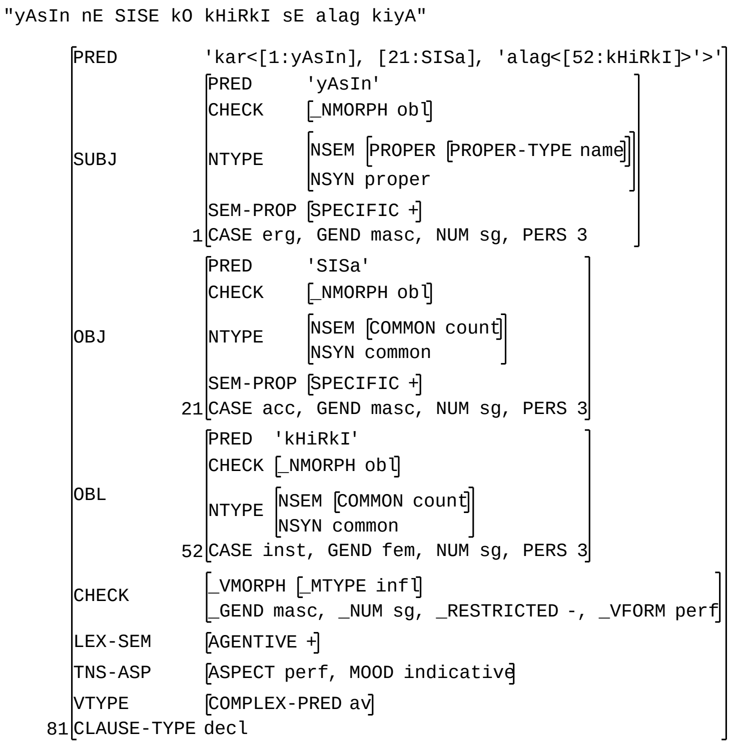
c. nAdiyah nE (mAzdUrON sE)
Nadya.F.Sg Erg (laborer.M.Pl.Obl Inst)
makAn ban-**vA**-yA
house.M.Sg.Nom be made-Caus-Perf.M.Sg
'Nadya had a house built (by the laborers).'

In (12b), the agent/subject argument *nAdiyah* is licensed by the causative morpheme, in (12c) the optional instrumental is additionally licensed by the *-vA* causative morpheme. The patterns of causation are highly dependent on verb class and are fairly complex. For a brief survey of the relevant patterns and details on the implementation of causatives within the Urdu ParGram grammar see Butt and King (2006).

2.6. Combinations of all of the above

We have now surveyed the major types of complex predication. All the types discussed above tend to be common in the languages of South Asia. Equally common is the possibility of combining the several different types of complex predications. We provide an example in (13).[7]

[7] Discussing all the various types of possible combinations and restrictions is beyond the scope of this paper as the combination-



```
"yAsIn nE SISE kO kHiRkI sE alag kiyA"

  PRED        'kar<[1:yAsIn], [21:SISa], 'alag<[52:kHiRkI]>'>'
              PRED      'yAsIn'
              CHECK     [_NMORPH obl]
  SUBJ        NTYPE     NSEM [PROPER [PROPER-TYPE name]]
                        NSYN proper
              SEM-PROP  [SPECIFIC +]
             1CASE erg, GEND masc, NUM sg, PERS 3
              PRED      'SISa'
              CHECK     [_NMORPH obl]
  OBJ         NTYPE     NSEM [COMMON count]
                        NSYN common
              SEM-PROP  [SPECIFIC +]
            21CASE acc, GEND masc, NUM sg, PERS 3
              PRED  'kHiRkI'
              CHECK [_NMORPH obl]
  OBL         NTYPE NSEM [COMMON count]
                    NSYN common
            52CASE inst, GEND fem, NUM sg, PERS 3
  CHECK       _VMORPH [_MTYPE infl]
              _GEND masc, _NUM sg, _RESTRICTED -, _VFORM perf
  LEX-SEM     [AGENTIVE +]
  TNS-ASP     [ASPECT perf, MOOD indicative]
  VTYPE       [COMPLEX-PRED av]
81CLAUSE-TYPE decl
```

Figure 4: LFG dependency (f-structure) analysis for (9)

(13) yAsIn nE laRkE kO gHar
Yassin.M.Sg Erg boy.M.Sg.Obl Acc house.M.Sg.Nom

ban-A-nE di-yA
make-Caus-Inf.M.Obl give-Perf.M.Sg
'Yassin let the boy build a house.'

As can be seen in Figure 5, this example consists of a causative in combination with a permissive. That is, the permissive 'give' introduces the subject argument 'Yassin' and takes a verbal predication as an argument. This verbal predication is in turn complex as it contains a causative morpheme, which licenses the indirect object 'boy'. Finally, the intransitive verb *ban* 'be made' is responsible for the object 'house'.

3. Other Verbal Complexes

Complex predicates are often confused with other verbal constructions such as modals and auxiliaries, although their syntactic distribution differs markedly. The CPs described above form verbal complexes which have special properties in terms of how the arguments of the overall clause are licensed — each part of the complex predicate contributes to the overall argument structure. The aspectual V-V CPs discussed in section 2.1. would seem to be an exception. However, as shown in Butt (1995), the light verbs here do have an effect on the overall argument structure in that they ultimately determine the case marking of the subject and place constraints on what kind of verbs they can combine with. This is very much unlike the behavior exhibited by auxiliaries (section 3.2.) and modals (section 3.1.), which can combine with mostly any verb.

In this paper and in our dependency bank, we provide some examples of modals and auxiliaries in order to alert researchers to the fact that not all verbal complexes are automatically complex predicates and as to how the structure of complex predicates contrasts with other verbal complexes.

3.1. Modals

Modality in Urdu/Hindi is mostly expressed constructionally and the language features only two dedicated modals, *cahiyE* 'need' (14) and *sak* 'can' (15). Modal expressions in Urdu have been analyzed as instances of either control or raising (Bhatt et al., 2011) and show no common behavior with complex predicates. We follow the general LFG analysis in that the modal verb is the main predicate and subcategorizes for an XCOMP (a non-finite complement) in the f-structure. That is, each verb (e.g. 'speak' and the modal in the examples below) is the head of its own clausal domain. The finite modal verb embeds the other, non-finite verb so that the structure is biclausal.

(14) yAsIn [kitAb paRH]
Yassin.M.Sg.Nom book.F.Sg.Nom read

sak-A
able-Perf.M.Sg
'Yassin was able to read a/the book.'

atory possibilities are impressive — all the types of the complex predicates discussed here can in principle be combined with one another, though there are restrictions on the order of combination.

```
"yAsIn nE laRkE kO gHar banAnE diyA"

    PRED      'dE<[1:yAsIn], 'A-CAUSE<[21:laRkA], 'ban<[52:gHar]>'>'>'
              PRED      'yAsIn'
              CHECK     [_NMORPH obl]
    SUBJ      NTYPE     [NSEM [PROPER [PROPER-TYPE name]]]
                        [NSYN proper                     ]
              SEM-PROP  [SPECIFIC +]
             1CASE erg, GEND masc, NUM sg, PERS 3
              PRED  'laRkA'
              CHECK [_NMORPH obl]
    OBJ-GO    NTYPE [NSEM [COMMON count]]
                    [NSYN common        ]
            21CASE dat, GEND masc, NUM sg, PERS 3
              PRED  'gHar'
    OBJ       NTYPE [NSEM [COMMON count]]
                    [NSYN common        ]
            52CASE nom, GEND masc, NUM sg, PERS 3
    CHECK     [_VMORPH [_MTYPE infl]          ]
              [_RESTRICTED -, _VFORM perf]
    LEX-SEM   [AGENTIVE +, GOAL +]
    TNS-ASP   [ASPECT perf, MOOD indicative]
    VTYPE     [COMPLEX-PRED vv-perm]
 106CLAUSE-TYPE decl, PASSIVE -, PERS 3
```

Figure 5: LFG dependency (f-structure) analysis for (13)

(15) yAsIn kO [kitAb paRH-nI]
Yassin.M.Sg Dat book.F.Sg.Nom read-Inf.F.Sg

cahiyE
need.Sg
'Yassin should read a/the book.'

On the surface, these verbal complexes may look very similar to complex predicates, however, their syntactic and semantic behavior differs markedly. For one, there is always a clearly modal meaning, in contrast to the complex predicates. For another, there are no complex selectional restrictions between modal and the non-finite verb and there is no merging of predicational domains — each keeps their own and the subject of the embedded clause is controlled according to standard patterns found crosslinguistically (Bresnan, 1982). In our dependency structures, the modal interpretation is captured (both at the f-structures and the triples) by a special feature encoding the type of modality, thereby further distinguishing them from complex predicate constructions.

3.2. Auxiliaries

Both auxiliaries and light verbs in Urdu can be used as main verbs. If one goes just by surface form, a proper identification of the syntactic status and behavior of the verbs becomes confusing. Auxiliaries differ from light verbs and main verbs in that they do not contribute any of their own arguments/participants to the clause. Auxiliaries are responsible only for contributing tense and aspect information to a clause. In Urdu, the auxiliary complex can become quite long and difficult to interpret, yet each piece is really only contributing information about the duration, iteration, etc. of an action. In the case of (16), for example, the main verb is *bOl* 'talk'. The verb *cal* 'walk' signals that the action is a long one, *jA* 'go' in its auxiliary form contributes an iterative dimension to the action, whereas *rah* 'stay' is the progressive marker.

(16) nAdiyah bOl-tI cal-I
Nadya.F.Sg talk-Impf.F.Sg walk-Perf.F.Sg

jA rah-I he
go stay-Perf.F.Sg be.Pres.3.Sg
'Nadya is talking (repeatedly, over a long time).'

In order to do justice to non-complex predicate verbal complexes such as in (17), our dependency bank encodes fine-grained aspectual information based on the features represented at the f-structures.

4. A reference dependency bank for complex predicates

The sentences contained in the reference dependency bank illustrate examples of all common CP types in Urdu as well as prototypical examples of auxiliary constructions and modals so that NLP researchers will be able to use the prototype analyses in order to model data in further South Asian languages without needing to rediscover that syntactic distributions and properties differ across these constructions (a situation that often arises at the moment).

The sentences included in our CP reference bank are first parsed with the Urdu ParGram grammar (Butt et al., 1999; Butt and King, 2007; Bögel et al., 2009). As a parse, the grammar produces a c(onstituency)-structure (generally known as the "tree" structure) and an f-structure, which encodes dependencies in the form of an AVM. For the purpose of a reference dependency bank, only the f-structures are of importance. These are banked using the LFG parsebanker (Rosén et al., 2009), where ambiguous analyses can be easily disambiguated via the manual selection of discriminants.

Our dependency bank is based on the triples format of PARC700 (King et al., 2003) which provides a theory-independent way of encoding the interrelationships within a sentence and also enables other parsers to evaluate their analysis against the reference bank. The XLE internal mechanism allows for the f-structure facts to be stored in the triples format. In addition, XFR rewrite rules can rewrite or flatten f-structure features so that the set of triples is adjusted (Crouch et al., 2012).

These adjustments are done in particular for the representation of the verbal complex in the triples. The general methodology is that every part of a complex predicate (either a V-V, a N-V or an ADJ-V CP) that contributes some argument structure is concatenated by an underscore to make clear that the whole complex is the main predicate of the clause (e.g. the triple `pred(root,dEkH_dE)` in (17)). In cases where a verb only contributes aspectual information and no arguments as in (1b), the information captured in the f-structure under `Aktionsart` is retained in the dependency triples.

Apart from the verbal information, we restrict the set of triples to predicate-argument relations and neglect the more detailed information in the f-structure. The predicate-argument structure is split into two parts in order to keep the dependency bank as transparent as possible: for one, we list the grammatical relations of the whole predicational domain, but we also indicate which part of the CP contributes what argument in the sentence. The parts of the CP are labeled consecutively, based on their linear position in the concatenated main clause predicate.

Based on the f-structure in Figure 2 for the permissive V+V CP, the resulting triples to be included in the reference dependency bank are shown in (17). The f-structure information has been reduced to only include the predicate-argument structure and the necessary CP type information. The complex predicate of the main clause is `dEkh_dE`, with its grammatical relations `subj,obj` and `obj-go`. The nature of the V-V is captured by the triple `complex-pred-type(dEkH_dE,vv-perm)`.

In order to represent the argument structure of each part of the complex predicate, the concatenated predicate is split up into its parts, with information about which arguments are being contributed by which part. Due to the number based on the linear precedence in the concatenated predicate, *dEkH* 'see' is labeled as the first part of the complex predicate (`cp-part1`) and the permissive light verb *dE* 'give' is labeled as the second part of the CP (`cp-part2`). Each part of the complex predicate is assigned a list of `arg1` and `arg2` facts that should not be confused with PropBank style argument numbering (Palmer et al., 2005), instead they are simply a naming convention for identifying the arguments of a predicate. For example, the finite verb, *dE* 'give' has two arguments, *nAdiyah* (`arg1`) and the verb `dEkh` 'to see' (`arg2`). *dEkh* in turn has also two arguments, *yAsIn* (`arg1`) and *kitAb* 'book' (`arg2`). As a convention, the highest predicate in the f-structure (here, `dE` 'give') is put first in the triples set, followed by the verb that it subcategorizes for (here, `dEkH` 'see').

```
(17) pred(root,dEkH_dE)
     subj(dEkH_dE,nAdiyah)
     obj(dEkH_dE,kitAb)
     obj-go(dEkH_dE,yAsIn)
     complex-pred-type(dEkH_dE,vv-perm)
     cp-part1(dEkH_dE,dEkH)
     cp-part2(dEkH_dE,dE)
     arg1(dE,nAdiyah)
     arg2(dE,dEkH)
     arg1(dEkH,yAsIn)
     arg2(dEkH,kitAb)
     asp(dEkH_dE,perf).
```

Similarly, for the N+V CP in Figure 3, we arrive at the triples dependency representation shown in (18). Again, the triples contain the concatenated predicate and its information about the grammatical relations as well as the arguments of the individual predicates.

```
(18) pred(root,yAd_kar)
     subj(yAd_kar,nAdiyah)
     obj(yAd_kar,kahAnI)
     complex-pred-type(yAd_kar,nv)
     cp-part1(yAd_kar,yAd)
     cp-part2(yAd_kar,kar)
     arg1(kar,nAdiyah)
     arg2(kar,yAd)
     arg1(yAd,kahAnI)
     asp(yAd_kar,perf).
```

The ADJ+V CPs receive a parallel dependency analysis to the other CPs. The resultative part of the ADJ+V is contained within the complex PRED (e.g., `alag_kar`='separate do'). The object to which the resultative property is applied is encoded as the overall object. This yields the set of triples in (19) for our example (9).

```
(19) pred(root,alag_kar)
     subj(alag_kar,yAsIn)
     obj(alag_kar,SISa)
     obl(alag_kar,kHiRkI)
     cp-part1(alag_kar,alag)
     cp-part2(alag_kar,kar)
     arg1(kar,yAsIn)
     arg2(kar,SISa)
     arg1(,alag,kHiRkI)
     complex-pred-type(alag_kar,av)
     asp(alag_kar,perf).
```

Our dependency bank also contains examples of complex predicates with more than two levels of embedding, such as the example in Figure 5 where the CP consists of three parts. The verb *ban* 'build' is causativized with an *-A* causative and then combined with the permissive light verb *dE* 'give', each of the parts contributing their own arguments.[8]

[8] The *-A* causative is represented with an underscore in the set of triples as it is not an independent lexical item but a morphological suffix of the verb (e.g. `cp-part2(ban_A_dE,_A)` in (20).

```
(20) pred(root,ban_A_dE)
     subj(ban_A_dE,yAsIn)
     obj(ban_A_dE,gHar)
     obj-go(ban_A_dE,laRkA)
     cp-part1(ban_A_dE,ban)
     cp-part2(ban_A_dE,_A)
     cp-part3(ban_A_dE,dE)
     arg1(dE,yAsIn)
     arg2(dE,_A)
     arg1(_A,laRkA)
     arg2(_A,ban)
     arg1(ban,gHar)
     asp(ban_A_dE,perf).
```

Further CP constructions and their representation in the triples format, in particular the agreeing N-V CPs, are presented and discussed in the on-line resource `http://ling.uni-konstanz.de/pages/home/pargram_urdu/main/Resources.html`.

5. Conclusion

This paper presented a seed bank for a complete reference dependency bank for CPs crosslinguistically. The reference dependency bank to date contains examples of all common Urdu/Hindi CP types we are aware of and it furthermore includes instances of other constructions which are not CPs, but which are often mistakenly analyzed on a par with CPs. The differences between these constructions is also clearly shown in the set of dependency triples which are theory-independent in nature and can serve as a gold standard in further applications for the resolution of complex predicates. Furthermore, the CP dependency bank is meant as a reference resource that NLP researchers can consult when working on a new language and can use for tasks such as figuring out an appropriate POS tagset for the language, constructing analyses for treebanking, chunking or extracting lexical resources.

6. References

Tafseer Ahmed and Miriam Butt. 2011. Discovering Semantic Classes for Urdu N-V Complex Predicates. In *Proceedings of the International Conference on Computational Semantics (IWCS 2011)*.

Tafseer Ahmed. 2011. Some issues related to Complex Predicates in Urdu/Hindii. In *Workshop on Complex Predicates in South Asian Languages*, ICON 2011, Chennai.

Rajesh Bhatt, Tina Bögel, Miriam Butt, Annette Hautli, and Sebastian Sulger. 2011. Urdu/Hindi modals. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of LFG11*, pages 47–67. Stanford: CSLI Publications.

Tina Bögel, Miriam Butt, Annette Hautli, and Sebastian Sulger. 2009. Urdu and the Modular Architecture of ParGram. In *Proceedings of the Conference on Language and Technology 2009 (CLT09)*.

Joan Bresnan. 1982. Control and complementation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 282–390. MIT Press, Cambridge, MA.

Miriam Butt and Tracy Holloway King. 2006. Restriction for Morphological Valency Alternations: The Urdu Causative. In *Intelligent Linguistic Architecturs: Variations on Themes by Ronald M. Kaplan*. Stanford: CSLI Publications.

Miriam Butt and Tracy Holloway King. 2007. Urdu in a Parallel Grammar Development Environment. *Language Resources and Evaluation*, 41(2):191–207.

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

Miriam Butt, Tracy Holloway King, and John T. Maxwell III. 2003. Productive Encoding of Urdu Complex Predicates in the ParGram Project. In *Proceedings of the EACL03: Workshop on Computational Linguistics for South Asian Languages: Expanding Synergies with Europe*.

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

Miriam Butt. 2010. The Light Verb Jungle: Still Hacking Away. In Mengistu Amberber, Brett Baker, and Mark Harvey, editors, *Complex Predicates in Cross-Linguistic Perspective*. Cambridge University Press.

Dick Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula Newman, 2012. *XLE Documentation*. Palo Alto Research Center.

Mary Dalrymple. 2001. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press.

Tracy Holloway King, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ron Kaplan. 2003. The PARC700 Dependency Bank. In *Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*.

Muhammad Kamran Malik, Tafseer Ahmed, Sebastian Sulger, Tina Bögel, Atif Gulzar, Ghulam Raza, Sarmad Hussain, and Miriam Butt. 2010. Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*.

Colin P. Masica. 1993. *The Indo-Aryan Languages*. Cambridge University Press.

Tara Mohanan. 1994. *Argument Structure in Hindi*. CSLI Publications.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1).

Victoria Rosén, Paul Meurer, and Koenraad de Smedt. 2009. LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. In Frank Van Eynde, Anette Frank, Gertjan van Noord, and Koenraad De Smedt, editors, *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133. LOT.