Citation for this document


Citation for the Multi-CAST collection

Haig, Geoffrey & Schnell, Stefan (eds.). 2015. Multi-CAST: Multilingual corpus of annotated spoken texts. (multicast.aspra.uni-bamberg.de/) (date accessed)
Contents

1 Notes on the GRAID annotations  .............................................................. 1
  1.1 Subordinate and relative clauses  ...................................................... 1
  1.2 Complex predicates  ...................................................................... 2
  1.3 Non-canonical subjects .................................................................. 4
  1.4 Complex noun phrases .................................................................. 4
    1.4.1 NP-internal classifiers, quantifiers, and demonstratives .............. 4
    1.4.2 Partitive modifiers within the NP  ................................................. 5
    1.4.3 Lexical modifiers within the NP .................................................. 6

References ................................................................................................. 6

Appendices ............................................................................................... 7
  A List of corpus-specific GRAID symbols ............................................. 7
  B List of abbreviated morphological glosses ....................................... 8
1 Notes on the GRAID annotations

This document contains notes on the implementation of the GRAID annotation conventions (Haig & Schnell 2014) in the Multi-CAST Persian corpus. It corresponds to version 1905 of the annotations, published in May 2019. Unless a more recent version of this document exists, it also applies to any later versions of the annotations.

1.1 Subordinate and relative clauses

In Persian, most subordinate clauses (including relative clauses) are introduced by the all-purpose complementizer *ke*, and involve finite verb forms governing a set of arguments basically identical to those of independent clauses. We therefore count them as clause units containing a normal (:pred). The example in (1) contains such a complement clause:

(1) a. *mard miāyad pāyin mibinad*
   man come.PRS.IND.3SG down 0 see.PRS.IND.3SG
   # np.h:s v:pred rv # 0.h:a v:pred
   ‘The man comes down [from the tree] and finds out...’

   b. *ke teki az sabdhā xāli ast*
   that one of basket.PL empty be.PRS.3SG
   #cc ke indef_other:s rn_adp rn_np:poss other:pred cop
   ‘...that one of the baskets is empty.’

When glossing relative clauses it is important to note that the head noun is usually systematically gapped in the relative clause (i.e. it cannot be overtly expressed). In such cases, we do not gloss a zero in the relative clauses, because speakers have no choice between zero and overt argument expression (following the rationale of glossing zeroes in Bickel 2003), with the result that in a large number of relative clauses there is no representative of a core argument in the GRAID annotation.

Example (2) illustrates subject relativization, where overt expression of the subject NP is systematically banned from the relative clause, while (3) illustrates systematic gapping of the object in object relativization:

(2) a. *mixorad be yek doxtari*
   hit.PRS.IND.3SG to one girl.INDEF
   # 0.h:s.cp v:pred adp ln_deti np.h:g
   ‘(He) runs into a girl...’

   b. *ke dāste az ān taraf barmigašte*
   that AUX.PST.3SG from that side return.PST.PTCP
   #rc ke aux adp ln_dem np:l v:pred
   ‘...who was coming back from the opposite direction.’

(3) a. *in mivehā =rā*
   this fruit.PL =ACC
   ln_dem np:p =rn_acc
   ‘these fruits...’
Relative clauses are frequently centre-embedded, in which case standard GRAID procedure is followed, indicating the right edge of the embedded clause with the symbol (%) (unless it coincides with the right edge of its matrix clause):

\[(4)\]

b. \(\text{ke} \text{ jam} \text{ mikonand}\)
\(\text{that} \text{ collected} \text{ do.PRS.IND.3PL}\)
\='#rc \text{ ke} \text{ 0.h:a other:lvc v:pred}\)
‘...that they gather’

In a small number of cases centre-embedded structures would have required complex (and controversial) syntactic annotation. In order to avoid undue complications, we treated the relevant strings as \(<\text{nc}>\), but annotated the matrix clause – to the extent that it is a syntactically well-formed clause – in the normal way:

\[(5)\]

b. \(\text{ke} \text{ dārand} \text{ miravand}\)
\(\text{that} \text{ AUX.PRS.3PL go.PRS.IND.3PL}\)
\='#rc \text{ ke} \text{ aux} \text{ v:pred} \%
‘...that are just leaving ...’

1.2 Complex predicates

Complex predicates (CPs) in Persian are conventionalized combinations of a non-verbal element with a light verb, which together create the predicate of a clause. Both the non-verbal element and the light verb contribute to the resulting semantics, but CPs are often not semantically compositional. Complex predicates raise a number of problems in connection with GRAID. The first issue is to decide on the transitivity value of the entire expression, as this determines whether
we gloss the subject with \(\langle s\rangle\) or \(\langle a\rangle\). We identify a verb as transitive if it has the ability to govern a direct object marked with the clitic \(\text{rā}\). We distinguish four possible coding scenarios, each glossed as follows:

a. **intransitive light verb + CP is intransitive**
   \(\text{rad šodan} \, \text{‘pass by’}\) ➔ subject glossed \(\langle s\rangle\)

b. **transitive light verb + CP is intransitive**
   \(\text{farār kardan} \, \text{‘escape’ (lit. ‘escaping do’)}\) ➔ subject glossed \(\langle s\_cp\rangle\)
   \(\text{zamin xorōn} \, \text{‘fall down’ (lit. ‘earth eat’)}\) ➔ subject glossed \(\langle s\_cp\rangle\)

c. **intransitive light verb + CP is transitive**
   \(\text{balad budan} \, \text{‘know’}\) ➔ subject annotated \(\langle a\_cp\rangle\)

d. **transitive light verb + CP is transitive**
   \(\text{yād gereftan} \, \text{‘learn’ (lit. ‘memory take’)}\) ➔ subject annotated \(\langle a\rangle\)

For the purposes of cross-corpus comparison, the additional underscores may be ignored and the \(\langle a\_cp\rangle\) and \(\langle s\_cp\rangle\) symbols included in the \(\langle a\rangle\) and \(\langle s\rangle\) categories respectively.

The second issue in annotating CPs is the status of the non-verbal element. CPs are typically highly conventionalized, and the non-verbal element is generally not referential, hence could be simply included into the predicate gloss. However, there are also borderline cases, and the class of CPs in Persian cannot be readily distinguished from other expressions involving indefinite or generic objects. We have generally applied a neutral \(\langle lvc\rangle\) ‘light verb complement’ gloss for these elements, which mainly serves the purpose of identifying complex predicates in the annotation in case researchers are particularly interested in their properties.

The example in (6) illustrates the annotation procedure. A special kind of CP involving non-canonical subjects is dealt with in the next section.

(6)

a. \(\text{bad čand tā az peserhā}\)
   \# other ln_qu class_np.h:a rn_adp rn_np.h:poss
   ‘Then some boys …’

b. \(\text{ke az hamān jā dāštand rad mišodand}\)
   \#rc ke adp ln_lex np:l aux other:lvc v:pred %
   ‘… who were passing by …’

c. \(\text{āmadand}\)
   \# other ln_qu class_np.h:a rn_adp rn_np.h:poss
   ‘… came …’

d. \(\text{komak =aš kardand}\)
   \# other:lvc =pro.h:p v:pred
   ‘… and helped him …’

e. \(\text{golābihā =rā jam kardand}\)
   \# 0 h:a np:p =rn_acc other:lvc v:pred
   ‘… gather up the pears, …’
1.3 Non-canonical subjects

In Persian, subjects can be uncontroversially defined in terms of (i) their ability to control agreement suffixes on the verbal predicate, and (ii) their lack of overt case marking. These morphological features also correlate with syntactic features such as the ability to control reflexives, or coreferential deletion. However, a set of predicates in Persian has NPs that show most of the typical properties of subjects, but lack the ability to control agreement suffixes on the verb. We refer to them as non-canonical subjects (NCS). Semantically, NCSs are generally experiencers, or some kind of external possessor or benefactive. Typically they occur with complex predicates (CPs), and the non-verbal element of the CP obligatorily carries a possessive clitic reflecting person and number of the NCS. Functionally, this is evidently a kind of "agreement", though the exponent of agreement is not a verbal suffix, but a possessive clitic. In this kind of construction, we gloss the possessive clitic in the same manner as other possessive clitics, and the NCS is glossed with the function gloss ⟨ncs⟩. If the NCS is not present in the clause, it receives a zero gloss in GRAID.

(7)  

(8)  

(9)  

1.4 Complex noun phrases

1.4.1 NP-internal classifiers, quantifiers, and demonstratives

The speakers make very frequent use of NPs of the type ‘three pieces (of) X’, involving a quantifier (often a numeral, but also indefinite expressions such as ‘one’, ‘some’, etc.), a classifier (e.g. tā ‘piece’), and a noun, in some cases linked to the entire expression with the preposition az ‘from’. These expressions lead to certain issues in analysis, particular in deciding on the head. Structurally, the classifier expression is the head, while semantically, the complement of the preposition az is the head.
When classifiers and quantifiers are combined in the NP, we gloss them \( \langle \text{ln\_class} \rangle \) and \( \langle \text{ln\_qu} \rangle \) respectively, while treating the lexical noun as the head, and adding the function gloss to it, as in (10) and (11):

\[
\begin{align*}
\text{(10)} & \quad \text{se} \quad \text{tā} \quad \text{pesarbaçeye} \quad \text{dīgar} \quad \text{nazdiktar} \quad \text{istāde} \quad \text{budand} \\
\# & \quad \text{ln\_qu} \quad \text{ln\_class} \quad \text{np\:h:s} \quad \text{rn\_lex} \quad \text{other} \quad \text{stand\:PST\:PTCP} \quad \text{AUX\:PST\:3PL} \quad \text{v\:pred} \quad \text{aux} \\
& \quad \text{‘Three boys were standing nearby.’} \\
\text{(11)} & \quad \text{hameye} \quad \text{golāihā} \quad \text{mirizad} \\
\# & \quad \text{ln\_qu} \quad \text{np\:s} \quad \text{v\:pred} \\
& \quad \text{‘All the pears spill out.’}
\end{align*}
\]

Analogously, we gloss NP-internal demonstratives with \( \langle \text{ln\_dem} \rangle \), as demonstrated in (12):

\[
\begin{align*}
\text{(12)} & \quad \text{bad} \quad \text{in} \quad \text{āqā} \quad \text{dobāre} \quad \text{miravād} \quad \text{bālāye} \quad \text{deraxt} \\
\# & \quad \text{other} \quad \text{ln\_dem} \quad \text{np\:h:s} \quad \text{other} \quad \text{v\:pred} \quad \text{adp} \quad \text{np\:l} \\
& \quad \text{‘Then he climbs up the tree again.’}
\end{align*}
\]

In the absence of a lexical head, the classifier or quantifier is treated as the head and receives the appropriate function gloss, as in (13):

\[
\begin{align*}
\text{(13)} & \quad \text{in} \quad \text{se} \quad \text{tā} \quad \text{dāştand} \quad \text{mirafand} \\
\# & \quad \text{ln\_dem} \quad \text{ln\_qu} \quad \text{class\:np\:h:s} \quad \text{aux} \quad \text{v\:pred} \\
& \quad \text{‘These three were leaving.’}
\end{align*}
\]

The same procedure is adopted for indefinite pronouns, where we use the gloss \( \langle \text{indef\_other} \rangle \):

\[
\begin{align*}
\text{(14)} & \quad \text{bad} \quad \text{yeki} \quad =\$ \quad =\text{rā} \quad \text{barmidārad} \\
\# & \quad \text{other} \quad \text{0\:h:a} \quad \text{indef\_other\:p} \quad =\text{pro\:poss} \quad =\text{rn\:acc} \quad \text{v\:pred} \\
& \quad \text{‘Then he picks up one of them.’}
\end{align*}
\]

\[
\begin{align*}
\text{(15)} & \quad \text{yeki} \quad \text{az} \quad \text{ān} \quad \text{baçehā} \quad \text{bā} \quad \text{sut} \quad \text{pesar} \quad =\text{rā} \\
\# & \quad \text{indef\_other\:h:a} \quad \text{rn\:adp} \quad \text{rn\_dem} \quad \text{rn\_np\:h:poss} \quad \text{adp} \quad \text{np\:obl} \quad \text{np\:h:p} \quad =\text{rn\:acc} \\
& \quad \text{sedā} \quad \text{mikonad} \quad \text{calling} \quad \text{do\:PRS\:IND\:3SG} \\
\quad & \quad \text{other\:lvc} \quad \text{v\:pred} \\
& \quad \text{‘One of the kids calls the boy by whistling.’}
\end{align*}
\]

1.4.2 Partitive modifiers within the NP

In several cases we find a lexically light expression (classifier, indefinite pronoun, quantifier, etc.) modified by a prepositional phrase, yielding expressions like ‘three of the boys’, and so on. In these cases we have treated classifiers or quantifiers as the head (and hence carrier of the function
gloss) in examples such as the following, taken from above. Where partitive expressions within the NP occur, they are considered (:poss):

(16) čand tā az peserhā
    a.few piece of boy.pl
    ln_qu class_np:h:a rn_adp rn_np:h:poss
    'a few of the boys’

(17) ke yeki az Ţän zanbilhā =rā gozāšt ruye
    that one of that basket.pl =acc put.pst.3sg on
    #rc other in_pro:p rn_adp rn_dem rn_np:poss =rn_acc v:pred adp
    doçarxe =aš
    bike =poss.3sg
    np:1 =pro.h:poss
    'He puts one of the baskets on his bike.’

(18) yek dāne az sabad=e golābihāyi ke
    one piece of basket=ezafe pear.pl.indf that 0
    ln_qu class_np:p rn_adp rn_np:poss rn_lex #rc ke 0:h:a
    çide bud
    pick.pst.ptcp aux.pst.3sg
    v:pred aux
    'one basket of pears that (he) had picked’

1.4.3 Lexical modifiers within the NP

Where lexical modifiers (adjectives or nouns) are included in the NP (generally linked via the ezafe particle), we have glossed them with ⟨rn_lex⟩:

(19) yeki az sabadhāy=e golābi =aš nist
    one of basket.pl=ezafe pear =poss.3sg neg.be.prs.3sg
    # indef_other.h:s rn_adp rn_np:poss rn_lex =pro.h:poss cop
    'One of the baskets is not here’

References


Appendices

A List of corpus-specific GRAID symbols

The following is a list of the non-standard GRAID symbols used in the annotation of the Multi-CAST Persian corpus. Please refer to the GRAID manual (Haig & Schnell 2014: 54–55) for an inventory of basic GRAID symbols.

Form symbols and specifiers

- ⟨class_np⟩ classificatory particle
- ⟨qu_np⟩ quantifier phrase
- ⟨indef_other⟩ indefinite pronoun

Function symbols and specifiers

- ⟨:lvc⟩ light verb complement
- ⟨:s_cp⟩ subject of an intransitive complex predicate with transitive light verb
- ⟨:a_cp⟩ subject of a transitive complex predicate with intransitive light verb

Subconstituent symbols

- ⟨_acc⟩ object postpositional particle የ; attaches to ⟨=rn⟩
- ⟨_adp⟩ adposition; attaches to ⟨rn⟩
- ⟨_class⟩ classificatory particle; attaches to ⟨ln⟩
- ⟨_dem⟩ demonstrative as determiner; attaches to ⟨ln⟩ and ⟨rn⟩
- ⟨_deti⟩ indefinite article as determiner; attaches to ⟨ln⟩
- ⟨_lex⟩ lexical modifier, usually an adjective or some other item of uncertain word class; attaches to ⟨ln⟩ and ⟨rn⟩
- ⟨_qu⟩ quantifier, subconstituent of a NP; attaches to ⟨ln⟩ and ⟨rn⟩

Other symbols

- ⟨ke⟩ complementizer ke
## B List of abbreviated morphological glosses

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>first person</td>
<td>2</td>
<td>second person</td>
</tr>
<tr>
<td>3</td>
<td>third person</td>
<td>4</td>
<td>negation</td>
</tr>
<tr>
<td>ACC</td>
<td>accusative</td>
<td>PFV</td>
<td>perfective</td>
</tr>
<tr>
<td>AUX</td>
<td>auxiliary</td>
<td>PL</td>
<td>plural</td>
</tr>
<tr>
<td>DEF</td>
<td>definite</td>
<td>POSS</td>
<td>possessive</td>
</tr>
<tr>
<td>EZAPE</td>
<td>ezafe</td>
<td>PRO</td>
<td>pronoun</td>
</tr>
<tr>
<td>IMP</td>
<td>imperative</td>
<td>PROSP</td>
<td>prospective</td>
</tr>
<tr>
<td>IND</td>
<td>indicative</td>
<td>PRS</td>
<td>present</td>
</tr>
<tr>
<td>INDF</td>
<td>indefinite</td>
<td>PST</td>
<td>past</td>
</tr>
<tr>
<td>IPFV</td>
<td>imperfective</td>
<td>PTCP</td>
<td>participle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SBJV</td>
<td>subjunctive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SG</td>
<td>singular</td>
</tr>
</tbody>
</table>
Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/