Multi-CAST

Persian

annotation notes

Shirin Adibifar

May 2019
v1.1

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

Australian Government
Australian Research Council

University of Bamberg

DFG

Multi-CAST *Multilingual Corpus of Annotated Spoken Texts*

Citation for this document

Adibifar, Shirin. 2019. Multi-CAST Persian annotation notes. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts.* (`multicast.aspra.uni-bamberg.de/`) (date accessed)

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)

The Multi-CAST collection has been archived at the *University of Bamberg*, Germany, and is freely accessible online at `multicast.aspra.uni-bamberg.de/`.

Multi-CAST Persian annotation notes v1.1 last updated 8 May 2019
This document was typeset by NNS with XeLaTeX and the *multicast3* class (v3.2.1).

Contents

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*	*sabadhā*	*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.' [mc_persian_g2-f-01_0018]

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*
	0	hit.PRS.IND.3SG	to	one	girl.INDF
#	0.h:s_cp	v:pred	adp	ln_deti	np.h:g

'(He) runs into a girl …'

b.

	ke	*dāšte*	*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.'

[mc_persian_g1-f-08_0011]

(3) a.

in	*mivehā*	*=rā*
this	fruit.PL	=ACC
ln_dem	np:p	=rn_acc

'these fruits …'

b.
	ke	*jam*	*mikonand*
	that 0	collected	do.PRS.IND.3PL
#rc	ke 0.h:a	other:lvc	v:pred

'… that they gather' [mc_persian_g1-f-05_0007]

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) a.
	ān	*se*	*tā*	*tačeyi*
	that	three	piece	kid.INDF
#	ln_dem	ln_qu	ln_class	np.h:a

'Those three boys …'

b.
	ke	*dārand*	*miravand*	
	that	AUX.PRS.3PL	go.PRS.IND.3PL	
#rc	ke	aux	v:pred	%

'… that are just leaving …'

c.
kolāh	*=rā*	*peydā*	*mikonand*
hat	=ACC	found	do.PRS.IND.3PL
np:p	=rn_acc	other:lvc	v:pred

'… find the hat.' [mc_persian_g2-f-04_0009]

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 ⟨nc⟩, but annotated the matrix clause – to the extent that it is a syntactically well-formed clause – in the normal way:

(5) a.
	kolāh	*=aš*	*=rā*
0	hat	=POSS.3SG	=ACC
# 0.h:a	np:p	=pro.h:poss	=rn_acc

'His hat, …'

b.
	ke	*didand*	*ruye*	*zamin*	*ast*	
	that	see.PST.IND.3PL	on	earth	is	
#nc	nc	nc	nc	nc	nc	%

'… which they saw lying on the ground, …'

c.
be	*=heš*	*bargardāndand*
to	=3SG	return.PST.IND.3PL
adp	=pro.h:g	v:pred

'… (they) returned to him.' [mc_persian_g2-m-11_0006]

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 ⟨:s⟩ or ⟨:a⟩. We identify a verb as transitive if it has the ability to govern a direct object marked with the clitic *rā*. We distinguish four possible coding scenarios, each glossed as follows:

a. intransitive light verb + CP is intransitive
   rad šodan 'pass by' → subject glossed⟨:s⟩

b. transitive light verb + CP is intransitive
   farār kardan 'escape' (lit. 'escaping do') → subject glossed ⟨:s_cp⟩
   zamin xordan 'fall down' (lit. 'earth eat') → subject glossed ⟨:s_cp⟩

c. intransitive light verb + CP is transitive
   balad budan 'know' → subject annotated ⟨:a_cp⟩

d. transitive light verb + CP is transitive
   yād gereftan 'learn' (lit. 'memory take') → subject annotated ⟨:a⟩

For the purposes of cross-corpus comparison, the additional underscores may be ignored and the ⟨:a_cp⟩ and ⟨:s_cp⟩ symbols included in the ⟨:a⟩ and ⟨:s⟩ 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 ⟨:lvc⟩ '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.

	bad	*čand*	*tā*	*az*	*peserhā*
	then	a.few	piece	of	boy.PL
#	other	ln_qu	class_np.h:a	rn_adp	rn_np.h:poss

'Then some boys …'

b.

	ke	*az*	*hamān*	*jā*	*dāštand*	*rad*	*mišodand*	
	that	from	same	place	AUX.PST.3PL	crossing	become.PST.IND.3PL	
#rc	ke	adp	ln_lex	np:l	aux	other:lvc	v:pred	%

'… who were passing by …'

c.

āmadand
come.PST.3PL
v:pred

'… came …'

d.

	komak	*=aš*	*kardand*
	help	=PRO.3SG	do.PST.3PL
#	other:lvc	=pro.h:p	v:pred

'… and helped him …'

e.

		golābihā	*=rā*	*jam*	*kardand*
	0	pear.PL	=ACC	collecting	do.PST.PL
#	0.h:a	np:p	=rn_acc	other:lvc	v:pred

'… gather up the pears, …'

f.

	dāxele	*zanbil*	*rixtand*	
0	inside	basket	pour.PST.3PL	0
# 0.h:a	adp	np:g	v:pred	0:p

'… and put them back in the basket.' [mc_persian_g2-f-07_0011]

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)

	çeşm	*=aş*	*in*	*sabad-hā*	*=rā*	*gereft*
0	eye	=POSS.3SG	this	basket.PL	=ACC	catch.PST.3SG
# 0.h:ncs	other:lvc	=pro.h:poss	ln_dem	np:p	=rn_acc	v:pred

'(He) caught sight of these baskets.' (lit. 'his eye took the baskets')
[mc_persian_g1-f-05_0005]

(8)

va		*çaşm*	*=aş*	*mioftad*	*be*	*golābihā*
and	0	eye	=POSS.3SG	fall.PRS.IND.3SG	to	pear.PL
# other	0.h:ncs	other:lvc	=pro.h:poss	v:pred	adp	np:obl

'(He) caught sight of the pears.' (lit. 'his eyes fall on the pears')
[mc_persian_g1-m-13_0012]

(9)

bad	*in*	*ham*	*havās*	*=aş*	*part*	*mīšavad*
then	3SG	ADD	attention	=POSS.3SG	separated	become.PRS.IND.3SG
# other	pro.h:ncs	other	other:lvc	=pro.h:poss	other:lvc	v:pred

'[His hat fell off] and then he got distracted.' (lit. 'he his attention became separated')
[mc_persian_g1-f-14_0013]

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 ⟨ln_class⟩ and ⟨ln_qu⟩ respectively, while treating the lexical noun as the head, and adding the function gloss to it, as in (10) and (11):

(10)

	se	*tā*	*pesarbaçeye*	*digar*	*nazdiktar*	*istāde*	*budand*
	three	piece	little.boy	other	closer	stand.PST.PTCP	AUX.PST.3PL
#	ln_qu	ln_class	np.h:s	rn_lex	other	v:pred	aux

'Three boys were standing nearby.' [mc_persian_g1-f-02_0015]

(11)

	hameye	*golāihā*	*mirizad*
	all	pear.PL	pour.PRS.IND.3SG
#	ln_qu	np:s	v:pred

'All the pears spill out.' [mc_persian_g1-f-01_0010]

Analogously, we gloss NP-internal demonstratives with ⟨ln_dem⟩, as demonstrated in (12):

(12)

	bad	*in*	*āqā*	*dobāre*	*miravad*	*bālāye*	*deraxt*
	then	this	man	again	go.PRS.IND.3SG	top.of	tree
#	other	ln_dem	np.h:s	other	v:pred	adp	np:l

'Then he climbs up the tree again.' [mc_persian_g1-f-01_0006]

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):

(13)

	in	*se*	*tā*	*dāştand*	*miraftand*
	this	three	piece	AUX.PST.PL	IPFV.go.PST.3PL
#	ln_dem	ln_qu	class_np.h:s	aux	v:pred

'These three were leaving.' [mc_persian_g1-m-04_0009]

The same procedure is adopted for indefinite pronouns, where we use the gloss ⟨indef_other⟩:

(14)

	bad	*yeki*	*=ş*	*=rā*	*barmidārad*
	then	one	=POSS.3SG	=ACC	pick.up.PRS.IND.3SG
#	other	indef_other:p	=pro:poss	=rn_acc	v:pred

'Then he picks up one of them.' [mc_persian_g1-f-01_0004]

(15)

	yeki	*az*	*ān*	*baçehā*	*bā*	*sut*	*pesar*	*=rā*
	one	of	that	kids	with	whistle	boy	=ACC
#	indef_other.h:a	rn_adp	rn_dem	rn_np.h:poss	adp	np:obl	np.h:p	=rn_acc

sedā	*mikonad*
calling	do.PRS.IND.3SG
other:lvc	v:pred

'One of the kids calls the boy by whistling.' [mc_persian_g1-m-13_0025]

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' [mc_persian_g2-f-07_0011]

(17)
	ke	*yeki*	*az*	*ān*	*zanbilhā*	*=rā*	*gozāşt*	*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	*=aş*
bike	=POSS.3SG
np:l	=pro.h:poss

'He puts one of the baskets on his bike.' [mc_persian_g2-f-07_0007]

(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' [mc_persian_g1-f-02_0008]

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*	*=aş*	*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' [mc_persian_g1-f-01_0018]

References

Adibifar, Shirin. 2016. Multi-CAST Persian. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (http://multicast.aspra.uni-bamberg.de/#persian) (Accessed 2019-03-08).

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4). 708–736.

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0)*. (http://multicast.aspra.uni-bamberg.de/#annotations) (Accessed 2019-03-08).

Haig, Geoffrey & Schnell, Stefan (eds.). 2016. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (http://multicast.aspra.uni-bamberg.de/) (Accessed 2019-03-08).

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 *rā*; 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

1	first person
2	second person
3	third person
ACC	accusative
AUX	auxiliary
DEF	definite
EZAFE	ezafe
IMP	imperative
IND	indicative
INDF	indefinite
IPFV	imperfective
NEG	negation
PFV	perfective
PL	plural
POSS	possessive
PRO	pronoun
PROSP	prospective
PRS	present
PST	past
PTCP	participle
SBJV	subjunctive
SG	singular

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/