هم بام

HAM BAM

The
Hamedan-Bamberg
Corpus of Contemporary
Spoken Persian

Corpus description

Geoffrey Haig

Mohammad Rasekh-Mahand

July 2022
v1.0

University of Bamberg

دانشگاه بوعلی سینا

Alexander von Humboldt
Stiftung/Foundation

هم بام

HAM BAM

The
Hamedan-Bamberg
Corpus of Contemporary
Spoken Persian

Citation for this document

Haig, Geoffrey & Rasekh-Mahand, Mohammad. 2022. HamBam corpus description. In Haig, Geoffrey & Rasekh-Mahand, Mohammad (eds.), *HamBam: Hamedan-Bamberg Corpus of Contemporary Spoken Persian*. (`https://multicast.aspra.uni-bamberg.de/resources/hambam/`) (date accessed)

Citation for the HamBam corpus

Haig, Geoffrey & Rasekh-Mahand, Mohammad (eds.). 2022. *HamBam: Hamedan-Bamberg Corpus of Contemporary Spoken Persian*. (`https://multicast.aspra.uni-bamberg.de/resources/hambam/`) (date accessed)

The HamBam corpus has been archived at the *University of Bamberg*, Germany, and is freely accessible online at `https://multicast.aspra.uni-bamberg.de/resources/hambam/`.

HamBam Corpus description v1.0 last updated 24 July 2022
This document was typeset by NNS with XeLaTeX and the *multicast3* class (v3.2.6).

Contents

1 About HamBam

1.1 Corpus objectives

HamBam, the *Hamedan-Bamberg Corpus of Contemporary Spoken Persian* (Haig & Rasekh-Mahand 2022), is an ongoing project aimed at creating an unrestrictedly accessible online corpus of contemporary spoken Persian. The design of the corpus follows the architecture and rationale of Multi-CAST (Haig & Schnell 2022), but with certain modifications. As in Multi-CAST, the texts are annotated using the free annotation software ELAN (ELAN developers 2022), which links sound files to annotation files, and the annotated data are available in various formats (sound files, ELAN annotation files, CSV files, XML). The data can also be listened to online via the preview function, or downloaded for processing in user-specific applications.

1.2 Corpus compilation

The speech samples were recorded, transcribed, and annotated in Hamedan by a team of researchers under the supervision of Mohammad Rasekh-Mahand, following training sessions with Geoffrey Haig in Hamedan in 2019 and later in Bamberg. A system of morphological glossing was developed that is geared to the specific requirements of Persian, and a simplified system of syntactic annotation was implemented, GRAID-L (L for ‘light’), based on the GRAID system (Haig & Schnell 2014). The systems have evolved over several years before being finalized in its current state. See Section 2 below for explanations. A subset of the HamBam texts has been used for a data set in the WOWA corpus (Izadi 2022).

1.3 Language characteristics

Most of the recordings come from residents of Hamedan, Iran, and environs. For some speakers there is evidence of the local Hamedani variety of Persian in their speech, which includes some striking features such as the object suffix *-e*, but most speakers use a fairly neutral variety of spoken contemporary Persian, though with a Hamedani accent. All texts would be fully comprehensible to any educated Persian speaker from any region of Iran, so we believe that the corpus provides a reasonable representation of contemporary spoken Persian, despite the obvious limitations in size.

Texts are distinguished approximately according to two degrees of formality, reflected in the file naming conventions. The majority of the texts have the prefix ⟨oh⟩ ‘oral history’, which indicates biographical anecdotes told in a relaxed domestic setting among family and friends. The remainder are ⟨ac⟩ ‘academic discourse’ (e.g. radio interviews, podcasts, etc.), which represent a more formal register. The addition of ⟨d⟩ to the first prefix (e.g. ⟨ohd⟩, ⟨acd⟩) indicates that the text is largely dialogue, as opposed to the default type of monologue.

1.4 Transcription and annotation

Recordings were initially segmented into utterance units, based on a combination of prosodic and syntactic cues. Utterance units are consecutively numbered, with the numbering appearing in the `utterance_id` tier in ELAN. Subsequently, we segment the entire text into clause units, comprising of a predicate and all items dependent on, or syntactically associated with, the predicate. These two steps replicate the structure of Multi-CAST corpora generally. In view of the diglossic nature of Persian, we have provided two transcriptions of the utterance: The

	tier name	*description*
A	utterance_id	sequential numbering of utterance units
B	utterance	broad phonemic transcription of the actual utterance
C	utterance_formal	more abstract, conservative rendering of the text with morphemes in a unified form (i.e. disregarding certain phonological processes that change the shape of morphemes)
D	grammatical_words	tokenized word units identified in the utterance_formal tier
E	gloss	contains a standardized morpheme-for-morpheme glosses, see Appendix A for the full list of tags used in this tier
F	graid-l	annotations with GRAID-L, a simplified version of the GRAID annotation system (Haig & Schnell 2014). Essentially the same set of tags and combinatorial syntax is used as in GRAID, but with two major differences: (i) referential zeroes ⟨0⟩ are not added, that is only those constituents that are overtly present in the utterance are coded; (ii) the system of clause boundary markers and tags has been simplified, mainly due to pervasive problems in differentiating between complement clauses and various kinds of adjunct clause in Persian; see Section 2.2 below for explanations.
G	utterance_translation	idiomatic English translation
H	comments	comments on the annotations or the text

Table 1 List of tiers in the HamBam annotation files.

first, in the utterance tier, contains a broadly phonemic transcription of the actual utterance as spoken. The second, utterance_formal, contains a more abstract representation that corresponds more closely to the written language, and the way Persian is spoken in very formal contexts. This second tier provides the foundation for the subsequent lower levels of annotation. The full set of tiers with explanations are provided in Table 1; (1) provides a fully annotated example showing all tiers.

(1) **A** ac_f_social_0006
B *xâteretun bâše sâle gozašte bud, ke...*
C *xâteretân bâšad sâle gozašte bud, ke...*

D	#	*xâter=etân*	*bâš-ad*	#	*sâl-e*	*gozašte*	*bud*	#	*ke*
E	#	mind=2PL	SBJ.be.PRS-3SG	#	year-EZ	past	be.PST.3SG	#	COMPL
F	#	other:lvc=pro.2:ncs	v:pred	#	np:pred	rn_other	cop	#	compl

H ‘You may remember, it was last year that...’

2 Notes on the annotations with GRAID-L

2.1 Clitic pronouns

Because clitic pronouns are usually referential expressions (but see below), they are placed in their own annotation cell, and if necessary have ⟨rn⟩ on them.

(2)

kelas	*=ešân*
class	=3PL
np:l	=rn_pro.hposs

‘in their class’

(3)

be	*=heš*
to	=3SG
adp	=pro.h:g

‘to him’

2.1.1 Exception: Clitic pronouns as agreement markers

An exception to the above rule is when clitic pronouns occur as a kind of agreement marker in non-canonical subject constructions ⟨:ncs⟩. In these cases, they are not considered referential and do not have their own annotation cell (or referent index):

(4)

qašang	*yâd=am*	*ast*
beautiful	remember=1SG	be.PRS.3SG
other	np:lvc=pro.1:ncs	v:pred

‘I remember very well.’

2.1.2 Exception: Reflexive pronoun xod- with possessive clitics

This is an obligatory combination (i.e. *xod-* always requires a pronominal suffix), and is treated as a single item without splitting the pronoun and base, as follows:

(5)

xod=aš	*raft*
self=3SG	go.PST.3SG
refl_pro.h:s	v:pred

‘He went.’

(6)

az	*hess=e*	*xod=am*
from	sense=EZ	self=1SG
adp	np:obl	rn_refl_pro.1:poss

‘from my senses’

2.2 Clause boundaries

In practice, it is extremely difficult to maintain a distinction between complement clauses and other kinds of subordinate or adverbial clauses, or even post-verbal relative clauses. Spoken Persian makes extensive use of the complementizer *ke*, which loosely connects a variety of more or less “embedded” clauses, but these items can also be added without any overt complementizer. In view of the multiple difficulties of interpretation, we have opted to simplify the system otherwise used in GRAID, as follows.

- ⟨#⟩ is the general marker for the beginning of a main clause, that is a clause that is considered independent or which cannot be unambiguously assigned to one of the other types below. In other words, this is the default clause boundary marker, and is by far the most frequent.
- ⟨#dc⟩ is a dependent clause. This covers any kind of clause considered to be syntactically dependent on another clause, which means it subsumes a number of items that are differentiated in standard GRAID (e.g. ⟨#cc⟩ for complement clauses or ⟨#ac⟩ for adverbial clauses, etc.). A clause could be annotated ⟨#dc⟩ because it fills an argument position of the other clause, or because it is semantically dependent (e.g. a purpose or cause, a temporal clause, a conditional clause, so it might begin with ‘if’, ‘when’, etc.). It may also be marked through a complementizer or the use of the subjunctive mood on the verb. ⟨#dc⟩ is also used for post-posed relative clauses (separated from their head noun, and following the verb). Where there was doubt, we have instead used the default clause marker ⟨#⟩.
- ⟨#rc⟩ is a relative clause that immediately follows its head noun. Relative clauses that are separated from their head noun are instead annotated ⟨#dc⟩.
- ⟨#ds⟩ marks clear cases of direct speech. This tag may combine with those listed above (e.g. ⟨#ds_dc⟩, ⟨#ds_rc⟩), though this occurs only rarely.
- The right-edge clause boundary marker ⟨%⟩ is only used to indicate the end of a centre-embedded relative clause ⟨#rc⟩ as in (7). There are very few examples of this in the texts.

(7)

	in	*bačče-yi*		*ke*	*âmad*		*xeyli*	*šeytân*	*bud*
	this	child-RM		COMPL	come.PST.3SG		very	demon	be.PST.3SG
#	ln_dem	np.h:s	#rc	other	v:pred	%	other	other:pred	cop

‘This child that came was very naughty.’

References

ELAN developers. 2022. *ELAN (Version 6.4)*. Nijmegen: Max Planck Institute for Psycholinguistics. (`https://archive.mpi.nl/tla/elan`).

Haig, Geoffrey & Rasekh-Mahand, Mohammad (eds.). 2022. *HamBam: Hamedan-Bamberg Corpus of Contemporary Spoken Persian*. (`https://multicast.aspra.uni-bamberg.de/resources/hambam/`) (Accessed 2022-07-24).

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

Haig, Geoffrey & Schnell, Stefan (eds.). 2022. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (`https://multicast.aspra.uni-bamberg.de/`) (Accessed 2022-07-09).

Haig, Geoffrey & Stilo, Donald & Doğan, Mahîr C. & Schiborr, Nils N. (eds.). 2022. *WOWA — Word Order in Western Asia): A spoken-language-based corpus for investigating areal effects in word order variation*. Bamberg: University of Bamberg. (`multicast.aspra.uni-bamberg.de/resources/wowa/`).

Izadi, Elham. 2022. Persian (New). In Haig, Geoffrey & Stilo, Donald & Doğan, Mahîr C. & Schiborr, Nils N. (eds.), *WOWA — Word Order in Western Asia*: *A spoken-language-based corpus for investigating areal effects in word order variation*, Updated 22 July 2022. Bamberg: University of Bamberg. (`multicast.aspra.uni-bamberg.de/resources/wowa/`).

Appendices

A List of abbreviated morphological glosses

1	first person
2	second person
3	third person
3SGX	non-standard third person singular verbal agreement *=aš*, found sporadically with the third person singular of past tense verbs (e.g. *raft=aš* ‘he went’), and with the copula *hast=aš* ‘it is’
CLF	classifier
COMPL	complementizer
COMPR	comparative
DEF	definite
EXCL	exclamation
EZ	ezafe
FOC	focus particle (e.g. *hâ* in *na ân ast hâ*)
FUT	future tense
IMP	imperative
IND	indicative
INDF	indefinite
INF	infinitive
NEG	negation/negated
OM	object marker (*=râ*)
PL	plural
PROG	progressive
PRS	present tense
PRV	preverbal particle
PST	past tense
PTCP	participle
Q	question particle (only one case, the *magar* in *magar mehrdâd zende ast?*)
REDUP	reduplication
RM	relative marker
SBJ	subjunctive
SG	singular
NC	not classifiable

B List of text and speaker metadata

file name	*date rec'd*	*spkr. gender*	*length mm:ss*	*genre*	*rec'd by*	*annotated by*	*also in WOWA*
ac_f_social	18-06-03	female	09:53	public speech, academic	Iran Abdi	Iran Abdi, Mitra Hossein-gholian	✘
ac_m_corona1	20-06-12	male	04:05	hygiene recommendations	(unknown, internet)	Fariba Sabouri, Maryam Pooyankhah	✓
ac_m_corona2	20-06-12	male	05:39	hygiene recommendations	(unknown, internet)	Elham Izadi, Iran Abdi	✓
ac_m_depression	20-05-06	male	03:34	storytelling	(podcast)	Fariba Sabouri, Maryam Pooyankhah	✓
acd_f_science	18-06-10	female	09:57	radio interview, academic	(Radio Iran)	Fariba Sabouri, Marayam Pooyankhah	✘
acd_m_education	18-06-05	male	05:01	radio interview, academic	(Radio Iran)	Rahele Izadifar, Mehdi Parizade	✘
acd_m_plane	20-07-15	male	03:56	scientific dialogue	(podcast)	Mehdi Parizadeh, Mehrdad Meshkinfam	✓
acd_m_ufo	20-07-15	male	03:25	scientific dialogue	(podcast)	Elham Izadi, Iran Abdi	✓
oh_f_accident	18-05-01	female	06:40	oral history, informal	Mehrdad Meshkinfam	Elham Izadi, Mehrdad Meshkinfam	✓
oh_f_amirali	20-07-15	female	01:01	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
oh_f_aunt	18-03-02	female	04:00	oral history, informal	Mehrdad Meshkinfam	Elham Izadi, Mehrdad Meshkinfam	✓
oh_f_childhood1	20-07-09	female	02:05	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
oh_f_childhood2	18-03-01	female	02:55	oral history, informal	Mehrdad Meshkinfam	Elham Izadi, Mehrdad Meshkinfam	✓

continue ⟶

	date	spkr.	length				also in
file name	*rec'd*	*gender*	*mm:ss*	*genre*	*rec'd by*	*annotated by*	*WOWA*
oh_f_class	19-12-03	female	02:42	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
oh_f_daryush	19-05-28	female	02:17	oral history, informal	Iran Abdi	Mehdi Parizadeh, Iran Abdi	✓
oh_f_istanbul1	20-07-25	female	02:46	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
oh_f_istanbul2	20-07-09	female	02:46	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✘
oh_f_marry	18-03-02	female	04:25	oral history, informal	Mehrdad Meshkinfam	Fariba Sabouri, Maryam Pooyankhah	✓
oh_f_mask	20-07-27	female	01:12	oral history, informal	Iran Abdi	Elham Izadi, Iran Abdi	✓
oh_f_parham	20-07-28	female	02:55	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
oh_f_pool	20-07-27	female	01:18	oral history, informal	Iran Abdi	Elham Izadi, Iran Abdi	✓
oh_f_taxi1	20-07-25	female	03:34	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
oh_f_uncle2	18-03-01	female	03:01	oral history, informal	Mehrdad Meshkinfam	Elham Izadi, Mehrdad Meshkinfam	✓
oh_f_uncle1	19-12-03	female	02:50	oral history, informal	Fariba Sabouri	Fariba Sabouri, Maryam Pooyankhah	✓
oh_f_university1	20-08-20	female	02:46	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
oh_m_corona3	20-06-14	male	00:52	hygiene recommendations	Fariba Sabouri	Fariba Sabouri, Maryam Pooyankhah	✓
oh_m_military2	20-07-28	male	01:42	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
oh_m_military1	20-07-10	male	06:13	oral history, informal	Mehrdad Meshkinfam	Mehrdad Meshkinfam, Mehdi Parizadeh	✓

continue ⟶

file name	*date rec'd*	*spkr. gender*	*length mm:ss*	*genre*	*rec'd by*	*annotated by*	*also in WOWA*
oh_m_music	19-07-08	male	05:27	oral history, informal	Mehdi Parizadeh	Mehdi Parizadeh, Iran Abdi	✓
oh_m_taxi2	20-07-10	male	02:51	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
oh_m_television	20-07-09	male	02:26	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
oh_m_university2	20-06-14	male	03:13	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
oh_m_university3	20-07-10	male	03:25	oral history, informal	Mehrdad Meshkinfam	Mehrdad Meshkinfam, Mehdi Prizadeh	✓
oh_m_usa	20-07-10	male	01:53	oral history, informal	Iran Abdi	Elham Izadi, Iran Abdi	✓

Table B.1 List of texts in HamBam and associated metadata as of July 2022.

هم بام

HAM BAM

The
Hamedan-Bamberg
Corpus of Contemporary
Spoken Persian

`https://multicast.aspra.uni-bamberg.de/resources/hambam/`