

H A M B A M

The  
Hamedan-Bamberg  
Corpus of Contemporary  
Spoken Persian

## *Corpus description*

*Geoffrey Haig*

*Mohammad Rasekh-Mahand*

*September 2022*  
v2.0



University of Bamberg



  
**Alexander von Humboldt**  
Stiftung/Foundation



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The HamBam corpus has been archived at the *University of Bamberg*, Germany,  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/>.

The entirety of HamBam, including this document, is published under the *Creative  
Commons Attribution 4.0 International Licence* (CC BY 4.0), unless noted otherwise. The  
licence can be reviewed online at [creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/).

*HamBam Corpus description* v2.0 last updated 1 September 2022  
This document was typeset by NNS with X<sub>Y</sub>L<sup>A</sup>T<sub>E</sub>X and the *multicast3* class (v3.2.6).

## Contents

<b>1</b>	<b>About HamBam</b> _____	<b>1</b>
1.1	Corpus objectives	1
1.2	Corpus compilation	1
1.3	Language characteristics	1
1.4	Transcription and annotation	1
<b>2</b>	<b>Notes on the annotations with GRAID-L</b> _____	<b>3</b>
2.1	Clitic pronouns	3
2.1.1	Exception: Clitic pronouns as agreement markers	3
2.1.2	Exception: Reflexive pronoun <i>xod-</i> with possessive clitics	3
2.2	Clause boundaries	4
	<b>References</b>	<b>5</b>
	<b>Appendices</b>	<b>6</b>
A	List of abbreviated morphological glosses	6
B	List of text and speaker metadata	7



# 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

<i>tier name</i>	<i>description</i>
<b>A</b> utterance_id	sequential numbering of utterance units
<b>B</b> utterance	broad phonemic transcription of the actual utterance
<b>C</b> 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)
<b>D</b> grammatical_words	tokenized word units identified in the utterance_formal tier
<b>E</b> gloss	contains a standardized morpheme-for-morpheme glosses, see Appendix A for the full list of tags used in this tier
<b>F</b> graid-l	<p>annotations with GRAID-L, a simplified version of the GRAID annotation system (Haig &amp; Schnell 2014). Essentially the same set of tags and combinatorial syntax is used as in GRAID, but with two major differences:</p> <p>(i) referential zeroes (&lt;math&gt;\emptyset&lt;/math&gt;) are not added, that is only those constituents that are overtly present in the utterance are coded;</p> <p>(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.</p>
<b>G</b> utterance_translation	idiomatic English translation
<b>H</b> 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:1 =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/](https://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/](https://multicast.aspra.uni-bamberg.de/resources/wowa/)).

## Appendices

### A List of abbreviated morphological glosses

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

## B List of text and speaker metadata

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

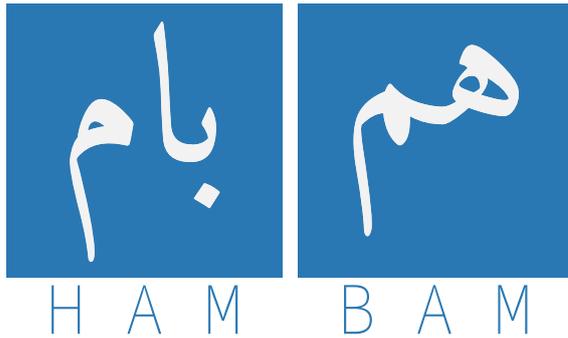
*continue* →

<i>file name</i>	<i>date rec'd</i>	<i>spkr. gender</i>	<i>length mm:ss</i>	<i>genre</i>	<i>rec'd by</i>	<i>annotated by</i>	<i>also in WOWA</i>
<i>oh_f_class</i>	19-12-03	female	02:42	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
<i>oh_f_daryush</i>	19-05-28	female	02:17	oral history, informal	Iran Abdi	Mehdi Parizadeh, Iran Abdi	✓
<i>oh_f_istanbul1</i>	20-07-25	female	02:46	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
<i>oh_f_istanbul2</i>	20-07-09	female	02:46	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✗
<i>oh_f_marry1</i>	18-03-02	female	04:25	oral history, informal	Mehrdad Meshkinfam	Fariba Sabouri, Maryam Pooyankhah	✓
<i>oh_f_marry2</i>	22-05-08	female	02:02	oral history, informal	Fariba Sabouri	<i>HamBam</i> team	✗
<i>oh_f_mask</i>	20-07-27	female	01:12	oral history, informal	Iran Abdi	Elham Izadi, Iran Abdi	✓
<i>oh_f_nosejob</i>	20-09-22	female	06:40	oral history, informal	Fariba Sabouri	<i>HamBam</i> team	✗
<i>oh_f_parham</i>	20-07-28	female	02:55	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
<i>oh_f_pool</i>	20-07-27	female	01:18	oral history, informal	Iran Abdi	Elham Izadi, Iran Abdi	✓
<i>oh_f_taxi1</i>	20-07-25	female	03:34	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
<i>oh_f_uncle2</i>	18-03-01	female	03:01	oral history, informal	Mehrdad Meshkinfam	Elham Izadi, Mehrdad Meshkinfam	✓
<i>oh_f_uncle1</i>	19-12-03	female	02:50	oral history, informal	Fariba Sabouri	Fariba Sabouri, Maryam Pooyankhah	✓
<i>oh_f_university1</i>	20-08-20	female	02:46	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
<i>oh_m_childhood3</i>	21-03-03	male	01:55	oral history, informal	Maryam Pooyan	<i>HamBam</i> team	✗
<i>oh_m_corona3</i>	20-06-14	male	00:52	hygiene recommendations	Fariba Sabouri	Fariba Sabouri, Maryam Pooyankhah	✓

*continue* →

<i>file name</i>	<i>date rec'd</i>	<i>spkr. gender</i>	<i>length mm:ss</i>	<i>genre</i>	<i>rec'd by</i>	<i>annotated by</i>	<i>also in WOWA</i>
<i>oh_m_football</i>	21-03-31	male	02:32	oral history, informal	Maryam Pooyan	HamBam team	✗
<i>oh_m_military2</i>	20-07-28	male	01:42	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
<i>oh_m_military1</i>	20-07-10	male	06:13	oral history, informal	Mehrdad Meshkinfam	Mehrdad Meshkinfam, Mehdi Parizadeh	✓
<i>oh_m_music</i>	19-07-08	male	05:27	oral history, informal	Mehdi Parizadeh	Mehdi Parizadeh, Iran Abdi	✓
<i>oh_m_taxi2</i>	20-07-10	male	02:51	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
<i>oh_m_television</i>	20-07-09	male	02:26	oral history, informal	Elham Izadi	Elham Izadi, Iran Abdi	✓
<i>oh_m_university2</i>	20-06-14	male	03:13	oral history, informal	Maryam Pooyan	Fariba Sabouri, Maryam Pooyankhah	✓
<i>oh_m_university3</i>	20-07-10	male	03:25	oral history, informal	Mehrdad Meshkinfam	Mehrdad Meshkinfam, Mehdi Prizadeh	✓
<i>oh_m_usa</i>	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 September 2022.



The  
**Hamedan-Bamberg**  
Corpus of Contemporary  
Spoken Persian

---

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