E - ISSN: 2953-8203
P - ISSN: 2953-819X

YEREVAN STATE
UNIVERSITY

JOURNAL OF

IRANIAN LINGUISTICS

Volume 2 - Issue 2 - 2025

JOURNAL OF IRANIAN LINGUISTICS

Volume 2 | issue 2

YEREVAN STATE UNIVERSITY PUBLISHING HOUSE

E - ISSN: 2953-8203
P - ISSN: 2953-819X

JOURNAL OF IRANIAN LINGUISTICS
VOLUME 2 | ISSUE 2

CONTENT

Foreword

Volume 2 / Issue 2

"A language is not just words. It is a culture, a tradition, a unification of a community, a whole history that creates what a community is. It is all embodied in a language." — Noam Chomsky

The five articles gathered in this issue of the Journal of Iranian Linguistics offer a compelling testament to the breadth and vitality of the field. Taken together, they traverse vast stretches of the Iranian linguistic world — from the steppes of Central Asia to the mountains of south-western Iran, from the coasts of Oman to the digital infrastructure of natural language processing — and engage with data ranging from medieval manuscript poetry to contemporary dialectal fieldwork. What unites them is a shared commitment to close, empirically grounded inquiry into languages and language phenomena that remain insufficiently understood.

The issue opens with **Shuan Osman Karim**'s investigation of a striking typological puzzle: the apparent similarity between the Balochi imperfective aspect marker a- and its functional near-equivalent in Central Kurdish. Karim brings new dialectal data from Kurdish to bear on the question first raised by Paul (2003), systematically evaluating the prospect of a shared etymology before ultimately rejecting it. In its place, he proposes a derivation from the so-called KAR construction, attested in Caspian languages with which Balochi is known to have been in contact. The article makes a significant contribution to the diachronic study of Western Iranian verbal morphology, demonstrating how parallel phonological processes and shared grammaticalization pathways can produce near-identical outcomes across historically distinct systems.

Muhammed Ourang and **Khalsa Al-Aghbari** turn to a different dimension of morphological creativity in their comparative study of reduplication in Lāri and Jibbāli. Despite belonging to entirely separate language families — Southwestern Iranian and Modern South Arabian Semitic, respectively — both languages make productive use of full reduplication to intensify adjectives and nouns. Beyond this shared pattern, the two languages diverge revealingly: Lāri partial reduplication serves a wide semantic spectrum encompassing emphasis, intensity, categorization, attenuation, and addition, while Jibbāli deploys reduplication to mark transitivity and convey aspectual continuity in Arabic loanwords. Drawing on original fieldwork data, the article constitutes the first systematic treatment of reduplication in Jibbāli and a significant contribution to the morphological documentation of Lāri.

Mortaza Taheri-Ardali, **Mansour Bozorgmehr**, and **Erik Anonby** contribute a methodologically innovative study in linguistic geography, challenging the widespread assumption that Kohgiluyeh va Boyer Ahmad Province is uniformly Lori-speaking. Through field survey and the interactive mapping tools of the Atlas of the Languages of Iran, the authors reveal a considerably more complex picture. In addition to seven varieties of Southern Lori, Ghashghāi Turkic, Khuzestāni Arabic, and Bakhtiari are spoken in the province's southern districts, while Persian has emerged as the first language of nearly a quarter of the population — with serious implications for the intergenerational transmission of the region's heritage languages.

Vahide Tajalli, **Mehrnoush Shamsfard**, **Yalda Yarandi**, **Mahtab Sarlak**, and **Arezoo Haghbin** bring computational linguistics to bear on Persian verbal multiword expressions (VMWEs). Adapting the international PARSEME annotation framework to the specific grammatical properties of Persian, the authors present guidelines tested on a corpus of 5,617 sentences. The study bridges universal annotation standards and the structural realities of Persian, with practical implications for syntactic parsing, machine translation, and semantic role labeling.

The issue closes with **Majid Tame**'s study of two proverb-like expressions in the Book of Zambasta, the celebrated Khotanese poetic composition from the Khotan region of present-day Xinjiang. Tame traces two sentences addressing futile effort and exertion, identifying their equivalents in several New Western Iranian languages where cognate proverbs remain in living use — a reminder of the deep continuities connecting the medieval Iranian literary heritage to the spoken traditions of the present day.

I would like to extend my sincere thanks to the editorial board for their continued guidance, and to the reviewers for their careful assessments.

We hope that this volume will prove stimulating to all scholars of Iranian linguistics and will encourage further research into the many dimensions of this richly varied field.

Vardan Voskanian
Editor-in-Chief
Journal of Iranian Linguistics

YEREVAN STATE UNIVERSITY PUBLISHING HOUSE

JIL 2/2 (2025): 6-41

Journal of Iranian Linguistics

Contact across the Iranian World: curious convergences between Kurdish and Balochi

Shuan Osman Karim

Julius-Maximilians-Universität Würzburg

doi.org/10.46991/jil/2025.02.01

Abstract: It has not gone unnoticed that the Balochi imperfective aspect marker *a-* is superficially similar in form to the Central Kurdish marker *a-* (*e-* in Kurdish orthography) and nearly identical in function. Micro variation shows further similarities. For instance, the way these formatives idiosyncratically attach to certain verbs: come, bring, etc. have striking similarities in both groups. Additionally, the forms of past imperfective and past conditional forms show further convergence with the Kurdish system, differing from the original constructions preserved in Gorani and Zazaki. In light of new data from Kurdish, I evaluate the prospect of a relationship between these formations and others, arguing that these constructions ultimately have different etymologies. However, their convergence is far from coincidence.

Keywords: Diachrony; Language Contact; Sound Change; Morphological Analogy.

Shuan Osman Karim
E-mail: shuan.karim@uni-wuerzburg.de
ORCID: https://orcid.org/0000-0002-9727-1637

Received: 07.11.2025
Revised: 22.12.2025
Accepted: 30.12.2025

Conflict of Interest
The authors declare no conflicts of interest.

Funding
This research did not receive any financial support

1. Introduction:

Balochi is a fascinating case in language contact. There are forms in Balochi that indicate contact with languages from all parts of the Iranian periphery, including Kurdish and Gilaki, Persian, and Indo-Aryan/Dravidian languages, at all historical stages. These data tell a story of nomadic migrations circling Iran from the northwest, across and down the southeast to Hindustan, back across the south of Iran, and then repeating the journey at least twice over the course of centuries. As such, it is not strange that Balochi has converged with languages following Karim's (2021) "Rubik's Cube' metaphor: moves on

one axis correspond to geographical migrations, while moves on other axes correspond to phonological and morphological changes shared by languages both genetically related and otherwise.

Examining the verbal systems of Balochi and Kurdish Zone languages, a set of superficial similarities shows convergences that set them apart from New Persian (and other varieties). A selection of these convergences is presented here in *Table 1*. Using Balochi as the base of comparison, there are a few immediately apparent patterns: The imperfective/indicative is marked by the prefix *a-*.[1] The same is true in most Northern, Central, and Southern Kurdish varieties, where the prefix is a variant of *d-, de-, e-, t-*, in Persian *mī-*, and in Most Gorani varieties *me-, m-*. Only a few varieties, e.g., Zazaki, feature a prefixless present imperfect indicative from the old active participle (< *-ant-). Some Gorani and Southern Kurdish varieties have lost these prefixes for purely phonological reasons, and Central Kurdish varieties have undergone phonological changes resulting from the loss of the initial dental consonant. However, in all varieties, even those with vastly different productive forms of the prefix, the original is preserved before h-initial verbs, e.g., 'come', 'bring', etc. Note that, although the etymon is not yet certain, the Balochi form follows the same pattern with a special imperfective formative *k-* occurring before h-initial verbs. However, in most varieties, the *k-* cooccurs with the productive *a-*. The only other parallel to this in *Table 1* is the Southern Kurdish variety of Qorwe, where younger speakers no longer associate the *t-* with the imperfective aspect and have begun to regularize the form by adding the *e-* prefix. Note that Kurdish and Balochi varieties do not use the irregular imperfective markers in the context of negation.

The other notable convergence between Kurdish and Balochi is the form of the past conditional and past imperfective, the protasis and apodosis of irreal conditionals, respectively. The past imperfective is just the simple past tense with the addition of the productive imperfective prefix, and the past conditional is the subjunctive prefix *bi-* and a suffix Kurdish *-a*, Balochi *-ên*. Note that similar strategies to these conditional forms are also found in other varieties. In Zazaki, the imperfective suffix is *-êni* and the conditional is the same with the prefix *bi-*. Likewise, Early New Persian (but not modern New Persian) had an imperfective suffix *-i*, and the conditional was the same form, optionally with the *bi-* prefix. The Gorani suffix cognate with that in Balochi,

[1] In Balochi, the imperfect markers have hitherto unknown etyma. These forms surface as the prefix *a-* in some varieties and the proposed enclitic *=a* attaching to the nearest preverbal matter in others (Nourzaei and Jahani 2012, for the geographic distribution of the particles, see). Additionally, there is a second imperfective prefix *k-* that attaches to specific verbs. Paul (2003) has suggested that there may be some relationship between the Kurdish imperfective prefix *a-* (*e-* in the Kurdish orthography) and the Balochi *a-*. Note that this assertion is not a strong claim. He merely recognizes that they are similar formatives with similar functions.

Zazaki, and Early New Persian *-ê(n)* attached to the present-tense or imperfective stem forms the past-imperfective, and, attached to the past-tense/perfective stem, forms the past conditional.

Table 1. Comparison of selected forms of the third-person singular in Western Iranic languages

	do.PRS.IND	come.PRS.AFF	come.PRS.NEG	do.PST.IPFV	do.PST.COND
NK Dihok	***ti***-*ke*	***ti***-*hê*	*na-ê*	***ti***-*kir*	***bi***-*kir*-***a***
CK Mukrî	***de***-*ka*	***d***-*ê*	*na-ê*	***de***-*kird*	***bi***-*kird*-***a***
CK Slêmanî	***e***-*ka*	***y***-*ê*	*na-ê*	***e***-*kird*	***bi***-*kird*-***a***
CK Kerkûk	***e***-*ka*	***d***-*ê*	*na-ê*	***e***-*kird*	***bi***-*kird*-***a***
SK Qorwe	***e***-*ka*	*(e-)****t***-*ê*	*na-ê*	***e***-*kird*	***bi***-*kird*-***a***
SK Kirmanşa	*ket*	***t***-*ê*	*ny-ê*	*kird-ya*	***bi***-*kird*-***ya***
Balochi	***a***-*kan-t*	*a-****k***-*ay-t*	*na-y-ay-t*	***a***-*kurt*	***b***-*kurt*-***ēn***
N Zazaki	*ker-en-o*	*y-en-o*	*nê-y-en-o*	*kerd-êni*	***bi***-*kerd*-***êni***
G Text	*ker-o*	***m***-*ê*	*ni-m-ê*	*ker-ê*	*kerd*-***ɛ***
G Pawe	***me***-*ker-o*	***m***-*ê*	*ni-m-ê*	*ker-ê*	*kerd*-***aê***
ENP	***mī***-*kun-ad*	*mī-āy-ad*	*ne-mī-āy-ad*	*kard-i*	*bi-kard-i*
NP Tehran	***mī***-*kun-e*	*mī-ā-d*	*ne-mī-ā-d*	***mī***-*kard*	*mī-kard*

The current study examines the convergence between these varieties, looking at what is known about these formatives from a historical comparative perspective. Ultimately, I demonstrate that although these formatives likely have different etyma, their superficial similarity may not be a mere coincidence. I examine the possibility of the relationship between Kurdish *e-* and Balochi *a-* (first proposed by Paul 2003) based on a proposal for the origin of the Kurdish imperfective markers in Southern Kurdish in Karim (2024). Ultimately, I reject the possibility that these are etymologically related. However, there are many similarities that point to another possible etymon, the so-called "KAR construction" (Vafaeian 2018).

These developments are relevant to the study of diachronic linguistics and Iranian languages because of what they suggest about convergence. Traditionally, language contact and 'Neo-Grammarian" sound change and morphological analogy are seen as entirely separate mechanisms of language change (following Campbell 2013). The physiology of the vocal tract leads to regular, exceptionless, phonetic developments, while psychological processes smooth out irregularities caused by those developments. The Balochi form show how languages across a large geographic area can undergo the same physiological and psychological processes, causing disparate formatives to converge in their general form and function. This can happen without direct contact, as the mechanisms that led to these mutual developments likely

began long before the relevant formatives were originally recruited into the respective verbal systems.

The proposal presented in this article is primarily based on patterns pervasive among Iranic languages that are too similar to ignore. Essentially, the same phonological changes affected broad swaths of the Western Iranian World. Different languages and varieties recruited different inherited forms for the emergent progressive construction, which underwent the progressive to imperfective cycle at varying times in their history. However, similar changes occurred to these formations due to regular sound shifts. I discuss the inherited Western Iranic forms in *Section 2* based on data from Fattah (2000), Axenov (2006), Matras et al. (2016), Belelli (2021), Mohammadirad (2020), Mahmoudveysi and Bailey (2013), Stilo (2007a), and Wal Anonby (2015). In *Section 3*, I outline the convergence of different patterns of the present tense in Western Iranic. I use the analysis of Kurdish in Karim (2024) as a template to evaluate whether the Balochi imperfective marker is likely related to the Kurdish one, and I propose a new etymon for the Balochi imperfective marker, based on the analysis that works for Kurdish but fails for Balochi (see *Section 3.5.2*). Likewise, I discuss convergence and developments of the past conditional and past imperfective following the Southern Kurdish model. Note that, thanks to Fattah (2000), we have sufficient knowledge of Southern Kurdish to observe all stages in the development from the inherited system extant in Zazaki today and in the oldest Early New Persian texts to the system in all of Central Kurdish and Balochi.

2. Previous Research

A notable pattern is evident among the verbal systems of Western Iranic languages. Each of the languages appears to have adopted the same pattern, regardless of the material recruited to perform each task. The pattern is as follows in *Table 2*. The basic template can be observed in Early New Persian: there is a single negation marker *na-* which has an allomorph *ne-* before the imperfective marker *mī*. The imperfective marker combines with the present-tense stem to form the basic indicative. It contrasts with the subjunctive, marked by the prefix *be-*, originally a marker of perfective aspect, which still serves this function in other Iranic languages, such as Pashto. The subjunctive prefix does not occur if there is any preverbal matter, i.e., a preverb, a light-verb complement, or the negation marker *na-*.

Table 2. Basic Western Iranic Verbal Paradigm with Early New Persian examples 'go'

Function	**Composition**	**ENP Example**
Non-past Indicative Affirmative	IPFV-STEM.PRS-APM	*mī-rav-ad*
Non-past Indicative Negative	NEG-IPFV-STEM.PRS-APM	*ne-mī-rav-ad*
Non-past Subjunctive Affirmative	PFV-STEM.PRS-APM	*be-rav-ad*
Non-past Subjunctive Negative	NEG-STEM.PRS-APM	*na-rav-ad*
Past Perfective Affirmative	STEM.PST-APM	*raft*
Past Perfective Negative	NEG-STEM.PST-APM	*na-raft*
Past Imperfective Affirmative	STEM.PST-IPFV-APM	*raft=i*
Past Imperfective Negative	NEG-STEM.PST-IPFV-APM	*na-raft=i*
Past Conditional Affirmative	PRF-STEM.PST-IPFV-APM	*(be-)raft=i*
Past Conditional Negative	NEG-STEM.PST-IPFV-APM	*na-raft=i*
Present Perfect Affirmative	PCPL-COP.PRS	*raft-e=ast*
Present Perfect Negative	NEG-PCPL-COP.PRS	*na-raft-e=ast*
Past Perfect Affirmative	PCPL-COP.PST	*raft-e=bud*
Past Perfect Negative	NEG-PCPL-COP.PST	*na-raft-e=bud*

The past-tense forms are built from the old past participle, now the past-tense stem. The bare stem is perfective, and it can become imperfective by the addition of the enclitic *=i* (see Lambton 1960). This form is used for recurring actions in the past, and in the apodosis of irreal conditionals. This form is also used in the protasis of irreal conditionals increasingly with the addition of the subjunctive/perfective prefix *be-*. The perfect tenses are formed by the new past participle ending in *-e* (< *-ag) and the enclitic copula.

Among the Western Iranic languages, the group that differs the most from this pattern is Gorani. In *Table 3*, the forms of Paweyane are presented. The forms of the present tense have the same structure observed in *Table 2*. Here, I refer to the non-past stem as imperfective, because all imperfective forms utilize this stem regardless of tense. Likewise, the past stem can be understood as both perfective and past. In this sense, the perfective past and the two perfect tenses are also constructed in the same way as the forms in *Table 2*. The difference lies in the past imperfective, which employs the imperfective stem with endings recognized as Middle Iranic optatives (see Skjærvø 2009; Aliyari Baboghani 2025). Just as in Middle Persian and Parthian, the optative of the copula could also combine with the perfective past stem to form the past conditional. Note that these forms are still fully inflected, i.e., not a frozen form like the *=i* of Early New Persian.

Table 3. Gorani Pawe Verbal Pardigm 'live'

Function	**Composition**	**Paweyane**
Non-past Indicative Affirmative	IPFV-IPFV(STEM)-APM	*me-jîɯ-o*
Non-past Indicative Negative	NEG-IPFV-IPFV(STEM)-APM	*ni-me-jîɯ-o*
Non-past Subjunctive Affirmative	PFV-IPFV(STEM)-APM	*bi-jîɯ-o*
Non-past Subjunctive Negative	NEG-IPFV(STEM)-APM	*ne-jîɯ-o*
Past Perfective Affirmative	PFV(STEM)-APM	*jîɯa*
Past Perfective Negative	NEG-PFV(STEM)-APM	*ne-jîɯa*
Past Imperfective Affirmative	IPFV(STEM)-OPT-APM	*jîɯ-ê*
Past Imperfective Negative	NEG-IPFV(STEM)-OPT-APM	*ne-jîɯ-ê*
Past Conditional Affirmative	IPFV(STEM)-COP.OPT-APM	*jîɯa-ê*
Past Conditional Negative	NEG-IPFV(STEM)-COP.OPT-APM	*ne-jîɯa-ê*
Present Perfect Affirmative	PCPL-COP.IPFV	*jîɯa(e)=n*
Present Perfect Negative	NEG-PCPL-COP.IPFV	*ne-jîɯa(e)=n*
Past Perfect Affirmative	PCPL-COP.PFV	*jîɯa(e)=bî*
Past Perfect Negative	NEG-PCPL-COP.PFV	*ne-jîɯa(e)=bî*

The Gorani system has undergone several developments; it can be understood as preserving the extant Middle Iranic situation in terms of imperfective and irrealis marking. The Middle Iranic optative endings were used for past-habitual actions. A periphrastic form consisting of the perfective stem (past-participle) and the optative of the copula was used for the protasis of irreal conditionals. Note that these were among the functions of the optative, dating back to the Old Iranic period.

Table 4. Optative/Imperfect: G Pawe, MP, Parthian, Zazaki

	Paweyane		**Middle Persian**		**Parthian**		**Zazaki**	
	SG	PL	SG	PL	SG	PL	SG	PL
1	*-ênê*	*-ênmê*	*-ên*	?	?	?	*-êni*	*-êni*
2	*-êşî*	*-êndê*	*-êš*	?	*-êndê*	?	*-êni*	*-êni*
3	*-ê*	*-ênê*	*-ê*	*-endê*	*-êndê*	*-endê*	*-êni*	*-êni*

Note that the forms of the imperfective in Gorani closely mirror what was observed for the imperfective/optative in Middle Persian. There are gaps in the attested paradigm for the first- and second-person plural forms, but all other forms match those of Gorani Pawe, with some phonological and analogical developments. The imperfective in Zazaki is formed with a single invariable exponent that does not inflect for person, number or gender. Despite gaps in the attested paradigm, this seems to be the system in the Middle Iranic Parthian language.

3. Mutual Developments: Imperfective prefixes

For some of the languages represented here, there is a complete historical analysis of the formatives involved in imperatives and conditionals. From that perspective, there is a clear roadmap for understanding the convergence in Balochi, which lacks the rich dialectal description that Kurdish has. Data from over 100 varieties (sourced from Matras et al. 2016; Fattah 2000) have preserved almost every stage in the development of these forms. I explore these convergences here.

Beginning with the present-tense forms. All of these varieties have begun with some variant of the common Western Iranic model described in *Section 2*: (NEG-)IPFV-STEM.PRS.APM. As the present tense is the exclusive domain of imperfectivity, the imperfective marker, after losing its original progressive meaning, is essentially bleached of all meaning. The only other present-tense forms are subjunctive, marked with the old perfective prefix *bi-*. As such, the only meaningful contrast was not aspectual but rather modal: indicative vs subjunctive. The various languages had imperfective markers from the locative, e.g., Kurdish *de-* and adverbials, e.g., Persian *mī-*, Gorani *me-*, etc.

In some of these languages, subsequent phonological changes led to the loss of the prefix in Certain contexts. These contexts are most clear in Gorani, where the changes only took place in the Hewramî core, and the stages have been preserved in the periphery (see Karim and Mohammadirad forthcoming). Some of these changes were shared by nearby Southern Kurdish and North Eastern Neo-Aramaic varieties (NENA). Additionally, the imperfective prefixes in these varieties share a common phonological shape CV-: NENA *ka-*, Kurdish *de-*, and Gorani *me-*.

3.1. Pretonic reduction

In Hewramî (Gorani) and Southern Kurdish, there was a pretonic reduction of the vowel *e, reducing to zero in open syllables or *i*, the epenthetic vowel, in closed syllables (Karim and Mohammadirad forthcoming; Karim 2024). Note the proto forms in (1). For each variety, I show an h-initial and another non-h C-initial root: Kurdish *ke-* 'make/do' and *hê-* 'come,' NENA *doq-* 'hold' and *hol* 'make/do,' and Gorani *ker-* 'make/do' and *(h)ar-* 'bring.'[2]

(1)	Kurdish		NENA		Gorani	
	*de-ke-m	*de-hê-m	*ka-doq-na	*ka-hol-na	*me-ker-û	me-(h)ar-û

Southern Kurdish, NENA Sanandaj and Hewramî underwent pretonic reduction (2). In a word-initial syllable that is unstressed, the vowel *e

[2] Note that the *h in the Gorani form is etymological ultimately going back to a PIE second laryngeal H_2.

reduces to *Ø* unless a triple consonant cluster is created, then the epenthetic vowel *i* appears.

(2)	Southern Kurdish		NENA		Gorani	
	*d-ké-m	*d-hế-m	*k-dóq-na	*k-hól-na	*m-ker-ű́	m-(h)ar-ű́

3.2. Cluster reduction

Then, the initial consonant clusters were reduced. *Ch became C (devoiced), and other C_1C_2 reduced to C_2. These changes can be understood in terms of regular sound changes and relative chronology

1. voicing assimilation: a preceding consonant takes the voicing value of the following consonant.

2. non-initial *h loss: word-medial *h is lost.

3. cluster reduction: word-initial consonant clusters are reduced to single consonants.

The result was no imperfective marker on the vast majority of verbs, but remnants of the old marker on h-initial verbs; see (3).

(3)		SK Kermanşa		NENA Sanandaj		Gorani Hewraman Text	
		*dké-m	*dhế-m	*gdóq-na	*khól-na	*mker-ű́	*mar-ű́
	1	*tké-m	*thế-m	*gdóq-na	*khól-na	*mker-ű́	*mar-ű́
	2	*tké-m	*tế-m	*gdóq-na	*kól-na	*mker-ű́	*mar-ű́
	3	*ké-m*	*tế-m*	*dóq-na*	*kól-na*	*ker-ű́*	*mar-ű́*

3.3. Stress attracting prefixes bleed pretonic reduction

Note that in the Iranic varieties, negative prefixes are stress-attracting, making the imperfective prefixes *de-* and *me-* non-pretonic and thus bleeding the conditioning context. In (4), I show the relevant changes. In the negative forms, the imperfective markers remain intact. Subsequently, the stress from the negation prefix shifts to the full stem vowel, and there is a post-vocalic lenition of *d in Kurdish. The *nm cluster in Hewramî reduces to *m*, but the dy (consonant-glide) cluster in Southern Kurdish remains. This creates the famous cheshirization of the negative marker in Hewramî: the segmental features of the morpheme are lost, while the suprasegmental property of stress remains hosted by the imperfective marker, lost in most other contexts (Karim and Mohammadirad forthcoming).

(4)	**SK Kermanşa**		**NENA Sanandaj**		**Gorani Hewraman text**	
	*de-ké-m	*ní-de-ke-m	*ka-doq-na	*la-ka-doq-na	*me-ker-ú́	*ní-me-ker-û
	*d-ké-m	*ní-de-ke-m	*k-doq-na	*la-k-doq-na	*m-ker-ú́	*ní-me-ker-û
	*d-ké-m	*n-yé-ke-m*	*k-doq-na	*la-k-doq-na	*m-ker-ú́	*n-mé-ker-û
	ké-m	*n-yé-ke-m*	*doq-na*	*la-doq-na*	*ker-ú́*	*mé-ker-û*

3.4. Leveling of allomorphy

In Kurdish varieties where the original negation marker did not have the allomorph *ni-* before the imperfective marker *de-*, the affirmative *di- [IPFV-] alternated with negative *ne-ye- [NEG-IPFV-], in some Southern Kurdish varieties. Some of these varieties generalized the proto-tonic version of the imperfective prefix (that in the negative context) just like in Old Irish (Russell 2017, 1286).[3]

3.5. Leveling of allomorphy

There are surface developments that are indicative of the changes described here: a present indicative marker is used only with h-initial verbs. In the case of both Kurdish and Gorani, the existence of varieties that preserve the original marker in all contexts shows that these are remnants of the original marker. All other verbs either have no marker or a reduced form due to sound change and morphological analogy. In Balochi, there is a regular imperfective/indicative marker, the prefix *a-* or the preposed enclitic *=a*. However, h-initial verbs also have a different marker only on those verbs; see *Table 5*.

Table 5. Balochi h-initial verbs

	PRS.IND.3SG	PRS.SBJV.3SG	INF
come	*(a-)ḵ-ey-t*	*b-ey-t*	*hātin*
bring	*(a-)ḵ-ār-īt*	*by-ār-īt*	*hāwurtin*
wait	*(a-)ḵ-ōšt-īt*	*b-ōšt-īt*	*hōštātin*
etc.			
do/make	*a-kan-t*	*b-kan-t*	*kurtin*

[3] In Old Irish, there were two different verbal stems: in the affirmative, the stress was on the stem, while in the negative, the stress was on the negation marker. The result was two different voicing patterns, e.g., *do·beir* 'gives' and *do·gaib* 'seizes,' vs. *ní tabair* 'doesn't give' and *ní digaib* 'doesn't seize.' For some verbs, the reduction creates truly suppletive forms, e.g., *do·sluindi* 'denies' vs *ní diltai*. In modern Irish, it is the so-called proto-tonic stems, the ones used in the context of negative, that were leveled to affirmative contexts, e.g., *díltaid*: *diúltaíonn sé* 'he refuses' *ní dhiúltaíonn sé* 'he doesn't refuse' (Russell 2017, 1286).

Unlike the vast majority of Southern Kurdish, the now productive marker *a-* can co-occur with the prefix *k-*. That being said, there are Southern Kurdish varieties where the *t-* marker with h-initial roots has been reanalyzed by younger generations as part of the imperfective stem, to which the productive marker can be added, e.g., Qorwe *e-tê-m* [IPFV-come.PRS.IPFV-1SG] by speakers under 30 and *tê-m* [come.PRS.IPFV-1SG] by older speakers (Mohammadirad and Karim 2025). This is clearly a recent innovation, representing a likely analogical progression: an opaque form is augmented or replaced by a productive form. Likewise, there are Balochi varieties where it seems that the regular marker *a-* does not cooccur with *k-*, e.g., some Coastal Balochi varieties.

Unfortunately, a consistent corpus of Balochi would be required to confirm this, which is not available. Many early Balochi scholars like Barker and Mengal (1969, 149) dismissed the prefix as a "meaningless" element and did not consistently mark it. Additionally, in some varieties, the prefix is lost if the preceding word ends with a vowel. Even a well-annotated and transcribed Balochi study could give the incorrect impression. There is one possible example of a variety where the markers *k-* and *=a* are incompatible from the Coastal Balochi variety of Korsar. The first text in Nourzaei (2017, appendix A) *Bādšāhay J̌anek* 'The King's Daughter,' recounted by a 45-year-old female speaker of Coastal Balochi, Zinat Jadgal from Korsar, never uses the two imperfective markers together in 181 lines (see 5). The second text from Coastal Balochi in Nourzaei (2017, appendix A) *Rahimbakshe Kessaw* 'The Story of Rahimbaksh,' recounted by a 58-year-old male speaker of Coastal Balochi, Rahim Shakalzahi from Nobandiyan, uses the markers together consistently.

(5) a. *mollā* *wat* *k-ay-Ø*
Mullah REFL IMPF-come.NPST-3SG
'the Mullah himself comes' (Nourzaei 2017, KD.f: 28b)

b. *edā* *k-ay-Ø* *hawrok-ayā*
here IPFV-come.NPST-3SG hawrok-LOC
'she comes there (lit. here) to Hawrok's [place]' (UT, Nourzaei 2017, 44)

c. *ē* *ǰanēn* *dar-k-ay-Ø* *raw-t*
DEM.PROX woman PV-IPFV-come.NPST-3SG go.NPST-3SG
otī *sar-ā* *zīr-ī*

REFL.GEN head-OBL take.NPST-3SG

'this woman went all alone (lit. went took her head)' (KD.f.CoB: 88a–88b, Nourzaei 2017, 130)

d. *dā̆ke ē ǰanēn k-ay-Ø*

you.know DEM.PROX woman IPFV-come.NPST-3SG

'You know, this woman keeps coming' (KD.f: 90–91c, Nourzaei 2017, 199)

Other varieties may prohibit the use of the *=a* and *k-* markers. Example (6) from a 60-year-old female from Bahukalat, clearly should have the marker *=a*, However, it is the only example from this variety:

(6) *ǰalā'hī baz'zag 'k-ay-t=ẽ*

Jalahi poor IPFV-come.NPST=cop.NPST.3=SG

'Poor Jalahi returns' (HA.f: 4–5), Nourzaei 2017, 206)

The texts from Nourzaei (2017) suggest that there is an incompatibility of *=a* and *k-* in some varieties. This introduces the possibility that, like Kurdish, where the dental prefix *d-/t-* and the productive *e-* prefix have the same etymon *de, the Balochi *a-* and *k-* prefixes are related, perhaps *ka-.

3.5.1. Parallel pathways from *Car to suppletive stems

According to Karim (2024), the Kurdish prefixes ultimate have their roots in the locative marker *de-* < Old Iranic *antara Middle Persian *dar*. This surfaces in the adpositional system of Central and Southern Kurdish varieties as the preposition *de*, identical to the imperfective prefix in northern Central Kurdish varieties. In some Southern Kurdish, the form *e* matching the prototonic allomorph also occurs. It also surfaces as the postposition *da* featuring a different vowel quality and the post-vocalic allomorph featuring the lenition of the *d in *da*. These are often combined as the circumposition *de* NP *da*.

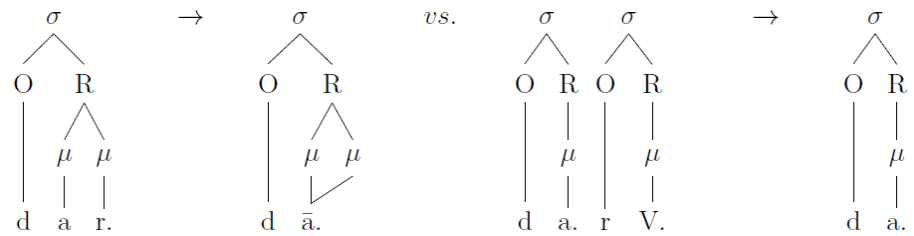
3.5.1.1. Parallel pathways from *Car to suppletive stems

There are some foundational assumptions required to arrive at the extant forms of the adpositions and the imperfective affixes as proposed by Karim (2024). It is generally recognized that the circumposition *de...da* consists of two elements that are the reflexes of the same etymon *antara. The development of this form must have gone through several well-documented

changes: (1) post-nasal voicing *andara; (2) the "rhythmic law" where final codas were lost *andar, and as in most Iranic languages (3) clipping to *dar. In the Hawar Kurdish orthography, this form would be *der*. Up to this point, all these changes are attested in other Iranic languages, including Persian.

In order for *der to develop into the extant postpositive element *da*, the final *r must have been lost with compensatory lengthening of the preceding vowel. In contrast, the prepositive element *de* the *r must have been lost without compensatory lengthening. We can understand the concept of compensatory lengthening in terms of Moraic theory (Hayes 1989). Essentially, syllabic nuclei and coda consonants carry moras while onset consonants do not.

Figure 1. Compensatory lengthening as moraic reasignment (σ: syllable; μ: morah; O: onset; R: rhyme)



The examples in Figure 1 show two different types of changes. The first shows the sequence *dar, which occurs at a syllable boundary. If a regular sound change causes the final coda consonant *r to be lost, the mora must be reassigned to the previous vowel. In contrast, if the final consonant is followed by a vowel, it resyllabifies as the onset of the following syllable. If that *r is lost, no moraic reassignment is possible. One possible reason for such a change is boundary reanalysis, where a formative is reinterpreted as part of a different morpheme. Examples from English include *a napron > *an apron*, *an ewt > *a newt*, etc.

Regardless of one's ultimate conclusion about the source of the Kurdish imperfective prefixes (for which no viable etymon has been proposed other than that in Karim 2024), these developments must be assumed to explain the prepositive and postpositive elements of the locative circumposition *de...da*. Additional changes are necessary to explain the form of the indicative/imperfective prefix *e-* (also the preposition *e* found in some Southern Kurdish varieties).

3.5.1.2. Deriving two formatives from one etymon

Recall that there are two reflexes of *dar (< *antara) in Kurdish, the preposition *de=* and the postposition *=da*. The latter lost its final *r with compensatory lengthening and the former without; see *Figure 1* in *Section 3.5.1.1*. In order for the form in Balochi to occur as *a-*, we must assume that the final *r was lost through boundary reanalysis, i.e., not conditioning compensatory lengthening. If this were true, the only possible trace of the original *r would be non-etymological *r*s at the beginnings of words. Note that Karim (2024) makes this assumption for Kurdish without giving examples, and the absence of proof is not proof of absence.

In Balochi, there is a clear possibility that points to this boundary reanalysis in the Minabi and Koroshi varieties (Nourzaei and Jahani 2012). In both Koroshi and Minabi, the imperfective marker is typically *a-* as in (7a). However, when a verb begins with *r*, instead of the typical marker *a-*, the imperfective marker *ar-* is employed; see (7b) and (7c).

(7) **a.** *a-žan-ant*

IPFV-arrive.PRS-3PL

'they arrive' (Koroshi, Nourzaei and Jahani 2012, 175, edited)

b. *ar-rapt-ad=en*

IPFV-go.PST=COP.PST-1PL

'we went' (Minabi, Nourzaei and Jahani 2012, 178, edited)

c. *ar-r-ant*

IPFV-go.PRS-3PL

'They go' (Koroshi Nourzaei and Jahani 2012, 174, edited)

This gemination points to two possibilities: (1) assimilation, (2) reanalysis. In the table of common Balochi sound changes described in Korn (2003, 71–75), there is no documented assimilation resulting in geminate *r*s. According to Korn (2003, 55), the majority of geminate clusters in Balochi come from foreign borrowed words. However, they can occur in native vocabulary through assimilation or "isometric substitution" where a long vowel is shortened, and its mora is reassigned to the following consonant, e.g., *jūrāb > *jurrāb* (Korn 2003, 209). Based on the presumption of a *dar-like formative which had the short vowel *a, we must reject "isometric substitution" as a possibility.

The existence of the *r* formative here should be taken as part of the likely etymon, as there are no known sources of the *rr* cluster that would produce

it in this environment. This formative could then, in principle, be reanalyzed as part of the *r-* initial roots. Nourzaei and Jahani (2012) also shows another allomorph in Koroshi Balochi. The form of the prefix with a long vowel *ā-* occurs sporadically in Nourzaei and Jahani's (2012) description before verbs beginning in *k*. This could, in principle, be a compensatory lengthening of the vowel after the loss of a coda consonant, which does occur in Balochi. However, this is uncertain especially as Nourzaei and Jahani (2012, 175–76) shows identical examples *a-k-āy-ant* and *ā-k-āy-ant* [IPFV-IPFV-come.PRS-3PL]. It is unclear if there is a systematic rule that explains this allomorph. In its occurrence with *h*-initial roots, there is another possibility. There are copious examples of k-fricative clusters undergoing metathesis *ks > *sk*, *kš > *šk* (see Korn 2003, 172–73 and corresponding sections). If this metathesis occurred with *h as well, the resulting *the sequence would become *āk* through regular sound changes (see Korn 2003, 252). This formative is not used consistently according to Nourzaei and Jahani (2012), and it is used with verbs with etymological *k*s as well as *h*-initial roots. As such, I refrain from making any claims regarding its provenance, except that there are these possibilities.

Due to the geminate *r*s in (7), I continue to entertain the possibility that the ultimate etymon ended in the consonant *r. In other words, it is possible that a proposed *C**ar*** could have become *C**a*** through the reanalysis of a morpheme boundary. Note that this must have been the case in Kurdish, where the etymology of the *de-* prefix is known, but it is also supported by these facts in Balochi.

3.5.1.3. Unimorphation

Recall that in Kurdish, the form of the imperfective prefix, *de-* in some varieties and *d-* in others, unimorphates with *h*-initial stems. This occurs in two ways, either with the *h deleted intervocalically (Central Kurdish) or with the *d devoicing in contact with the following *h (Southern Kurdish).

In the vast majority of Balochi varieties, *h*-initial roots get an imperfective prefix *k-*, which may be interpreted as a part of a suppletive imperfective stem. The Balochi verbs that commonly take the *k*-type imperfective marker are *(k)āyag* 'to come,' *(k)ārag* 'to bring,' *(k)illag* 'to leave, abandon, let go,'

(k)oštag[4] 'to stand,' *(k)andag*[5] 'to laugh,' and *(k)uškinag*[6] 'to listen, hear' (Barker and Mengal 1969, 133–34). The first three of these verbs in Balochi correspond nicely to the set of verbs in Kurdish that have developed suppletive imperfective forms, e.g. *hatin/tê-* 'to come,' *hil-* 'leave (it),' and *hawirdin/têr-* 'to bring.' With the exception of the forms that are not cognate, the Balochi *k*-stems are parallel to the Kurdish suppletive *t-/d*-stems.

From the perspective of the suppletive imperfective stems in Kurdish, it is not possible that the Balochi *k-* has descended from *dar. There is no evidence in Balochi of a pretonic shortening (*dehê- [come.IPFV.PRS-] > *dⁱhê-), eventually leading to the devoicing of the *d element (*dhê- > *tê-*) like that of Southern Kurdish. As such, we must assume a *k formative, not a devoiced *g or similar form. A prefix *ka could, in principle, have fused to the stem with the loss of the *h* intervocalically. Note that this is what happened in Central Kurdish varieties, where there was no pretonic shortening, creating the conditioning environment for voicing assimilation. For example, CK Kerkûk has *dê*. The loss of intervocalic *h that led to *deê (h → Ø / C,V__) and the vowel hiatus resolution strategy (e → Ø / __V, V__) are well attested in Central and Southern Kurdish. A similar development in Balochi is possible, e.g., *kaheyt > *kaeyt > *keyt*.

At this point, we must reject the *dar etymon (which was never really a possibility). However, we can reevaluate this as *kar. The same structures that allow Kurdish *dar to become the forms *e-* and *d-* allow for *kar to become Balochi *a-* and *k-*.

3.5.2. The etymon *kar

In this section, I propose a new possible unified etymon for the Balochi imperfective markers *=a*, *k-*, *a-*, and *ar-*. It should be clear that this proposal

[4] The Kurdish verb *standin/sta-* lacks a suppletive present-tense form, and its past stem does not begin with *h*. According to Korn (2005) and Cheung (2006), *aa-staH is the source of Balochi *ōšt-*. The preverb *aa ultimately derives from PIE *H_2eu (Beekes 2011, 247), consistent with Kümmel (2014), who reconstructs *h* as the Balochi reflex of PIE *H_2 in initial position (e.g. Balochi *hīz* 'leather' < *H_2iĝ-). By contrast, the Kurdish equivalent with the *Haa- prefix, *westan* 'to stop', loses the initial syllable and does not preserve the laryngeal.

[5] Balochi *(k)andag* 'to laugh' corresponds to Kurdish *kenîn* and Persian *xandan* (Cheung 2006, 443). From this comparative evidence, it is unlikely that the Balochi form would feature an etymological *h*. If this line of reasoning is sound, it is possible that a verb with an inherited initial *k-* could be reanalyzed in this class of verbs. Korn (2005) has both *kand* and and *hand* as variants which points to this type of reanalysis.

[6] Balochi *(k)uškinag* 'to listen, hear,' (Korn 2005) suggests that the verb is built from the word for 'ear' Av. *us* (Korn 2005, 350). This word ultimately goes back to *H_2us-; the laryngeal consonant corresponds to an *h* in both Balochi and Kurdish (Kümmel 2014).

matches parallel developments found in Kurdish and respects the internal diachronic phonology of Balochi. However, it is based on limited data. Unfortunately, as stated in *Section 3.5*, many early Balochi scholars like Barker and Mengal (1969, 149) did not consistently mark the imperfective marker. Despite the paucity of reliable data, recent documentary work by Maryam Nourzaei has enabled tentative reconstructions.

The most common form of the imperfective marker is the prefix *a-* (or the proposed enclitic *=a*. Sometimes this marker appears as the long allomorph *ā-*. Based on these facts, we can reconstruct *ă as the original formative without specifying the length. Based on the findings of Korn (2003), there are phonological contexts where a long *ā could be shortened and where a short *a could be lengthened.

In addition to the *a, there is a secondary imperfective marker used with h-initial verbs. Recall that h-initial roots in Kurdish have integrated the dental of modal origin into the imperfective stem (Fattah 2000, 399). However, where the Kurdish forms integrate a dental *t-* or *d-*, Balochi varieties incorporate a *k-*. Due to the absence in Balochi of pretonic shortening that resulted in the Kurdish *d* forms becoming *t*, it is safe to assume that there was formerly a *k formative that was part of the imperfective marker: *ka or *ak. The simplest explanation would be an original *ak formative as both *a-* and *a-* coexist in this order in many varieties. However, preservations in Kurdish have shown us how the suppletive imperfective prefix *t-* unimorphates with the present tense stem: Qorwe (> 35) *t-ê-d* [IPFV-come.PRS-3SG] *tê-d* [come.PRS.IPFV-3SG]. In younger generations (< 35 years), the regular imperfective marker of the same origin is added back to this now opaque form: *tê-d* [come.PRS.IPFV-3SG] *e-tê-d* [IPFV-come.PRS.IPFV-3SG]. The "smoking gun" that would allow us to decide between *ka and *ak would be a Balochi variety where *k-* and *a-* were incompatible. There is no reason why these formatives would be incompatible unless they were from the same etymon with different allomorphs in different contexts, as in Kurdish. The regularization of the more common allomorphs to what is now a suppletive imperfective stem represents a leveling restoring regularity to the system. However, the lack of a variety with this incompatibility does not point to *ak in the same way that the existence of such a variety would point to *ka. See *Section 3.5* for a possible example of a variety where these formatives are incompatible.

In order to produce the most commonly used marker *=a/a-*, we must assume the loss of the *k in *ka-*. There is no phonological rule that would account for the following descriptions from Korn (2003). One might assume, as (Kolyâi, Qorwa (C.D.G.), and Bayray Fattah; 2000, 372) proposed for certain Southern Kurdish varieties, that there was an additional boundary reanalysis that fused the initial consonant of the prefix to the preceding word;

see the Southern Kurdish example (8), showing *=y* the post-vocalic allophone of the prefix *d-*. However, in Balochi, there are a massive number of non-etymological final *k*s. It is not possible to determine if they have come about through boundary reanalysis or by suffixation of the diminutive/evaluative *-k*. However, it is possible.

(8) *xormâ=y* *xwa-m* (Bayray: Xayrsuni & Musi)
date=IPFV eat.PRS-1SG
'I eat dates' (Fattah 2000, 372)

The existence of termination with *r*-initial verbs in Koroshi may give us a clearer idea of the shape of the original formative. I propose that the original *kar formative underwent a boundary reanalysis where the geminate cluster was created, which was interpreted as part of the stem, reduced to a single *r* when the prefix was removed. Similar evidence was used to reconstruct the Semitic definite article from the presentative [particle], or its reflexes *han/hal* (Pat-El 2009, 47). Both the Hebrew and Arabic definite articles *hā-* and *al-* assimilate the final n/l to a following consonant. As the Balochi form is restricted to doubling only /r/s, I tentatively propose /r/ as the original formative.

The *kar proclitic is attached to the *r*-initial stem. Subsequently, reanalysis of the morpheme boundary resulted in the loss of/r/from the proclitic. This is well attested in the world's languages. E.g., English *an apron* < *a *napron and *a newt* < *an *ewt.[7]

Let us assume an original *kar formative as the source for the Balochi imperfective markers. In that case, there is a likely template for the development from the verb 'to do' to a marker of progressive aspect (and ultimately imperfective aspect, following Deo 2015). In fact, *kar 'do' as a source of a progressive/imperfective marker is attested in the Caspian region among languages with which Balochi is known to have been in contact. For instance, among the Caspian languages, there are two common patterns for marking progressive and imperfective aspects (as well as proximative): the DAR and KAR constructions (Vafaeian 2018). The DAR construction is from the locative marker (just as Karim 2024, proposed for Kurdish), and KAR construction is from the verb 'to do.' According to (Stilo 2001), different Gilaki varieties form their progressive aspect with various strategies: a locative suffix (e.g. 9a and 9d), an invariable prefix *kəra* (9b), and a compound formation consisting of both (e.g. 9c and 9e). According to Korn (2019), Balochi features some grammatical convergence with Gilaki.

[7] In English, this boundary shift and others like it are likely the reason for the allomorphy between *a* and *an*, as the loss of /n/ is not a regular sound change.

(9) **a.** *gift-án-dər-əm*
take-INF-LOC-1SG
Western Gilaki: 'I am taking' (Stilo 2001)

b. *kəra gir-ám*
KAR take.PRS-1SG
'Western Gilaki: 'I am taking' (Stilo 2001)

c. *kəra gift-án-dər-əm*
KAR take-INF-LOC-1SG
'Western Gilaki: 'I am taking' (Stilo 2001)

d. *git-é-dər-əm*
take-INF-LOC-1SG
Eastern Gilaki: 'I am taking' (Stilo 2001)

e. *git-e-ká-dər-əm*
take-INF-KAR-LOC-1SG
Eastern Gilaki: 'I am taking' (Stilo 2001)

The Balochi imperfective marker could be a retention of part of an original proclitic *kar=* that reanalyzed its *-r* to be part of or conditioned by the following word, leaving a marker *ka=* in its place. A hypothetical Balochi imperfective marker *ka= was assimilated to *h* initial verbs just as Kurdish *de-* was assimilated to a subset of the same verbs. In Kurdish, there was a sandhi variant of *de* which appeared as *(y)e-* after the negative marker or even vowels more broadly. This sandhi variant was extended to non-applicable contexts. This is a bit speculative for Balochi. However, the negated forms of the present indicative never feature the *k* formative, e.g., Koroshi Balochi *a=nay-ay* [IPFV=NEG-come.PRS.3SG] *nay-aškont-a* [NEG-hear.PST-PCPL] (Nourzaei et al. 2015, 137, 227). This, coupled with the fact that the negative marker does not precede the imperfective marker (e.g. *a=na-twān-t* 'it cannot' [IPFV=NEG-can.PRS-3SG]) (Nourzaei and Jahani 2012, 175), lends some weight to the possibility that the two imperfecives *a-* and *k-* have the same etymon, the unimorphation of the imperfective prefix and the stem being blocked by an intervening negative marker. This word order is consistent with what is observed in Gilaki, e.g. *kára bišt-án-deí-yə* 'he is frying [KAR fry-INF-LOC-NEG-COP.3SG]' (Stilo 2001).

The development from *kar to Balochi *a-* and *k-* is parallel to what Karim (2024) proposed for Kurdish: (1) The original form *kar loses the *r through boundary reanalysis. Koroshi and Minabi Balochi give us the possible

conditioning context for this, r-initial verbs: *kar + *rant *kar.rant give us *ka.řant, separable into *ka and *(r)rant. Note that in many Caspian varieties this is the final stage in what Stilo (2018) calls short-form progressive, e.g., *zü=mun* ***ko*** *š-e* [early=1PL PROG go-INF]. (2) The *ka formative merges with h-initial roots: *ka + *(h)ayt *kaayt becomes *kayt* because of h syncope and vowel hiatus resolution. This *ka-zant 's/he knows' *kayt* 's/he comes' stage is parallel to CK Mukrî *de-zanê* '*s/he knows' dê* 's/he comes'. (3) The varieties with separable *ka- prefixes would then have to lose their *k through another boundary reanalysis, i.e., with a preceding k-final word. Balochi nouns ending in *k* are common and plentiful: *bačak + *ka + *rawt 'the boy goes' could easily group *bačak-ka *rawt, leading to a reanalysis *bačak-ka [boy-IPFV] > *bačakk-a* [boy-IPFV]. This possibility is bolstered by the fact that this set of nouns features a geminate final *kk*, e.g., *bačakk* 'boy,' *jinikk* 'girl,' *kuçekk* 'dog,' *tupekk* 'gun,' etc. Korn (2005: 169) states that the suffixes with the geminate *kk* still need to be explained and that they likely come from something like *a-ka-ka [THEME-DIM-DIM], presumably with the second *ka being a second diminutive/evaluative ending. Phonologically this works for my proposal as well, i.e., *a-ka-ka [THEME-DIM-PROG].[8]

If the adaptation of the *kar construction is a result of contact with the Caspian languages (as suggested for the case system by Korn 2019), it must have happened very early in the development of Balochi, i.e., early enough to have affected all varieties and subgroupings. The linguistic genealogy of Balochi is clear. It is a so-called "northwestern" Iranian language. Its nearest modern relatives are Kurdish, Tati, Ṭāleši (Elfenbein 1989) According to Elfenbein (1989), "the Baluch tradition of a migration to their present habitat from the west in the 7th-8th centuries a.d. has an echo of history in it, strengthened by the linguistic connections of Baluchi, and one is led to the assignment of the original home of the Baluch to somewhere just east or southeast of the central Caspian region". I caution against the use of historical data, which is in its nature more malleable than linguistic affiliation. However, historical works such as Dashti (2012) locate the Origins of the Baloch similarly in the Caspian, suggesting a connection with Bālāsaāgan in present day Azerbaijan, suggesting a migration through the Caspian region from West to East passing through all the relevant genealogically related contact languages.

The *kar construction is well attested as a preverbal formative in Tati: Kelasi, Koluri, Gandomabi, Shali, Hezarrudi; Gilaki: Rashti; and Taleshi: Masulei

[8] The Balochi boundary reanalyses that I propose are similar to the way that the English indefinite article *an* lost its nasal to vowel-initial words, becoming *a* or the possessive *mine* [main] lost its nasal to the Nicknames *Ed* and *Elly*, rendering *my* [mai], *Ned*, and *Nelly*.

and Masal-Sandermani, etc. (Vafaeian 2018). This is exactly the proper syntactic prerequisite for the developments I have proposed here for Balochi. There are even varieties in the Caspian that have gone through some of the reductions that must have taken place in Balochi, e.g., Charozh Talyshi: *æmæ zü=mun* ***ko*** *š-e* [we early=1P1 PROG go-INF]. However, the absolute chronology is unclear, and independent innovation is possible. The development of the verb 'to do' into a marker of progressive aspect is attested independently in other languages, e.g., Tucano (Brazil; see Bybee et al. 1994; West 1980, etc.).

If this is indeed contact, Caspian to Balochi, the *kar type progressive was active in the region before the Balochi left the Caspian which must have been in the late Sassanian period as, following Elfenbein (1989), "many areas of Kermān and Sīstān may have been at least partially occupied by Baluch migrants by the 8th century, for at the time of the Arab conquest of Kermān in 644, it is stated by later geographers that they came into contact with large numbers of Qwfṣ and Blwṣ, "Kōč and Balōč"". At this point, it is impossible to tell whether contact or independent developments are more likely. If further study identifies more lexical or morphological contact phenomena, the case for contact will be strengthened significantly.

4. Mutual developments: past-Imperfective and conditional forms

The next point of convergence is the relationship between the past-imperfective and conditional forms. In the relevant languages, these formatives are often constructed from the same material. For instance, in Goranî, the same suffix that marks imperfective aspect when attached to the present tense stem marks conditional when attached to the past stem, e.g., Paweyane: *d-ê* [give.prs-pst.ipfv.3sg] 's/he used to give' vs. *da-ê* [give.pst-pst.ipfv.3sg] 'has (smn.) given s/he/it'. Likewise, in Southern Kurdish varieties that distinguish imperfectve stems, they match the conditional, e.g., Qesiri Şîrîn: *der-hat-ya-n* [out-come.PST-IPFV-3PL] 'I used to emerge' vs. *der-bi-hat-ya-n* [out-SBJV-come.PST-IPFV-3PL] 'had I emerged.' Other varieties have lost the stem identity in favor of the imperfective prefix discussed above, e.g., Qorwe: *e-hat-in* [IPFV-come.PST-3PL] vs. *bi-hat-a-n* [SBJV-come.PST-IPFV-3PL]. As such, the developemnt of the imperfective prefix and the past conditional markers are intrinsically linked.

As we have seen in *Section 1*, the past-imperfective and past-past conditional originally had the same form. This was true of the Old and Middle Iranic optative, and it carried over into New Iranic languages. This is in its original form in Gorani, where the imperfective stem combines with the optative endings, creating the imperfective, used for irrealis and habitual past actions.

The optative of the copula combines with the past-perfective stem to create the past conditional, used in the protasis of irreal conditionals; see (10).

(10) Hewram Text

eger *řeza=şa* *ne-bî-ɛ*

if satisfaction.F=3PL:NC NEG-be.PST.COND.3SG:S

ne-đ-ênê

NEG-give.PRS-IPFV.3PL:A

'If they didn't agree, they wouldn't give her.' (Mohammadirad forthcoming, 294)

4.1. Persian developments

In the Middle Iranic Parthian language, the form of the optative became a single invariant form *-ende*. By Early New Persian, the form of the optative, already rare in Middle Persian, had lost most forms of its conjugation, becoming a frozen particle *=i*, and it only attached to the fully inflected past stem. See (11), where the form ending in the particle *=i* marks past habitual action, past conditionals, and irrealis.

(11) Early New Persian (Lambton 1960, 161)

gar *ān-hā* *ke* *mī-goft-am=i* *kard-am=i*

if DEM.DIST-PL REL IPFV-say.PST-1SG=HAB do.PST-1SG=COND

neku_sirat *o* *pārsā* *bud-am=i*

good.of.character and pious be.PST-1SG=IRR

'If I had done those things which I used to say, I would have been of good character and pious.'

At some point, this form becomes compatible with the perfective prefix *be-* (later the subjunctive prefix) to emphasize the conditional reading; see (12).

(12) Early New Persian (Lambton 1960, 162)

tā *be-dānest-am=i* *ze* *došman* *dust*

to PFV-know.PST-1SG=COND from enemy friend

zendegāni *do* *bār* *bāyest=i*

life two times must.PST.3SG=IRR

'Life would be needed twice over to know friend from foe.'

In modern New Persian, this construction has been replaced by the simple past tense with the imperfective prefix *mī-*; this is used in all functions: past imperfective, past conditional, and irrealis; see (13).

(13) New Persian (Rubinčik 2001, 245)

agar mī-dānest-am be šomā mi-goft-am

if IPFV-know.PST-1SG to 2PL IPFV-tell.PST-1SG

'If I had known, I would have told you.'

4.1.1. Zazaki forms

In Zazaki, we observe a state of affairs that closely matches that of Early New Persian. The past imperfective, irrealis, and conditional are all marked by the simple past form, with an invariable suffix *-êni*. The subjunctive prefix *bi-* is mandatorily added to this form when it is used as a conditional; see (14).

(14) Şeyxan Zazaki (Paul 2003, 181)

yûwerna bi-vat-êni mi nê-rot-êni

someone SBJV-say.PST-IPFV 1SG.OBL NEG-sell.PST-IPFV

'If someone else had said so, I would not have sold it.'

4.1.2. Kurdish developments

The Zazaki situation closely mirrors the Southern Kurdish forms. In *Table 6*, these stages can be observed changing away from a Zazaki-like system. Many Southern Kurdish varieties have an inherited system in which the past imperfective/irrealis form retains the inherited ending. The past conditional is identical in form, with the addition of the *bi-* prefix; see examples from Arkewazî Xaneqîn, Qesri Şîrîn, Kirmanşa, and Şerwan Kileway.

The verb 'to be' cannot typically occur with the *bi-* prefix (although this has changed in some varieties). See the Qasri Şîrîn example where the forms of the past imperfective and conditional are identical. In other varieties, the ambiguity of form is corrected by adding a second imperfective suffix to the conditional form, e.g., Bîlewar. Note that the Bîlewar imperfective is doubly marked for imperfectivity with the prefix *e-* causing the post-vocalic lenition of *b.

Other varieties have replaced the suffixed form of the imperfective with the simple past stem and the imperfective prefix from the present system. In these varieties, the suffix only remains as part of the conditional form, as in Wermizyar and Melikşay Rasulwan. However, some of these varieties have

retained the form from the formally ambiguous roots with a reduplication of the conditional marker, e.g., Melikşay Xamîs.

Table 6. Evolution of Kurdish forms Conditionals

		PST.IPFV	**PST.COND**
Arkewazî Xaneqîn	'emerge'	*der=çig-ya-n*	*der=bi-çig-ya-n*
		out=go.PST-IPFV-3PL	out=SBJV-go.PST-IPFV-3PL
Qesri Şîrîn	'emerge'	*der=hat-ya-n*	*der=bi-hat-ya-n*
		out=go.PST-IPFV-3PL	out=SBJV-go.PST-IPFV-3PL
Kirmanşa	'cut'	*biř-ya-n*	*buř-ya-n*
		cut.PST-IPFV-3PL	cut.SBJV.PST-IPFV-3PL
Şerwan K	'do'	*kird-ya-n*	*bi-kird-ya-n*
		do.PST-IPFV-3PL	SBJV-do.PST-IPFV-3PL
Qesri Şîrîn	'be'	*bî-ya-n*	*bî-ya-n*
		be.PST-IPFV-3PL	be.PST-IPFV-3PL
Bîlewar	'be'	*e-ü-a-n*	*bü-a-t-a-n*
		IPFV-be.PST-IPFV-3PL	be.PST-IPFV-HI-IPFV-3PL
Wermizyar	'come'	*di-hat-in*	*b-at-a-n*
		IPFV-come.PST-3PL	*SBJV*-come.PST-COND-3PL
Melikşay R	'throw'	*di-xist-in*	*bi-xist-a-n*
		IPFV-throw.PST-3PL	*SBJV*-throw.PST-COND-3PL
Melikşay X	'throw'	*di-xist-in*	*bi-xist-a-g-a-n*
		IPFV-throw.PST-3PL	*SBJV*-throw.PST-COND-HI-COND-3PL
CK KerKûk	'come'	*e-hat-in*	*bi-hat-in-a-ye*
		IPFV-come.PST-3PL	SBJV-come.PST-3PL-COND-COP.3SG
NK Pertek	'hold'	*di-girt-in*	*bi-girt-a-n-a*
		IPFV-hold.PST-3PL	SBJV-hold.PST-COND-3PL-COND-

Moving northward through Central Kurdish, the secondary addition of the particle *-a* takes place after the personal endings (like Early New Persian *=i*), and there is a frozen third-person singular form of the copula added *-a-ye* on the past conditional. In Northern Kurdish, the past conditional ending *-a* is followed by personal endings, just as in Southern Kurdish, but the secondary addition of the particle also occurs, yielding multiple exponence *-a*-APM-*a*. The imperfective is formed by adding the present-tense imperfective endings to the simple perfective past stem in all Northern and Central Kurdish varieties. There is only one possible exception to this development. In an

unknown Northern Kurdish variety documented by Bulut (2000), the imperfective still occurs with the suffix *-a*; see (15).[9]

(15) Northern Kurdish (Bulut 2000, 158)

mişk	*nan-ê*	*me*	*di-xwar-a*
mouse	bread-EZ.M.SG	1PL.OBL	IPFV-eat.PST-IPFV.3SG

'The mouse was eating our food.'

4.1.3. Balochi forms

In Balochi, there is little documented variation between imperfective forms and conditional forms. These forms are comparatively rare in the languages, and they are notoriously difficult to elicit. The forms in Matras et al's (2016) Kurdish dialect corpus variably show the forms of the past conditional and the past-perfect conditional, i.e., the past stem with the past conditional of 'to be.'

There are, however, some assumptions we can make based on the forms in Balochi. Firstly, the past conditional is formed by the past stem, e.g., *šut* [go.PST], the subjunctive prefix *b-*, the reflex of the old optative suffix *-ēn*, and affix person markers: *b-šut-ēn-un* [SBJV-go.PST-COND-1SG] (Axenov 2006, 199). However, this form can also occur without the prefix; see (16).

(16) Balochi Turkmen (Axenov 2006, 191)

aga	*zānt-ēn-un*	*šmā-rā=a*	*gušt-un*
if	know.PST-COND-1SG	2PL-OBL=IPFV	tell.PST-1SG

'If I had known, I would have told you.'

Assuming that the form came into Balochi with all of its inherited Middle Iranic functions, it stands to reason that the conditional form without the subjunctive prefix once had irrealis and habitual function. That being said, there are no known Balochi varieties that have an imperfective ending in *-ēn*. Of course, the same could have been said about Kurdish before Fattah's (2000) study of Southern Kurdish dialects, showing the full evolution from COND = SBJV-IPFV(STEM) to a system where COND = SBJV-IPFV(STEM) but the innovative IPFV = IPFV-PST(STEM).

[9] While the dialectal provenance remains unclear, this example from Bulut (2000) suggests that the suffixal imperfective may be more widespread in Northern Kurdish than currently documented, warranting future investigation.

Table 7. Stages of past imperfective and conditional by language

IPFV	COND	
IPFV(STEM)-OPT	PFV(STEM)-OPT	Gorani
PFV(STEM)-OPT	SBJV-PFV(STEM)-OPT	Zazaki, Southern Kurdish
IPFV-PFV(STEM)-OPT	SBJV-PFV(STEM)-OPT	Southern Kurdish (Northern Kurdish?)
IPFV-PFV(STEM)	SBJV-PFV(STEM)-OPT	Northern, Central, and Southern Kurdish, Balochi
IPFV-PFV(STEM)	IPFV-PFV(STEM)	Modern New Persian

These convergences are shown in *Table 7*, where I use IPFV(STEM) to refer to the non-past stem and PFV(STEM) to refer to the past-tense stem. This reflects the etymology of these forms, but it is also synchronically true in the Gorani group. In other varieties, such as Kurdish, Persian, and Balochi, where the division is temporal (past/non-past) rather than aspectual (perfective/imperfective), I continue to use this language to reflect the use of the same forms. Additionally, I use OPT to refer to the suffixes used in these forms, although the Kurdish forms, for instance, do not necessarily have the same etymon. Note that the etymological account of Skjærvø (1997) leads much to be desired for the origin of the *-ēn* marker, and this disagrees with Karim (2024) who proposed that the Kurdish suffixes are originally locatives ultimately from the same etymon as the prefix forms: *de-* and *-ya* from *dar. None of these etymological considerations is relevant to the pattern discussed in this section.

5. Summary of changes

- Iranic languages have an imperfective marker that attaches to the past-tense stem of a verb to distinguish the perfective [PST.PFV-AGR] from the imperfective [**IPFV-**PST.PFV-AGR]. This marker is present in all varieties in *Table 8* as *mî-*, *de-*, *e-*, *e=*, *=e*, *me-*, or *=a*. It is missing from SK Sencewî and the Goranî varieties of Hewreman Text and Zerde, where the stem or stem extension differentiates the imperfective from the past perfective.
- In the present tense, being the exclusive domain of the imperfective, the same imperfective marker does not contrast the perfective but rather the subjunctive: [SBJ-PRS-AGR] vs [IPFV-PRS-AGR]. Note that the subjunctive prefix is fairly uniformly *bí-*, *be-*, or *b-*. The indicative prfix matches the past imperfective with a few variations The Gorani variety of Zerde has this marker in the present tense *me-* but not in

the past. In the Goranî of Hewreman Text, phonemic stress placement alone signifies mood in the present tense The stem *ker-* without stress is indicative, and *kér-* with stress is subjunctive. According to Karim and Mohammadirad (forthcoming), this stress placement is the result of "Cheshirization" of the prefix *bí-* on the subjunctive. In other words, regular sound changes led to the loss of all segmental features of *bí-*, leaving only its suprasegmental features, i.e., stress.

- There are various imperfective markers generally falling into several types: (1) a dental type with *t* or *d*, an *m* type, usually with *mî-* from *ham-aiwa 'always, same time, one time' (Windfuhr 2009, 26) or *ma-* from *ham- with the augment *a- (Yoshida 2009, 296).

- Some varieties, closely related to those with dental forms (*t* or *d*), have, instead, vocalic markers *a-, e-,* etc. This is most striking in *Table 9* CK Germiyanî, SK Qorwe, SK Sencawî, and Balochi, where there are regular vocalic markers or even no regular marker as in SK Sencawî. However, each of these varieties has either a remnant of the original dental marker *t/d*. To these, we may add the Balochi *k-*, possibly from *kar 'do.'

Table 8. Imperfective contrasts 'do'

	[PST.PFV.1SG]	[**IPFV**.PST.1SG]	[SBJ.PRS.1SG]	[**IPFV**.PRS.1SG]
NP (Tehran)	*kærd-æm*	***mî**-kærd-æm*	*be-kon-æm*	***mî**-kon-æm*
CK (Mukrî)	*kird=im*	***de**=m kird*	*bi-ke-m*	***de**-ke-m*
CK (Germiyanî)	*kird=im*	***e**=m kird*	*bi-ke-m*	***e**-ke-m*
SK (Qorwe)	*kird-im*	***e**-kird-im*	*bi-ke-m*	***e**-ke-m*
SK (Sencabî)	*kird-im*	*kird-**ya**-m*	*bi-ke-m*	***Ø**-ke-m*
L (Hersîn)	*kird-im*	***=e me**-kird-**ya**-m*	*bi-ke-m*	***=e me**-ke-m*
G (Zerde)	*kerd=im*	*ker-**ên**-ê*	*bi-ker-û*	***me**-ker-û*
G (Text)	*kerd=im*	*ker-**ên**-ê*	*Ø-kér-û*	*Ø-ker-**ű***
B (Turkman)	*kurt-un*	***=a** kurt-un*	*b-kan-īn*	***=a** kan-īn*

- Varieties with consonantal forms and their sisters with vocalic markers alike preserve the consonants with *h*-initial verbs.

- In a small set of varieties where the general imperfective marker differs from the form of the imperfective marker preserved on *h*-initial verbs, the regular imperfective marker is added back to synchronically opaque form: [IPFV-PRS-AGR] > [PRS.IPFV-AGR] > [IPFV-PRS.IPFV-AGR]. Table 9:

Table 9. Imperfective contrasts 'come'

	[PST.PFV.1SG]	[IPFV.PST.1SG]	[SBJ.PRS.1SG]	[IPFV.PRS.1SG]
CK (Mukrî)	*hat-im*	***de**-hat-im*	*b-ê-m*	***d**-ê-m*
CK (Germiyanî)	*hat-im*	***e**-hat-im*	*b-ê-m*	***dê**-m*
SK (Qorwe: > 35)	*hat-im*	***e**-hat-im*	*b-a-m*	***t**-ê-m*
SK (Sencabî)	*hat-im*	*hat-**ya**-m*	*b-a-m*	***tya**-m*
SK (Qorwe: < 35)	*hat-im*	***e**-hat-im*	*b-a-m*	***e-t**-ê-m*
B (Turkman)	*āt-un*	***=a k**-āt-un*	*by-ā-īn*	***=a k**-ā-īn*

There are several key points from *Table 8*. In the present tense, the most common way to express indicative/imperfective is by a prefix. The only exceptions here are the Hewramî (Gorani) variety of Text and the Southern Kurdish variety Sencewî, which features zero marking. According to Karim (2024), this is the result of regular sound changes: pretonic reduction *dekem > *d^{i}kem followed by cluster reduction *d^{i}kem > *kem*. Note that the same thing happens in the neighboring Hewramî varieties, e.g., G Text: *mekerű́ > m^{i}kerű́ > *kerű́*. However, since the same change affects the stressed subjunctive marker as well, both forms are rendered without prefixes. The result is that stress placement is the only marker of indicative/imperfective in the present tense (see Karim and Mohammadirad forthcoming). In the past tense, the general strategy for imperfective marking is also a prefix. However, there are varieties with suffixes like SK Sencewî, G Text, and G Zerde, as well as those with both prefixes and suffixes, such as the Laki variety of Hersîn.

There are several differences between what is observed in *Tables 8* and *9*. For past tense forms, there are essentially no differences except for in Balochi, where, in addition to the typical preposed enclitic marker *=a*, there is an additional marker *k-*. This *k-* formative only occurs with *h*-initial roots. In the present tense, the situation is a bit more interesting. Balochi shows the same pattern seen in the past tense, featuring the imperfective formative *k-* which only occurs in the imperfective, cf. the simple past tense *āt-un* 'I came' and present subjunctive *by-ā-īn* '(that) I come.' However, in Central Kurdish varieties, there is no pretonic reduction of vowels, as observed in Southern Kurdish. Central Kurdish varieties exhibit the regular contraction of the sequence *dehê- to *dê-*, characterized by the loss of intervocalic *h and vowel coalescence. In Southern Kurdish varieties, there was the pretonic reduction of the vowel of the indicative marker. In these varieties, the sequence *dhê- resulted in the devoicing of the *d *tê-*. This occurs in all varieties regardless of whether or not subsequent changes cause the original marker *de-* to change form on all other verbs.

Note that all varieties have what is essentially a suppletive imperfective stem with h-initial verbs beginning with *d-, t-,* or *k-*. There is one additional odd development in the Southern Kurdish variety of Qorwe, where the suppletion is beginning to be regularized. Speakers from younger generations have begun to extend the regular imperfective marker *e-* to the stems with the dental reflex (from the same etymological source, following Karim 2024), becoming *etim* 'I come' where older generations maintain the suppletive *tim*. Without knowing anything else about the system, it is clear that there is at least a superficial similarity between what is observed in Qorwe and Turkman Balochi: both have a regular imperfective marker, *=a* in Qorwe and *e-* in Turkman Balochi, respectively. There is a suppletive marker on a small subset of verbs *k-* and *t-*, respectively. These markers occur together in these varieties *=a k-* and *e-t-*. The Kurdish data make it clear that the combination *e-t-* is an innovative feature and not a reconstructable part of the etymon.[10] The development of an imperfective stem or even a present-tense stem with the remnants of the imperfective/indicative marker is a widespread feature in Iranic languages, e.g., 'Jowš. *a-pic-am* 'I twist,' *but a-t-ār-am* 'I bring'; Qohr. *a-k-ūn* 'I fall,' but *a-t-ār-ūn* 'I bring'; Tāri. *a-ker-ō* 'I do,' but *a-t-ār-ō* 'I bring'; Aby. *e-kar-ān* 'I do,' but *e-t-özmar-ān* 'I count'; Ards. *e-ker-ō* 'I do' but *e-t-oroš-ō* 'I sell.' See discussion in Windfuhr (1991, 249–50); Krahnke (1976, 182–87) also comes to a similar conclusion. Just as the move Southward in the Kurdish-speaking region, Stilo (2007b) shows that moving Eastward amongst the Provincial Dialects, the imperfective/indicative marker disappears in most contexts, e.g., '*biri, biri, bira* 'I, you, he takes/carries away,' or *der-k-i, der-k-i, der-k-a* 'I, you, he falls,' but the original *-t* still shows up as a remnant before vowels (in the durative tenses): (present) *tāri, tārém (= t-ār-i, t-ār-ém)* 'I, we bring' or *tosi (= t-os-i)* 'I get up' and imperfect *ši-t-ārt* 'he would bring.' These should be parsed as *a-t-ār-am* [IPFV-IPFV-bring.PRS-1SG.A] or even more accurately as *a-tār-am* [IPFV-bring.PRS.IPFV-1SG.A] with the *t* now part of a present imperfective stem. This analysis is additionally suggested by further examples that show the dental form becoming part of the present-tense stem regardless of aspect or mood, e.g., 'Qohrudi *tengas-*

[10] Central Kurdish varieties allow clitic person markers to intervene between the prefix and the verb, marking the direct object of a transitive verb. As such, the opportunity for the prefix and the h-initial stem to unimorphate was diminished in Central Kurdish transitive verbs. I refer to the increased likelihood for words and paradigm cells with high token frequency to be irregular. See Bybee (2003; 2002), Corbett et al. (2001) for a discussion of the role of type/token frequency in preserving suppletion. This is approximately parallel to the wat that the merger of the *k-* prefix with the stem did not occur in the context of negation where the proximity was blocked syntactically. Note that in Southern Kurdish, where no intervening clitic is possible, transitive and intransitive verbs alike show this suppletive imperfective stem, e.g., *hawirdin* 'to bring' Šėrwân Kıļâwây *tyar-*, Warmâwa *tyâr-*, and Zurbâtya, *d(ı)yârım*.

/tangašt vs. Meyme *enges-/angašt* 'look,' Soi *angis-* [pres.] ... Thus while the Qohrudi present tense is *a-tangisun* 'I look,' almost identical to Soi *a-t-angisom*, the Qohrudi subjunctive, *bátengisun* (*bá-tengis-un*) and the preterit *batangaštun* (= *ba-tangašt-un*) show that the initial *t-* is now part of the root (but not in Soi).' More recent data from Meym. shows the same developments observed in Qohrudi, e.g., *be=ş-ter-da* [SBJ=3SG-bring.PRS-2PL], where the *t* of *ter-* [bring.PRS-] is not etymological.

This pattern is pervasive and requires explanation. As it happens, Kurdish varieties are among the best-documented, despite also being understudied. Works such as Fattah (2000) and corpora like Matras et al. (2016), despite their flaws, allow for a considerable degree of certainty in commenting on the development of these formations. The findings from Karim (2024) are summarized in *Section 3.5.1* as a template for exploring the Balochi developments.

- The forms of the past imperfective likely started out as the inherited forms of the optative or similar construction covering habitual and irrealis functions when attached to the imperfective stem, and conditional functions when attached to the perfective stem.
- The shift from an aspectual split to a tense-based split caused the conditional form to be used for both imperfective and conditional functions.
- The perfective prefix *bi-*, later subjunctive, was then added to distinguish the conditional functions and innovative imperfective prefixes, e.g., *dar, *kar, *ham, *ham-ēw, to distinguish imperfective.
- The form of the past imperfective loses its suffixal form, remaking the imperfective on the basis of the past stem with imperfective prefixes alone.
- The form of the conditional is lost, its functions subsumed by the past imperfective.

6. Conclusion

Paul (2003) suggested that the Balochi imperfective prefix *=a/a-* is possibly related to the Central Kurdish imperfective marker *e-*. This statement has provided me with a set of changes based on which I could evaluate the Balochi formatives. My conclusion is that they are not possibly related. However, the proposal has failed in certain ways that point to the ultimate etymon.

Balochi is a Northwestern Iranic language related to Kurdish, Gilaki, Zazaki, Tati, Talyši, etc. Their nomadic history has led to their current location in the far southeast of the Iranic-speaking zone. However, there have been several waves of contact with and borrowing from other (North)western Iranic languages. These proposed borrowings include developments in the nominal system, e.g., the Balochi and Gilaki case system (following Korn 2019). However, the borrowings may include the developments in the verbal system described here, i.e., the imperfective marking from the *KAR construction of Vafaeian (2018) and the relationship between imperfective and conditional stems.

Linguistic data point to an original Balochi homeland in the far northwest of the Iranian world, and historical accounts point to modern-day Azerbaijan on the Caspian coast (see Elfenbein 1989 and Dashti 2012). Their migration followed in clockwise fashion, beginning in the late Sassanian period, just predating the Arab conquest. It was in the time immediately preceding this migration that the contact between Balochi and the Caspian languages could have been sufficient for the *KAR construction to spread between the languages: Gilaki, Tati, Talyši, and Balochi.

Because of this shared history and history of contact, it is not unreasonable to propose a contact-induced transfer between Caspian languages or even Kurdish at an early point in their development. I addressed these possibilities in *Section 3.5.2*. Another approach is that of mutual independent developments, a possibility that is strengthened by their genealogical closeness. A set of related languages may be, as I have informally referred to it, “cooking with the same ingredients”. In a recruitment cycle such as the progressive-to-imperfective cycle (Deo 2015), the initial stage is an optional progressive periphrasis that specifies an act-in-progress reading. It could exist alongside what Hopper (1991) refers to as layered forms: more than one gram to express the same function, e.g., English *will* vs *gonna* [FUT]. *DAR, *KAR, *HAM, and *STAN (not discussed here) progressives were likely present in the early ancestors of these varieties. Each of those types occurs in multiple regions across the Iranian world. Varieties of Tajik and Southern Kurdish had a *STAN progressive, the modern reflex of which cannot be confused with the *DAR, *KAR, and *HAM types (see Jeremiás 1993). Likewise, varieties with *HAM markers, New Persian, Laki, Gorani, etc., cannot be confused with the *DAR, *KAR, and *STAN types. Only what I propose are *DAR and *KAR types become superficially similar because of their close phonological shape. The base word order and any other sound changes that took place, e.g., consonant lenition, boundary reanalysis, hiatus resolution, etc., were in the DNA of all the groups.

I propose that the Balochi imperfective marker is the so-called “kar construction” discussed by Vafaeian (2018). There are a few issues that seem

to stand in the way of that proposal: (1) Multiple markers can stack in what seems to be the wrong order, e.g., nazzīk-a k-ayt [near-IPFV IPFV-come.PRS.3SG] (Axenov 2006, 247). However, this is not problematic, as can be observed in Kurdish forms that have reanalyzed the prefix as the imperfective stem, e.g., Qorwa *e-t-ê-m* [IPFV-IPFV-come.PRS-1SG]. Parsing this as two imperfective markers is diachronically accurate, but a better synchronic parsing is *e-tê-m* [IPFV-come.PRS.IPFV-1SG] with the productive *e-* prefix attaching to a suppletive imperfective stem. This is not unlike the English word *child-r-en*, which, from a diachronic perspective, features two plural markers [child-PL-PL]. Given the diversity of forms observed in Kurdish, it would be unsurprising to find some Balochi varieties where *=a* and *k-* could not cooccur (further research necessary).

(2) The Balochi imperfective marker is, at least in some dialects, a discontinuous morpheme compared to the prefix in Kurdish. Paul (2003) shows the imperfective marker as the prefix *a-* while Barker and Mengal (1969) and Korn (2005) have the preposed enclitic *=a*. (Nourzaei and Jahani 2012) show that the proclitic is more common in Balochi, with the enclitic becoming more common in Eastern varieties. However, this is not an issue as the Southern Kurdish varieties of Bisıtun, Cıhr, Harsin, Pâyrawand feature discontinuous imperfective marker *=a* alongside the prefix *ma-*.

The extant evidence for the origin of the Balochi construction is limited. Based on the *k-, ar-, a-,* and *=a* variants, it seems that a locative origin may not work for Balochi. Assuming that a *k* initial prefix is the ultimate root of both parts of the Balochi imperfective construction, some version of the "*kar* construction" found in the Caspian languages like Gilaki is a good possibility. There is no clear evidence for an etymological link between the Kurdish and Balochi constructions. However, it is an interesting coincidence that the Kurdish verbs that broadly bear the suppletive imperfective marker *t-* are cognate with a subset of the Balochi verbs that bear the suppletive imperfective marker *k-*. This is likely a phonological phenomenon and a coincidence. However, this coincidence could be the only clue to the ultimate origin of the construction *kar.

In Kurdish, the "Smoking Gun" pointing to a unified etymon for the imperfective prefixes *de-* and *e-* is the fact that they are incompatible on verbs that preserve the d formative in some environments, e.g., CK Kerkûk *e-ke-m* [IPFV-do.PRS-1SG] vs. *d-ê-m* [IPFV-come.PRS-1SG]. This is bolstered by the fact that the places where they do occur together are few and favored by younger speakers, e.g., SK Qorwe *e-t-ê-m* [IPFV-IPFV-come.PRS-1SG]. These are obviously the innovative forms. In Balochi, the system is the opposite. In nearly all known varieties, the affixes occur together, e.g., Turkmen Balochi *-e k-e-yt* [-IPFV IPFV-come.PRS-3SG], reflecting a longer time depth since the grammaticalization of these forms that matches what is currently understood

about Balochi migration. However, my search for a "Smoking Gun," a variety that doesn't allow both imperfective markers, was successful. Unfortunately, this was only a single speaker of Coastal Balochi from Korsar. That speaker uses the regular marker with all verbs except for the set with the *k-* prefix. It is unlikely that this is an idiosyncrasy of this speaker, as there is no reason to exclude a marker only with these verbs if they were there originally. However, a description of this variety, with data from multiple speakers, is necessary to confirm that it is systematic.

In a recent study of the t-form prefixes across Iranian languages, Aliyari Babolghani (2025) recognizes many of the developments described here: (1) the longtime depth of the related construction; (2) combinations with the old imperfective/optative forming the new past imperfective and conditional; (3) the Balochi forms *-a/a-* are likely not related and show no trace of dental elements; etc.[11]

The forms of the past conditional in Balochi and Kurdish exhibit a development already well-documented in the history of New Persian, as well as in Zazaki and Gorani. The past conditional is the form of the imperfective/irrealis with the addition of a perfective/subjunctive marker. However, the original form of the imperfective has been replaced synchronically by a productive imperfective marker and the regular unaugmented past-tense stem. There is no known relic of the use of the past-tense stem with the *-ēn* suffix for imperfective in Balochi. However, some relics show this cline of development in Kurdish, and the pattern is robustly attested across the Iranian World. I would not be surprised if a wide-ranging corpus study of Balochi varieties would reveal remnants of the older construction, as recently discovered in Southern Kurdish and possibly in Northern Kurdish by Bulut (2000). Obviously, further documentary work on Balochi dialects is necessary. Additionally, the theoretical motivations for these developments require clarification. The semantic motivation for a marker of imperfective combining with a perfective/subjunctive marker to specify a conditional is not clear. Additionally, it is unclear whether

[11] Aliyari Babolghani (2025)'s analysis differs from Karim (2024)'s only in the etymology of the t-form markers. Karim (2024) notes for places like Qorwe older forms like *tê* 'he comes' get regularized by younger speakers to *e-tê* featuring the productive imperfective marker next to the historical one. a Flattened description of the language gives the false impression that the eymon should have been *Vt not *tV. This led or misled Noorlander and Stilo (2015) to assume an original *at-, and the same logic has led Aliyari Babolghani (2025) to propose the preposition *ati as the ultimate etymon. I reject Aliyari Babolghani (2025)'s etymology based on phonological (*ati would render *hê* through regular sound changes) and semantic development issues ('through' > PROG is not a documented cline for emergent progressive periphrases). However, a long rebuttal of his analysis is beyond the scope of the current article. What is important to the current article.

morphological leveling alone is responsible for the formative losing its imperfective function in favor of the conditional.

BIBLIOGRAPHY

Aliyari Babolghani, S. (2025), "The Imperfect with the t-Type Prefix in New Iranian and Its Connection to the Old Iranian Augmented Imperfect Optative", *DABIR*, vol. 12, no. 1, pp. 1–39. https://doi.org/10.1163/29497833-20230031.

Axenov, S. (2006), *The Balochi Language of Turkmenistan, a Corpus-Based Grammatical Description* (Studia Iranica Upsaliensia 10), Uppsala: Acta Universitatis Upsaliensis.

Barker, M. A.-R., and Mengal, A. Kh. (1969), *A Course in Baluchi*, Montréal: McGill University's Institute of Islamic Studies.

Beekes, R. S. P. (2011), *Comparative Indo-European Linguistics*, 2nd ed., revised and corrected by M. de Vaan, Amsterdam/Philadelphia: John Benjamins.

Belelli, S. (2021), *The Laki Variety of Harsin: Grammar, Texts, Lexicon*, Bamberg Studies in Kurdish Linguistics (BSKL), vol. 2, ahead of print, Bamberg: University of Bamberg Press. https://doi.org/10.20378/irb-51703.

Bulut, Ch. (2000), "Indirectivity in Kurmanji", in: *Evidentials: Turkic, Iranian and Neighbouring Languages*, ed. by L. Johanson, and B. Utas, Berlin/Boston: De Gruyter Mouton, pp. 147-184. https://doi.org/10.1515/9783110805284.147.

Bybee, J. (2002), "Word Frequency and Context of Use in the Lexical Diffusion of Phonetically Conditioned Sound Change", *Language Variation and Change*, vol. 14, no. 3, pp. 261–290. https://doi.org/10.1017/S0954394502143018.

Bybee, J. (2003), "Mechanisms of Change in Grammaticization: The Role of Frequency", in: *The Handbook of Historical Linguistics*, ed. by B. D. Joseph, and R. D. Janda, Maiden/Oxford: Blackwell Publishing, pp. 602–623.

Bybee, J., Perkins, R., and Pagliuca, W. (1994), *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago/London: University of Chicago Press.

Campbell, L. (2013), *Historical Linguistics: An Introduction*, 3rd ed., Edinburgh: Edinburgh University Press.

Cheung, J. (2006), *Etymological Dictionary of the Iranian Verb*, vol. 2 (Leiden Indo-European Etymological Dictionary Series), Leidein: Brill.

Corbett, G. G., Hippisley, A., Dunstan, B., and Marriott, P. (2001), "Frequency, Regularity and the Paradigm: A Perspective from Russian on a Complex Relation", in: *Frequency and the Emergence of Linguistic*

Structure, ed. by J. Bybee and Paul J. Hopper, Amsterdam/Philadelphia: John Benjamins, pp. 201-226.

Dashti, N. (2012), *The Baloch and Balochistan: A Historical Account from the Beginning to the Fall of the Baloch State*. Bloomington, Indiana: Trafford Publishing.

Deo, A. (2015), "The Semantic and Pragmatic Underpinnings of Grammaticalization Paths: The Progressive to Imperfective Shift", *Semantics and Pragmatics*, vol. 8 (October), Article 14, pp. 1–52. https://doi.org/10.3765/sp.8.14.

Elfenbein, J. (1989), "BALUCHISTAN Iii. Baluchi Language and Literature" *Encyclopaedia Iranica*, vol 3, pp. 633-644.

Fattah, I. K. (2000), *Les Dialectes Kurdes Méridionaux: Étude Linguistique Et Dialectologique* (Acta Iranica 37), Leuven: Peeters.

Hayes, B. (1989), "Compensatory Lengthening in Moraic Phonology", *Linguistic Inquiry*, vol. 20, no. 2, pp. 253–306.

Hopper, P. J. (1991), "On Some Principles of Grammaticization", in: *Approaches to Grammaticalization: Vol. I. Theoretical and methodological issues* (Typological Studies in Language 19/1), ed. by E. C. Traugott, and B. Heine, Amsterdam/Philadelphia: John Benjamins, pp. 17-36.

Jeremiás, É. M. (1993), "On the Genesis of the Periphrastic Progressive in Iranian Languages", in: *Medioiranica: Proceedings of the International Colloquium Organized by the Katholieke Universiteit Leuven, 21st–23rd May 1990,* ed. W. Skalmowski, and A. Van Tongerloo, Leuven: Peeters, pp. 99–116.

Karim, Sh. O. (2021), *The Synchrony and Diachrony of New Western Iranian Nominal Morphosyntax*. Ohio State University Ph.D. dissertation. http://rave.ohiolink.edu/etdc/view?acc_num=osu1638313342357598.

Karim, Sh. O. (2024), "Can Insignificant Evidence Yield a Significant Result? Sub-Grouping Southern Kurdish Based on Imperfective Allomorphs" in *Current Issues in Kurdish Linguistics II*, ed. by A. Grond and S. Gündoğdu, Wien: Praesens Verlag, pp. 7-38.

Karim, Sh. O., and Mohammadirad M. (forthcoming), "Demorphologization and Remorphologization: The Development of a Progressive Prefix to a Marker of Negation in Hewramí", *Journal of Historical Linguistics*.

Korn, Agnes (2003), "Balochi and the Concept of Northwestern Iranian", in: *The Baloch and Their Neighbours. Ethnic and Linguistic Contact in Balochistan in Historical and Modern Times*, ed. by C. Jahani and A. Korn, in cooperation with G. Gren-Eklund, Wiesbaden: Reichert Verlag, pp. 49-60.

Korn, A. (2005), *Towards a Historical Grammar of Balochi: Studies in Balochi Historical Phonology and Vocabulary*, Wiesbaden: Reichert Verlag.

Korn, A. (2019), "Digging Through Layers of Language Contact: Elements of Diglossia and Multilingualism in Balochi", presented at the *2nd North American Conference on Iranian Linguistics (NACIL 2)*, University of Arizona.

Krahnke, K. J. (1976), *Linguistic Relationships in Central Iran.* University of Michigan. Ph.D. dissertation.

Kümmel, M. J. (2014), "The Development of Laryngeals in Indo-Iranian", Paper read at the conference *The Sound of Indo-European 3,* Opava.

Lambton, A. K. S. (1960), *Persian Grammar*. Reprinted with corrections. Students' ed., Cambridge: Cambridge University Press.

Mahmoudveysi, P., and Bailey, D. (2013), *The Gorani Language of Zarda, a Village of West Iran: Texts, Grammar, and Lexicon*, Wiesbaden: Reichert Verlag.

Matras, Y., et al. (2016), "The Dialects of Kurdish", in: *Web Resource, University of Manchester*. http://kurdish.humanities.manchester.ac.uk/.

Mohammadirad, M. (2020), *Pronominal Clitics in Western Iranian Languages: Description, Mapping, and Typological Implications*. Ph.D. thesis, Sorbonne Nouvelle – Paris 3 Ph.D. dissertation. https://tel.archives-ouvertes.fr/tel-02988008.

Mohammadirad, M. (forthcoming), *A Grammar of Hewramî*. Language Science Press, forthcoming.

Mohammadirad, M., and Karim, Sh. O. (2025), "The Development of Imperfective and Subjunctive Marking in Hewramî", *Linguistics*, vol 63, no. 5, pp. 1265-1292. https://doi.org/10.1515/ling-2023-0247.

Noorlander, P. M., and Stilo D. (2015), "On the Convergence of Verbal Systems of Aramaic and its Neighbours. Part I: Present Based Paradigms", in: *Neo-Aramaic and Its Linguistic Context*, ed. by Lidia Napiorkowska et al., Piscataway, New Jersey: Gorgias Press, pp. 426-452. https://doi.org/10.31826/9781463236489-026.

Nourzaei, M. (2017), *Participant Reference in Three Balochi Dialects: Male and Female Narrations of Folktales and Biographical Tales* (Studia Iranica Upsaliensia 31), Uppsala: Acta Universitatis Upsaliensis.

Nourzaei, M., and Jahani C. (2012), "The Distribution and Role of the Verb Clitic =a/a=in Different Balochi Dialects", *Orientalia Suecana*, vol. 61, pp. 170-186.

Nourzaei, M., Jahani C., Anonby E., and Ahangar, A. (2015), *Koroshi: A Corpus-Based Grammatical Description* (Studia Iranica Upsaliensia 13), Uppsala: Acta Universitatis Upsaliensis.

Pat-El, N. (2009), "The Development of the Semitic Definite Article: A Syntactic Approach", *Journal of Semitic Studies*, vol. 54, no. 1, pp. 19–50. https://doi.org/10.1093/jss/fgn039.

Paul, L. (2003), "The Position of Balochi Among the Western Iranian Languages: The Verbal System", in: *The Baloch and Their Neighbors: Ethnic and Linguistic Contact in Balochistan in Historical and Modern Times*,

ed. by C. Jahani, A. Korn, and P. Titus, Wiesbaden: Reichert Verlag, pp. 61-71.

Rubinčik, Jurij (2001), *Grammatika Sovremennogo Persidskogo Literaturnogo Jazyka* [Grammar of the Modern Persian Literary Language], Moscow: Izdatel′stvo vostočnoj literatury. In Russian.

Russell, P. (2017), "The Evolution of Celtic", in: *Handbook of Comparative and Historical Indo-European Linguistics*, Band 1, ed. by J. Klein, B. Joseph, and M. Fritz, Berlin/Boston: De Gruyter Mouton, pp. 1274-1297.

Skjærvø, P. O. (1997), "On the Middle Persian Imperfect", in *Syntaxe Des Langues Indo-Iraniennes Anciennes: Colloque International, Sitges (Barcelona), 4–5 Mai 1993* (Aula Orientalis Supplementa 6) ed. by E. Pirart, Barcelona: Editorial AUSA, pp. 161-188.

Skjærvø, P. O. (2009), "Middle West Iranian", in *The Iranian Languages*, ed. by G. Windfuhr, New York/London: Routledge, pp. 196-278.

Stilo, D. (2001), "Gilan x. Languages", *Encyclopædia Iranica*, vol. 10, fasc. 6, pp. 660–668.

Stilo, D. (2007a), "ISFAHAN Xix. JEWISH DIALECT", *Encyclopædia Iranica*, vol. 14, fasc. 1, pp. 77–84.

Stilo, D. (2007b), "Provincial Dialects", *Encyclopedia Iranica*, vol. 14, fasc. 1, pp. 93-112.

Stilo, D. (2018), "The Caspian Region and South Azerbaijan: Caspian and Tatic", in: *The Languages and Linguistics of Western Asia: An Areal Perspective*, ed. by G. Haig, and G. Khan, Berlin/Boston: De Gruyter Mouton, pp. 659-824. https://doi.org/10.1515/9783110421682-019.

Vafaeian, Gh. (2018), *Progressives in Use and Contact: A Descriptive, Areal and Typological Study with Special Focus on Selected Iranian Languages*. Stockholm University Ph.D. dissertation.

Wal Anonby, Ch. (2015), *A Grammar of Kumzari: A Mixed Perso-Arabian Language of Oman*, Ph.D. thesis, Leiden University.

West, B. (1980), *Gramática Popular Del Tucano*, Bogotá: Ministerio de Gobierno, Instituto Lingüístico de Verano.

Windfuhr, G. (1991), "Central Dialects", *Encyclopedia Iranica.*, vol 5, pp. 242-252.

Windfuhr, G. (2009), "Dialectology and Topics", in *The Iranian Languages*, ed. by G. Windfuhr, New York/London: Routledge, pp. 5-42.

Yoshida, Y. (2009), "Sogdian", in *The Iranian Languages*, ed. by G. Windfuhr, New York/London: Routledge, pp. 279-335.

YEREVAN STATE UNIVERSITY PUBLISHING HOUSE

JIL 2/2 (2025): 42-65

Journal of Iranian Linguistics

Reduplication in Lāri and Jibbāli: A Structural and Semantic Study

Muhammed Ourang*
The University of New South Wales
Khalsa Al-Aghbari
Sultan Qaboos University

doi.org/10.46991/jil/2025.02.02

Abstract: The focus of this study is two-fold. Firstly, it describes the most common types of reduplication in Lāri and Jibbāli. Secondly, it discusses their semantics. Reduplication is a morphological process whereby a base or part of it is repeated to express a meaning different from the base. The study explores the formation of 60 Lāri and Jibbāli reduplicative forms collected through a questionnaire distributed to 10 bilingual Lāri speakers and from 2 native Jibbāli speakers. Although Lāri and Jibbāli belong to distant language families, the study shows that both languages employ full reduplication to accentuate adjectives or nouns. Lāri partial reduplication express meanings as diverse as emphasis, intensity, categorization, attenuation and addition whereas Jibbāli uses reduplication to mark transitivity and to convey continuity in borrowed bi-consonantal Arabic verbs.

Keywords: Lāri, Jibbāli, total reduplication, partial reduplication, semantics, Iranian languages.

Muhammed Ourang
E-mail: m.ourang@unswalumni.com
ORCID: https://orcid.org/0000-0003-4270-5693
Khalsa Al-Aghbari
E-mail: khalsah@squ.edu.om
ORCID: https://orcid.org/0009-0004-2313-7939

Received: 26.06.2025
Revised: 30.11.2025
Accepted: 19.12.2025

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support.

1. Introduction

Reduplication is a productive word formation process cross-linguistically. According to Jin and Fang (2019: 420), reduplication is a "process that involves completely or partially repeating a stem". Broselow and McCarthy (1983) argue that reduplication is a particular type of affixational morphology

in which the affixes are phonologically unspecified and receive their full phonetic representation by repeating adjacent segments. Olsson (2015: 1) extends the definition to include the repetition of elements such as whole words, parts of words, stems or roots to achieve new grammatical or semantic functions. The repeated material in reduplication is conventionally known as 'the reduplicant'. Moreover, Urbanczyk (2007) states that the scope of what can be replicated, how reduplication collaborates with different morphemes and phonological processes, and such implications can be communicated through reduplication, all factor into making the investigation of reduplication a huge and rich area of linguistic examination.

Speaking cross-linguistically, Kauffman (2015: 22) states that reduplication is very common in Austronesian, South Asian, African languages, the Caucasus, Indo-European, Indonesian, Turkish and Arabic languages. He also argues that this process enhances, emphasises, amplifies, enlarges, diminishes, adds number or changes verb tense to bring about significant meaning changes or shades of meaning. Inkelas and Downing (2015: 502) have classified reduplication into various types depending on the unique features of the root, stem or word, which could be phonological or morpho-semantic. Reduplication can either be full or partial. Full reduplication is the repetition of the entire base (O'Grady & de Guzman 1997) while in partial reduplication, a portion of the base is repeated. Rubino (2005: 11) argues that partial reduplication comes in a variety of types that range from a simple consonant gemination or vowel lengthening to a copy of syllables. This paper is an attempt to address the existing types of reduplication in Lāri and Jibbāli and highlight their structures and meanings. Despite belonging to different language families (Southwestern Iranian and Semitic respectively), Lāri and Jibbāli coexist within the same broader linguistic area stretching from southern Iran to southern Oman. This geographical and sociolinguistic proximity makes them valuable for a cross-linguistic comparison. In addition, this study is a continuation of previous work (Al Aghbari & Ourang 2017) conducted on morphology of Jibbāli and Lāri. It is hoped that this topic of morphological studies will highlight differences in morphology between these two languages.

The paper is structured as follows: section §2 describes the literature review on Lāri and Jibbāli, section §3 will look into the studies on reduplication in these two languages, section §4 focuses on the data collection process and research methodology. In section §5, we discuss the most common reduplication patterns in Lāri including total (§5.1) and partial (§5.2) and then the semantic analysis of the forms will be presented in section §6. Section §7 constitutes the main part of the analysis of reduplication in Jibbāli, outlining the most common types of reduplication including full (§7.1) and

partial reduplication (§7.2). Finally, section §8 concludes the paper with some remarks.

2. Previous Research

This section will look at some of the previous studies conducted on Lāri (§2.1) and Jibbāli (§2.2) which is followed by a brief discussion of the research on reduplication in these languages (§3).

2.1. Lāri

Lāri belongs to the Southwestern (SW) branch of Iranian language family which is spoken mainly in three provinces of Fars, Hormozgan and Bushehr (Iran) and by Lāri speakers in diaspora in some Arabic-speaking countries such as United Arab Emirates (UAE), Kuwait and Qatar. Figure 1 below illustrates the region in Iran and other countries where Lāri language is spoken. It is estimated that approximately 150,000 people speak Lāri as their mother tongue though this issue needs more investigation. The Lāri language[1] has numerous varieties including Evazi, Xonji, Bastaki, Aheli, Gerāshi, Lāri, Buchiri, and so on which have some phonological, morphological or syntactical differences but they are mutually intelligible to speakers of different areas, thus they can smoothly communicate with each other.

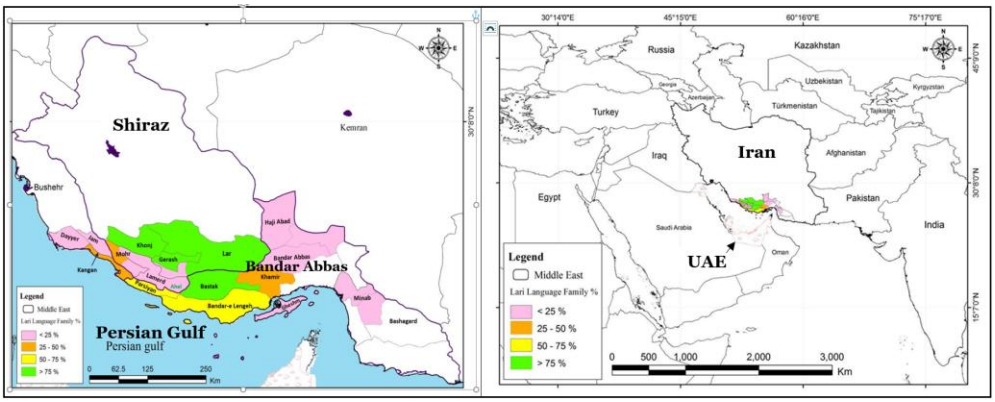


Figure 1. The areas where Lāri is spoken in Iran and other neighbouring countries (Ourang, 2022)

[1] It should be noted that the term "Lāri language" is used here as a cover term for all the dialects spoken in the regions shown in Figure 1. Accordingly, when reference is made to a specific dialect, terms such as 'Būchiri dialect', 'Bastaki dialect', 'Lāri dialect', or 'Aheli dialect' are used- each representing a member of 'the Lāri language family'.

Previous research on the Lāri language has explored various aspects, including etymology, noun and verb morphology, and syntax, significantly contributing to the study of the language and its dialects. Eqtedārī's works (1964, 1992, 2005) survey the vocabulary of seven Lāri dialects. Skjærvø (1989) compares Lāri (Lārestānī) with Kumzārī and Baškardī. Mahmoodian (2007), Vos̱ūqī (2001), and Xorram-Rūz (2011) explore the historical, cultural, geographical, and socioeconomic contexts of Lāri dialects. Moridi (2007, 2009, 2011) analysed the structure of nouns and adjectives in Lāri and proposed its classification. Xonjī (2009) examined the Xonji dialect in Fars Province. Salāmī (2004–2011) compiled data from various Lāri dialects through questionnaires, producing an eight-volume work with sample sentences, semantic categories, and Persian equivalents. In addition, Dabir-Moghaddam's (2013) two-volume *Typology of Iranian Languages* provided new insights into the syntax of Lāri and Xonji, detailing their verb and noun phrases as well as word order.

In addition to Iranian researchers, non-Iranian scholars have also contributed to the study of Lāri. Kamioka and Yamada (1979) compiled a list of one thousand basic Lāri words through questionnaires and interviews with native speakers. Later, Kamioka et al. (1986) collected and categorized nouns related to food, clothing, and kinship during fieldwork in Lar and Xonj. Similarly, Molčanova (1982/2000) provided a diachronic and detailed description of the Lāri dialect. So far, Ourang (2022) has presented the first reference grammar of the Aheli dialect of the Lāri language. The grammatical analysis is framed within descriptive linguistics and relies on extensive fieldwork interviews. The grammar offers an overview of the Aheli dialect, including a description of its sound system, morphology, and syntax, the data of which were collected via interviewing the Aheli speakers.

2.2. Jibbāli

Jibbāli is a Semitic language and one of the six Modern South Arabian languages. Modern South Arabian languages constitute a primary branch of West Semitic (Huehnergard & Rubin 2011) which occupy an eastern branch of Semitics (Rodgers 1991; Hetzron 1997; Faber 1997; Ratcliffe 1992 among a host of Semitists). Jibbāli, which is also called Shahri or Shahret, is spoken in the mountainous areas of Dhofar, a governorate in Oman. Johnstone identifies three dialectal varieties of Jibbāli on the basis of their geography in Dhofar: Eastern, Central and Western (Johnstone 1981: xii; Hofstede 1998:14). There is noticeable growth in the number of studies that explore the various aspects of Jibbāli's linguistics. Moreover, works that document

the morphological behaviour of the language are numerous, and some is written in languages other than English (Dufour 2016; Lonnet 2006) [2].

Matthews (1969) explores the nature of the deleted /m/ which results in a nasalized vowel in Jibbāli. He argues that the deleted /m/ before a vowel word initially and medially is a determiner in Jibbāli. Lesalu's (1945) research paper serves as a comparative study of the Semitic 'body parts' vocabulary. It reveals how Modern South Arabian languages express body parts using words different from those widely used in other Semitic languages. Johnstone (1980) argues that gemination which occurs in nouns marked for the definite article and with certain forms of the causative verbs is fairly a recent development in central and eastern dialects of Jibbāli. He observes that in both these sets, only certain sounds are geminated. He further claims that gemination as a morphological feature distinguishing meanings of words has long been lost from these dialects. Nakano's book (1986) is a semi-dictionary of Mehri, Socotri and Jibbāli in which the equivalents of some English words and expressions are listed for the three languages.

The latest work conducted on Jibbāli is a comparative cultural glossary of the six Modern South Arabian languages (Morris et al. 2019). The glossary comprises 345 terms that are of essential cultural significance in the region. The authors state that the glossary is meant to serve as a comparison of culturally significant terms across the six languages, and to reveal the nuances of certain cultural terms in each of them.

The following section presents a review of the existing literature on reduplication and reduplicated nominals in Semitic languages. To the best of our knowledge, no prior studies have specifically examined reduplication in Jibbāli. This paper therefore seeks to constitute the first systematic investigation to shed light on this understudied linguistic phenomenon.

Butts (2011: 84) argues that reduplicated nominal patterns are attested in Chadic, Cushitic, Egyptian, and Semitic. However, in Semitic, it is variously manifested in phonology, verbal morphology, nominal morphology and even syntax. However, the primary focus of his study was on reduplication that is found in nominal derivation. This type occurs when a noun "is morphologically derived by duplicating one or more root consonants". The range of semantics exhibited by the Semitic reduplicated nominal is wide as it covers diminutives, intensives, sound-symbolic terms, and bodily or personal characteristics, including defects and colour terms. Although there is no specific mention to the Jibbāli, Butts confirms that "The pattern

[2] For an extensive review of linguistic works done on Jibbāli, we refer the readers to Al Aghbari (2012).

*C1C2C3C2C3 occurs throughout West Semitic [which Jibbāli belongs to] but is absent from East Semitic" (p. 99).

In their study, Castagna and Al-Amri (2025) present under-described verbal classes in a variety of Jibbāli, more specifically the dialect spoken in the coastal town of Sadḥ. The study also presents a summary containing the relevant morpho-phonological phenomena involved in the formation of these verbal paradigms. The last study to be reviewed on Semitic reduplication is a PhD dissertation exploring bi-consonantal reduplication in Amharic. The study comprises a survey of the semantic categories that these reduplicative forms take, revealing that 'impairment of gait' and 'dressing up fancy' are common.

3. Research on Reduplication in Lāri

In recent years, there has been an increasing amount of literature on reduplication in Iranian languages (Gilaki by Fayyāżī 2013; Kurdish by Rasekh Mahand and Mohammadirad 2013; Persian by Vāḥedī Langarūdī and Yūsefīrād 2004; and Šeqāqī, 2000 inter alia), However, very few studies have examined the reduplication process in Lāri language and its dialects. Javadpour (2018: 147), in her Master's thesis, has studied the reduplication forms in Lārestani (as she calls the language) on the basis of Marantz's theory (1982). She has categorised the 210 reduplicated words into total proper reduplication and partial reduplication and concluded that former is more frequent in Lārestani language (over 50%). Javadpour has also found that "partial reduplicated forms contain a conjugation form" and only made 10% of the total reduplicated forms. Diyānat (2020) employed a descriptive-analytic and phonological framework, investigating a corpus of 114 examples collected from native speakers and the author's own linguistic intuition. The findings reveal that Lāri exhibits both total and partial reduplication, aligning with cross-linguistic patterns where reduplication serves grammatical and semantic functions such as emphasis, plurality, repetition, and intensification. The study highlights the role of phonological constraints in shaping reduplicative forms in Lāri and underscores the importance of documenting such processes in endangered languages. Diyānat's work will be referenced in the discussion and findings sections (§5.1-§5.2 and §6).

On the other hand, the authors are not aware of any previous works on the reduplication in Jibbāli. Therefore, this humble work can serve as introductory work to this intriguing linguistic phenomenon. In the following sections, we will analyse the data collection methods (§4), after which the structural analysis and semantic analysis of reduplication forms in Lāri (§5 and §6, respectively) and Jibbāli (§7) are presented.

4. Data collection and research methodology

The Lāri data were collected through a questionnaire distributed to ten Lāri subjects in the city of Lar, Iran. Respondents were males and females from a different age range; some of them were competent in Persian, the official language of Iran. One of the authors is a native speaker of Lāri. He used his own intuition to assess the collected data. He also transcribed the collected forms. The Jibbāli data were collected by two Jibbāli speakers enrolled in a linguistics undergraduate course. These speakers transcribed the forms and classified them based on their types and semantics. For the classification of the Lāri reduplication, the Leipzig Glossing Rules for inter-linear glosses. For the analysis of the Jibbāli reduplication using the CV type.

The following section will list the most common reduplicative forms in both Lāri and Jibbāli, with the discussion highlighting the Lāri reduplication followed by the Jibbāli reduplicative forms.

5. Structural analysis of reduplication

Reduplicative forms in Lāri, similar to other languages, consist of a base which belongs to one of the different grammatical categories (verb, noun, adjective, adverb, etc.), and a reduplicant which is a lexical category rather than a grammatical one. We divide the reduplication into total reduplication and partial reduplication, each of which explained below. Given that the primary focus of this paper is the examination of partial reduplication, a brief discussion of total reduplication will be included for context. Semantically, the forms can denote repetition, frequency, increase/decrease of quantity, quality and intensity emphasis, irregularity, reduction, dispersion, abundance, and so on. In the following sections, total reduplication (§5.1) and partial reduplication (§5.2) will be examined.

5.1. Total reduplication

As Javadpour (2018) mentions, in this type of reduplication, the base is repeated. The total reduplication can be categorised into total non-additive reduplication, which is made up of a reduplicant and a base only and total additive reduplication in which there is a grammatical morpheme at the end of the form (final) or in the middle of the form (medial). These types are explained in detail below with total non-additive reduplication (§5.1.1) and total additive reduplication (§5.1.2):

5.1.1. Total non-additive reduplication

This is where the entire root or word is duplicated, but the second part of the reduplicated form is not a simple repetition of the first[3]. It means it is altered in terms of phonological shifts like vowel or consonant changes. In the following example, the base *kæpæ* 'cracked' is added to the reduplicant *kæpæ* 'cracked' which forms *kæpæ-kæpæ* which denotes an intensity of the object, namely 'smashed'.

(1) kæpæ-kæpæ:

telefʊn=ʊm	*ʊftæ-Ø*	***kæpæ-kæpæ***	*bu-Ø*
mobile=1PC	fall-3SG	cracked-cracked	become.PST-3SG

'my mobile fell down and smashed' (lit. became cracked all over)

(2) petæku-petæku

guʃt-ija	***petæku-petæku***	*bʊ-vʊl-Ø*
meat-PL	small-small	IMP-cut.NPST-2SG

'cut the meat in very small [pieces]'

5.1.2. Total additive reduplication

In this kind of reduplication, a morpheme is added either at the end of the reduplicated form (final, §5.1.2.1) or in the middle of the base and the reduplicant (medial, §5.1.2.2):

5.1.2.1. Final added total reduplication

In this type of reduplication, a morpheme such as /i/, /ʊ/, /j/ are added to the end of the reduplicated form as shown in below where *j* 'ADJVS' is added to the end of the reduplicant as an adjectiviser morpheme:

(3) *kælæ-kælæ-**j***
hole-hole-ADJV

'holey, porous'

[3] Javadpour (2018) concludes that, in her sample of 210 forms, total non-additive reduplication constitutes the majority of reduplicated forms in the Lāri dialect, whereas medial-added total reduplication represents the smallest proportion of forms in the dialect.

In the following example, the suffix *æku* 'NMLS' is added to the final position of the reduplicative form to turn *pæl* 'wing/feather' into *pæl-pæl-æku* 'butterfly/earrings':

(4) *pæl-pæl-**æku***
wing-wing-NMLS
'butterfly/earrings'[4]

5.1.2.2. Medial added total reduplication

When an affix such as *væ* or *æ* intercedes the base and the reduplicant, the resulting form is called medial added total reduplication. In the example below, the infix *væ* intervenes between the base *vela* 'time' and reduplicant vela 'time and makes the resulting form as shown below:

(5) *vela-**væ**-vela*
time-INF-time
'time after time, again and again'

The infix *a* can also intercede between the base and the reduplicant as shown below:

(6) *gæp-^{a}-gæp*
wide-INF-wide
'throughout' (lit. for all width)

5.2. Partial reduplication

In most of the partial reduplication forms, there is a medial morpheme which comes between the two reduplication elements (base and reduplicant). Furthermore, the reduplicant differs from the base in both phonological and morphological features. The reduplicant is meaningless in some cases as shown in (11) and (12) below. In this type, /ʊ/ intervenes between the base and its reduplicant. So, partial reduplication is discontinuous in Lāri since a segment is inserted between the reduplicant and base[5]. In Lāri, the formation of most partial reduplication types is not simple[6], rather it is complex. As

[4] Javadpour (2018: 84-85) has mentioned only 'butterfly' as the meaning for the reduplicative form in the Lāri dialect.

[5] Other Iranian languages such as Kurdish, which belongs to the Western branch of the Iranian language family, uses infix /u/ and /e/ between the base and the reduplicant in full reduplication forms (*kæt-ʊ-per* 'torn apart', *deng-e deng* 'full of air'). However, no infix is reported between the elements of partial reduplication forms (Mirmokri & Seifori 2016: 168-169).

[6] Rubino (2005: 15) defines 'simple reduplication' as the one in which the reduplicant matches the base from which it is copied without phoneme changes or additions.

Rubino (2005: 15) defines, "in complex reduplications there is some different phonological material, such as a vowel or consonant change or addition". Reduplicative forms in Lāri can fall under a noun (*xar-ʊ-xæʃar* 'woods'), an adjective (*tak-ʊ-tela* 'sparkling clean') or rarely an adverb (*zeʃt-ʊ-ziba* 'disgracefully'). We can divide the partial reduplication into prefixal and suffixal, in the former of which the reduplicant comes before the base while the in the latter the reduplicant comes after the base. We have specified these types in front of each example. It should be mentioned that majority of forms which are partially reduplicated belong to the suffixal category[7]. Semantically, most of partial reduplication forms have been lexicalised in the Lāri language (Javadpour 2018: 108-109). The existing types of partial reduplications with representative examples in Lāri are shown below:

a) Noun + -ʊ- + Noun → Noun (suffixal)

(7) *xar-ʊ-xæʃar* **(suffixal)**
thorn-and-firewood
'woods'

(8) *tir-ʊ-tærkæ* **(suffixal)**
timber-and-stick
'timbers'

(9) bæl-ʊ-ʃæl **(suffixal)**[8]
sand-and-mud
'soil'

(10) *sʊk-ʊ-pʊk* **(suffixal)**
spur-and-REDUP
'spur; goad (verb: to goad someone into an action)

In the examples above, both the base and the reduplicant are meaningful and can semantically stand alone. Other examples include *dur-ʊ-dærvazæ* (door-and-gateway) 'gate' and *ʤær-ʊ-dæ:va* (argument-and-fighting) 'fighting'.

[7] Javadpour (2018) contends that partial reduplication forms cannot be classified as 'prefixal' or 'suffixal' due to the presence of an intermediary morpheme, identified as a 'coordinating conjunction morpheme,' which separates the base from the reduplicant. As a result, the reduplicant does not function as an affix. Consequently, she proposes referring to these forms as 'rhyming compounds/reduplicated forms'.

[8] Naseri and Razmdideh (2022), in their study on Ghayeni, refer to forms in which the initial consonant of the reduplicant differs from that of the base as 'echoic reduplication.' An example of echoic reduplication in Lāri is sʊk-ʊ-pʊk ('goad') as illustrated in example (10).

However, there are cases where the reduplicant is meaningless and the resultant reduplication is semantically dependant on the meaning of the base:

(11) *beʃ-ʊ-bax* **(suffixal)**
child-and-REDUP
'children'

(12) *del-ʊ-dʊm* **(suffixal)**
belly-and-REDUP
'abdominal pain'

Reduplication in Lāri can either be pre- or post-reduplication. In the former type, the reduplicant precedes the base while in the latter one the reduplicant occurs after the bases. The reduplicant cannot occur in medial position in this language. Examples (11) and (12) show post-reduplication, while example (13) shows a pre-reduplication form:

b) Noun + -ʊ- + Adjective → Adjective

(13) *tak-ʊ-tela* **(prefixal)**
REDUP-and-gold
'sparkling clean'

c) Noun + -ʊ + Preposition → Noun

(14) *ɣæp-ʊ-dəʊr* [9] **(suffixal)**
left-and-around
'surrounding'

d) Adjective + -ʊ + Adjective + (NOM) → Noun

In this formula, the reduplicant is made of an adjective and probably a nominaliser as shown below:

(15) *pak-ʊ-pelæʃt-i* **(suffixal)**
clean-and-dirty-NOM
'cleaning'

(16) *næk-ʊ-bæz* **(suffixal)**
good-and-bad

[9] Ghaniabadi et al. (2006: 5) calls "imitative co-compounds" in which the second member is meaningless and phonologically similar to the first but not in a way that is derivable by rule.

'arrangement'

e) Adjective + -ʊ + Adjective → Adjective

(17) *zeʃt-ʊ-ziba* **(suffixal)**
ugly-and-pretty
'disgraceful'

(18) *gæt-ʊ-gʊt* **(prefixal)**
REDUP-and-big
'very big; enormous'

(19) *særd-ʊ-perd* **(suffixal)**
cold-and-REDUP
'very cold' (referring to the food temperature)

As seen in (15)-(17), the semantic relationship between the elements inside the reduplication is antonymous or oppositional. For example, in (17), *zeʃt* means 'ugly' while *ziba* means 'pretty'. This will be explained in detail below. There are other structures in which the initial consonant of the base is repeated and the rest of the reduplicant is unpredictable:

f) C_1y_1+ -ʊ + C_1y_2 → Noun

(20) *per-ʊ-pak* **(prefixal)**
REDUP-and-clean
'pruning'

Another productive reduplicative type in Lāri follows the type C_1*æk-ʊ-*C_1y_1 as in *bæk-ʊ-beʃ* 'child and other people like them'[10]. It should be noted that the type adds the meaning of 'and related stuff/people' to the base. Šeqāqī (2000) argues that the concept of categorisation is inferable form these forms. Phonologically, the base and the reduplicant share mainly the first phoneme (*per-ʊ-pak* 'pruning'). Some examples are given below:

g) C_1æk + -ʊ + C_1y_1 → Noun

(21) *bæk-ʊ-beʃ* **(prefixal)**
REDUP-and-child
'child and people like them'

[10] This is similar to what is known as *shm*-reduplication which is lent to English from Yiddish Jews from NY. (Kauffman 2015: 1) argues that *fancy-shmancy* is another reduplication from Yiddish. In Lāri, there are some cases which denote the meaning of playfulness, *axtung-paxtung* as a reduplication form for the unknown language.

(22) *nækʊ-nu* **(prefixal)**
REDUP-and-bread

'bread and other grains'

(23) *ʤækʊ-kʤaɪl-ija* **(prefixal)**
REDUP-and-young-PL

'youths and other people in the same range of age'

(24) *dækʊ-dʊrəʊ* **(prefixal)**
REDUP-and-untruth

'lies and other untruth statements'

When the initial consonant of the base is /k/ as in *kʊrʊna* in (25), then /t/ replaces /k/ in the reduplicant:

(25) *kætʊ-kʊrʊna* **(prefixal)**
REDUP-and-Corona_virus

'Corona (virus) and relevant transmittable diseases'

Kauffman (2015, p. 5) argues that languages such as Malay employ reduplication forms to express diversity and collectivity (e.g., *sayur* 'vegetable' → *sayur-sayur* 'various vegetables'). As illustrated above, the type C_1*æk-ʊ-*C_1y_1 can be extended to any other new words lent to Lāri. It expresses the meaning of diversity.

h) C_1y_1 + C_1y_2 → Noun

(26) *xæssæ-xʊni* **(prefixal)**
tired-REDUP

'break'

In some forms such as (26), the initial consonant of the base is repeated and the reduplicant is not predictable. There is not any infix between the base and the reduplicant. Therefore, *xæssa-xʊni* 'break' is a post-reduplication without the infix, and the reduplicant *xʊni* is not predictable.

In some partial reduplication types, the grammatical category of the base does not change in the output (27) whereas in some cases the reduplicative form has a different category compared to the base (28):

(27) *bækʊ-bæʃkar* **(prefixal)**
REDUP-and-tillage

'tillage'

(28) *tak-ʊ-tela*
REDUP-and-gold **(prefixal)**
'sparkling clean'

As seen, both the base *bæ/kar* and reduplication form *bæk-ʊ-bæ/kar* in (27) belong to the category of nouns meaning 'tillage'. In (28), the base *tela* is a noun meaning 'gold' while the resultant form *tak-ʊ-tela* is an adjective meaning 'sparkling clean'[11]. Reduplication can occur in the categories of nouns, verbs, adjectives and adverbs (as shown in the examples above) but the prepositions cannot function as the base in reduplicative forms.

Similar to other Iranian languages such as Persian (Maḥmūdī Baxtīyārī and Ẕolfaqār Kondorī 2015), reduplication forms are not used to form plurals in Lāri[12]. The reduplication forms are neither used to express verbal features (such as tenses, aspect, in/transitivity or causative) nor they are employed to create a comparative adjective[13]. The reduplication form, depending on the category, can be marked for different features of the class. For example, the form *næk-ʊ-bæz* (**good-and-bad**) 'arrangement' which is a noun can take plural marker *-ija*, definiteness marker *-əʊ* and be modified by an adjective, a demonstrative pronoun or appear in relative clauses. Below, we show a reduplicative form in its plural form (29) and in the context (30) for the Lāri language:

(29) *næk-ʊ-bæz-ija* **(suffixal)**
good-and-bad-PL
'arrangements'

(30) *æz* *sʊb* *ta* *ʊla* *næk-ʊ-bæz-ija* *æ-kerdæz-ʊm*
from morning till now good-and-bad-PL IPFV-do.NPST-1SG
'I am organising the arrangements from morning up to now'

As seen in (30), the reduplicative form *næk-ʊ-bæz* 'arrangement' as a non-verbal element together with the verb *kerdæ* 'to do' form a compound verb. The verb *kerdæ* 'to do' in Lāri is preceded by nouns, adjectives or prepositions to form compound verbs. Most of the reduplicative forms can participate in

[11] The change of category is similar to other Iranian languages like Persian (Šeqāqī 2000).

[12] Kauffman (2015: 1) mentioned languages such as Malay where reduplication is used to make plurals as in the example *bunga* 'flower → *bunga-bunga* 'flowers.'

[13] In Lāri, full reduplication is used to derive temporal adverbials: *zuz* 'early'→ *zuz-a-zuz* 'quickly'; *gærm* 'warm'→ *gærm-a-gærm* 'warmly, but partial reduplication does not have this feature in the language.

the formation of compound verbs, the meaning of some is non-compositional (i.e. idiomatic):

(31) *xæssæ-xʊni* *æ* *dur* *kerdæ*
tired-REDUP to outside to_do
'to have a break' (lit. to take the tiredness outside [of the body])

Reduplication in nouns make up the largest proportion of the corpus in Lāri, whereas adverbs comprise the lowest percentage of the corpus. The only example of reduplication which functions as an adverb is *zeʃt-ʊ-ziba* 'disgracefully' which is illustrated in formula e) of example (17) above. Adjectives comprise the second most frequent class of reduplication forms. Rasekh Mahand (2009) reported the same results on the frequency of nominal reduplications in Persian.

6. Semantic analysis

Semantically, reduplication can be divided into three categories depending on the component which is meaningful:

a) Both base and reduplicant are meaningful:

(32) *xar-ʊ-xæʃar*
thorn-and-firewood
'woods'

As shown above, the base and the reduplicant can semantically stand alone. The meaning of the reduplication is compositional and is a subcategory of the reduplication. For example, in (32) above, the meaning of *xar-ʊ-xæʃar* encompasses the meaning of both *xar* 'thorn' and *xæʃar* 'firewoods'. Similarly, *tir-ʊ-tærkæ* 'timbers' includes the meaning of *tir* 'timber' and *tærkæ* 'stick'. Nonetheless, there are some cases where the parts of reduplication have an independent meaning, but they have different meaning if they are used individually. Consider the following example:

(33) *del-ʊ-dʊm*
heart-and-tail
'abdominal pain'

As seen, the meaning of the reduplicative form 'abdominal pain' is different from the meaning of its constituents: *del* 'heart' and *dʊm* 'tail'.

b) The base is meaningful but the reduplicant is meaningless:

(34) *ʊlʊs-ʊ-tæmbæl*
REDUP-and-lazy
'lazy, sluggish'

This example shows a pre-reduplication where the reduplicant occurs before the base. The reduplicant can appear after the meaningful base (post-reduplicant) as in *xæssæ-xʊni* 'break' which is analysed in example (26) above. However, there is no reduplication form in which both base and reduplicant are meaningless in Lāri.

Furthermore, when the reduplicant is meaningless as in (34) above, the base determines the meaning of the reduplicative form which encompasses a range of meanings of intensity[14] (*tak-ʊ-tela* 'sparkling clean'), similarity (*tir-ʊ-tærkæ* 'timbers and other similar materials'), categorisation (*fæk-ʊ-færmu* 'instructions and relevant commands'), emphasis (*læ:m-ʊ-xæʃ* 'very soft'), and attenuation[15] (*asækəʊ-asækəʊ* 'slowly').

The meaning of the resultant reduplicative form varies from iconic (*ʊlʊs-ʊ-tæmbæl* 'lazy') to the potentially idiomatic/ironic meanings (*zeʃt-ʊ-ziba* 'disgracefully'). Thus, the meaning of the reduplication is sometimes compositional as it can be derived from the meanings of the inputs involved or idiomatic/metaphorical, which is unpredictable by the elements of the base and the reduplicant. The idiomatic concept of the output can be typically ascribed to the idiomaticity of the base. For example, in the example below, *zæ:ræ* 'valour' [16] determines the metaphoric meaning of the reduplication form:

(35) *zæk-ʊ-zæ:ræ*
Redup-and-valour
'valour'

Khanjan and Alinezhad (2010), analysing Persian, argue that the idiomatic/metaphorical semantics of reduplicative forms results from the semantic extension of the base and its reduplicant involved. In some cases, the resulting output is semantically related to one of the elements. This finding is in line with what we stated about the Lāri examples (34) and (35) where the base defines the meaning of the output. However, the meaning of

[14] Or enhancement/amplification (Kauffman 2015: 4). Typically, full reduplication forms express intensity in Lāri as in *xæl-xæl* 'very hot' or *zuz-a-zuz* 'quickly'.
[15] Typically, in full reduplication forms.
[16] *Zæ:ræ* 'valour' originally means 'yellow bile', which is secreted from the gall bladder. Traditionally, Lāri speakers thought that fear caused rupture in the gall bladder and lead to secretion of the yellow bile. In this case, people used to say that the person has 'a rupture in his gall bladder'.

the reduplication is not always predictable from the meaning of its constituents as shown in (36).

(36) *ze∫t-ʊ-ziba*
ugly-and-pretty
'disgracefully'

Ghaniabadi et al. (2006) argue that the meaning of reduplication does not have to be attributed to the reduplicant but can be posited as a property of the construction as a whole. This applies to the reduplicative forms such as *C1æk-ʊ-C1y1* as in *ʧæk-ʊ-ʧe:run* 'furnace and other ovens' or *ʧæk-ʊ-ʧaji* 'tea and other beverages.'

As illustrated in (15)-(17), there are cases where the reduplicative elements have antonymous relationship to each other and connect to each other. Another example is *pak-ʊ-pelæ∫ti* (**clean-and-dirty-NOM**) 'cleaning' in which *pak* 'clean' and *pelæ∫t* 'dirt' juxtapose each other using /ʊ/ as the connecting element.

Investigating reduplication forms in Kurdish, Mirmokri & Seifori (2016, p. 175) made the following observations about the base and the reduplicant which have their own independent meaning:

a) If the parts of the reduplication are synonymous, the reduplication shows increase or emphasis. In Lāri, forms such as *læ:m-ʊ-xæ∫* (**soft-and-fine**) 'very soft' indicate an emphasis in the meaning.

b) If the parts of the reduplication have contrastive meanings, they refer to a general concept. This relationship is called 'uncoordinated relationship' and the meaning of the resultant reduplication is attached to one of the parts (either the base or the reduplicant) though they are opposite in meaning. These reduplication forms are called 'incompatible reduplications.' In Lāri, we can give an example of *ze∫t-ʊ-ziba* (**ugly-and-pretty**) 'disgracefully'.

7. Jibbāli reduplication

In what follows, we present two types of reduplication in Jibbāli: Full reduplication (§7.1) and partial reduplication (§7.2).

7.1. Full reduplication

Full reduplication exists in Jibbāli, the Dhofari language in Oman. It is used to intensify the meaning expressed by the adjectives; for example, instead of

saying ‘beautiful’, the meaning after full reduplication will be ‘very beautiful’. To illustrate, ‘ɣabz’ means ‘boring’, and when people fully replicate the form ‘ɣabz ɣabz’, the meaning changes to ‘very boring’. Full reduplication in Jibbāli is not limited to adjectives but can extend to apply to other lexical categories. In the following Table 1, we list several examples of full reduplication. It is worth noting that a few forms in the table belong to Dhofari Arabic which is often code-switched by the two informants of the study.

Table 1. Full reduplication in Jibbāli

Base	**Gloss**	**Reduplicative forms**	**Gloss**
ɣabz	‘boring’	ɣabz ɣabz	‘very boring’
ħiluu	‘beautiful’	ħiluu ħiluu	‘very beautiful’
ħmaas	‘exciting’	ħmaas ħmaas	‘very exciting’
kiʃxah	‘stylish’	kiʃxah kiʃxah	‘very stylish’
murhiq	‘exhausting’	murhiq murhiq	‘very exhausting’
haadii	‘quite’	haadii haadii	‘very quite’
mʕasˁsˁab	‘angry’	mʕasˁsˁab mʕasˁsˁab	‘very angry’
tˁwiil	‘tall’	tˁwiil tˁwiil	‘very tall’
qsˁiir	‘short’	qsˁiir qsˁiir	‘very short’
frħaan	‘happy’	frħaan frħaan	‘very happy’

Another type of reduplication in Jibbāli involves full reduplication of verbs which are borrowed from the Arabic language to express the ‘continuity or progression of the action’ expressed by the verb. For example, ‘*Ahmed haz atˁaawlah*’ means that ‘he shook the table’, while ‘*Ahmed hazhaz atˁaawlah*’ means that he was shaking the table for some time. In Table 2, we present several examples of this type of reduplication.

Table 2. Two-consonant verbs’ reduplication in Jibbāli

Base	**Gloss**	**Reduplicative forms**	**Gloss**
haz	‘shook’	hazhaz	‘shaking’
far	‘escaped’	farfar	‘escaping’
fasˁ	‘pressed’	fasˁfasˁ	‘pressing’
gar	‘dragged’	gargar	‘dragging’
Kab	‘poured’	kabkab	‘pouring’
sˁab	‘poured’	sˁabsˁab	‘pouring’
ʕaðˁ	‘bit’	ʕaðˁʕaðˁ	‘bitting’
ʃam	‘smelled’	ʃamʃam	‘smilling’
ʕad	‘counted’	ʕadʕad	‘counting’
raʃ	‘sprayed’	raʃraʃ	‘spraying’

7.2. Partial reduplication

The third type of reduplication in the Dhofari dialect is partial reduplication. In this process, the last consonant is reduplicated in order to make the verb a transitive verb. For example, '*Ahmed marra*' means he walked or passed in that path, while '*Ahmed marrar ʔlʔɯraaq lilt^{ʕ}ullaab*' means he passed the papers to the students. In Table 3 below, we list various examples of partial reduplication in transitive verbs. It is noteworthy to state that the base verbs can be used both as intransitive verbs in most contexts and as transitive verbs in other contexts, while the verbs after reduplication can be used only as transitive verbs.

Table 3. Partial reduplication in Jibbāli

Base	Gloss	Reduplicative forms	Gloss
ħassa	'feel'	ħassas	'make someone feel a feeling'
madda	'stretch'	maddad	'supply something to someone or stretch something'
fakka	'release'	fakkak	'turn something into separate pieces'
ʃamma	'smell'	ʃammam	'let someone smell something'
massa	'touch'	massas	'let someone touch something'
ɡarra	'drag'	ɡarrar	'let something drag something else'
ħabba	'love'	ħabbab	'make someone love something or someone else'
laffa	'change his way'	laffaf	'change the way of something or someone else'
hazza	'shake'	hazzaz	'shake someone or something else'
marra	'pass'	marrar	'let someone pass or pass something'

8. Conclusion

Reduplication is a common morphological process in both Lāri and Jibbāli. To illustrate, in Lāri, it is used productively to add meanings like emphasis (*pæræ-pet* 'torn away'), categorisation (*mæk-ʊ-me:mu* 'guests and other people') or addition (*ʤær-ʊ-dæ:va* 'fighting') in the language. There are some cases where the reduplication form can add a playfulness meaning as in *zæk-ʊ-zum* 'Zoom and other [ridiculous!] applications like it'. As Kauffman (2015: 1) mentions, reduplication is a "form of seasoning that salts and peppers the language". In Jibbāli, reduplication is also widely used. This is illustrated by the use of reduplication in bases that belong to different lexical categories listed in the paper. These are adjectives to express intensification or the

meaning of 'very', verbs to express the meaning of continuity and transitivity. The latter meaning is mainly used for borrowed verbs from Arabic.

In Lāri, reduplication forms are mainly nouns (*bæl-ʊ-ʃæl* 'soil') and adjectives (*tak-ʊ-tela* 'sparkling clean') but there are some cases of reduplication functioning as adverbs (*zeʃt-ʊ-ziba* 'disgracefully'). Reduplication forms can be made by combining nouns, adjectives or prepositions. There are also other formulae in which the reduplicant is made by the first consonant of the base and another variant part: *gæt-ʊ-gena* (**REDUP-and-mad**) 'a mad person'.

One of the reduplication forms which is very productive in Lāri is C_1*æk-ʊ-*C_1y_1 which is employed to localise foreign words as exemplified by *væk-ʊ-vʊrd* 'Microsoft Word and similar software products.' Moreover, the output is not always grammatically predictable on the basis of the constituents, and there is a change in grammatical category of reduplication in some cases as in *tela* 'gold' which is a noun but *tak-ʊ-tela* (**REDUP-and-gold**) 'sparkling clean' is an adjective.

Semantically, the base is typically meaningful while the reduplicant cannot stand alone. However, there are cases with both components meaningful. The meaning of the outcome can be iconic (*bæk-ʊ-balæk* 'alcove') or idiomatic (*zæk-ʊ-zæ:ræ* 'valour'). The relationship between the base and the reduplicant can be either synonymous or contrastive which indicate an emphasis or general meaning respectively.

There is a clear similarity between Lāri and Jibbāli in terms of the availability of full reduplication to emphasize adjectives or nouns. However, full reduplication is mainly used to emphasize adjectives in the Dhofari language. Intriguingly, Jibbāli uses reduplication to mark transitivity in verbs. Intriguingly, Jibbāli uses reduplication to mark transitivity in borrowed verbs from Arabic. It also replicates bi-consonantal verbs borrowed from Arabic to express the continuity or progression of the action.

BIBLIOGRAPHY

Al Aghbari, Kh. & Ourang M. (2017), "Description of Number, Person and Tense Features in the Verbal Morphologies of Jibbāli and Lari", *Journal of Arts and Social Sciences (JASS)*, vol. 8, no. 2, pp. 5-12.

Broselow, E, and McCarthy, J. J. (1983), "A theory of internal reduplication", *The Linguistic Review,* vol. 3, no. 1, pp. 25-88.

Butts, A. M. (2011), "Reduplicated Nominal Patterns in Semitic", *Journal of the American Oriental Society*, vol. 131, no. 1, pp. 83–108.

Castagna, G., & Al-Amri, S. (2025), "The morphology of some under-described verbal classes in the eastern Jibbali/Shehret variety of Sadḥ (Seṯḥ)", *Journal of Semitic Studies,* vol. 70, no. 1, pp. 403-430.

Diyānat, L. (2020), "Do-gān-sāzī dar zabān-e Lārī" [Reduplication in the Lārī language], *Majalle-ye zabān-šenāsī va gūyeš-hā-ye Īrānī* [Journal of Linguistics and Iranian Languages], vol. 5, no. 1, pp. 83-105. In Persian.

Dufour, J. (2016), *Recherches sur le verbe subarabique modern*. Habilitation sous la direction de M. Gilles Authier, EPHE. Ecole pratiqe des hautes etudes.

Eqtedārī, A. (1964), "Lahje-ye Fīšvarī" [The dialect of Fishvar], *Farhang-e Īrān-zamīn*, no. 11, pp. 71–92. In Persian.

Eqtedārī, A. (1992), *Lārestān-e kohan va farhang-e Lārestanī* [Old Lārestān, and Lārestānī culture], Tehran: Jahān-e Moʿaṣer. In Persian.

Eqtedārī, A. (2005), *Zabān-e Lārestānī: jostār-ī dar zabān va farhang-e mardom* [The Lārestānī language: Research in people language and culture], Tehran: Hamsāye. In Persian.

Faber, A. (1997), "Genetic Sub-Grouping of the Semitic Languages", in: *The Semitic Languages*, ed. R. Hetzron, London: Routledge, pp. 3-15.

Fayyāżī, M. S. (2013), "Tekrār dar guyeš-e Gīlakī" [Reduplication in Gilaki dialect], *Pažūheš-hā-ye zabān-šenāsī-ye taṭbīqī* [Journal of Comparative Linguistic Researches], vol. 3, no. 6, pp. 135–159. In Persian.

Ghaniabadi, S., Ghomeshi, J., and Sadat-Tehrani, N. (2006), "Reduplication in Persian: A morphological doubling approach", in: *Proceedings of the 2006 Annual Conference of the Canadian Linguistics Association*, ed. C. Gurski, and M. Radišić, Canadian Linguistic Association, pp. 1–15.

Hofstede, A. (1998), *Syntax of Jibbali*, Ph.D. dissertation, University of Manchester.

Huehnergard, J., and Rubin, A. D. (2011), "Phyla and waves: Models of classification of the semitic languages", in: *The Semitic Languages: An International Handbook*, ed. S. Weninger; in collaboration with G. Khan, M. P. Streck, and J.C.E. Watson, Berlin/Boston: De Gruyter Mouton, pp. 259-278.

Inkelas, S., and Downing, L. J. (2015), "What is reduplication? Typology and analysis part 1/2: The typology of reduplication", *Language and Linguistics Compass,* vol. 9, no. 12, pp. 502–515.

Javādpūr, F. A. (2018), "Barrasī-ye Farāyand-e Tekrār dar Zabān-e Lārestānī" [*Reduplication in Larestani language*] [Unpublished master's thesis], Allameh Tabataba'i University. In Persian.

Jin, J., and Fang, Z. (2019), "A comparative study on reduplication in English and Chinese", in: *Proceedings of the 2019 5th International Conference on Humanities and Social Science Research* (Series: Advances in Social Science, Education and Humanities Research, vol. 319), ed. X. Du, Ch. Huang, and Y. Zhong, Paris: Atlantis Press, pp. 420–424.

Johnstone, Th. (1980), "Gemination in the Jibbali Language of Dhofar", *Journal of Arabic Linguistics,* vol. 4, pp. 61-71.

Johnstone, Th. (1981), *Jibbāli lexicon*, London: Oxford University Press.

Kamioka, K., and Yamada, M. (1979), *Lārestāni studies 1: Lāri basic vocabulary* (Studia Culturae Islamicae, no. 10), Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

Kamioka, K., Rahbar, A., and Hamidi A. A. (1986), *Comparative basic vocabulary of Khonji and Lāri. Lārestāni studies 2* (Studia Culturae Islamicae, no. 30), Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

Kauffman, Ch. A. (2015), *Reduplication reflects uniqueness and innovation in language, thought and culture*, New York: York College of Pennsylvania.

Khanjan, A., and Alinezhad, B. (2010), "A morphological doubling approach to full reduplication in Persian", *SKY Journal of Linguistics*, vol. 23, pp. 169–198.

Lonnet, A. (2006), "Les langues sudarabiques modernes", *Faits de Langues*, vol. 27, no. 1, pp. 27-43.

Matthews, Ch. (1969), "Modern South Arabian Determination-A Clue Thereto from Shaḥri", *Journal of the American Oriental Society*, vol. 89, no. 1, pp. 22-27.

Mahmoodian, S. (2007), *Encyclopedia Larestanica*. Richmond, VA: Brandylane Publishers.

Maḥmūdī Baxtīyārī, B., and Ẕolfaqār Kondorī, Z. (2015), "Do-gān-sāzī-ye kāmel dar zabān-e Fārsī: yek bar-rasī-ye peykare-bonyād" [Full reduplication in Persian language: A corpus-based study], *Pažūhešnāme-ye pardāzeš va modīrīyyat-e eṭṭelāʿāt* [Iranian Journal of Information Processing and Management], vol. 31, no. 1, pp. 147–161. In Persian.

Mirmokri, M., and Seifori, S. (2016), "On the reduplication in Kurdish language", *International Journal of Kurdish Studies*, vol. 2, no. 3, pp. 165–178.

Molčanova, E. (2000), "Gūyeš-hā-ye Lārī", trans. from Russian to Persian by M. Ehsani, *Nāme-ye Farhangestān* [Academy's Letter], vol. 5, no. 2, pp. 183–187. Original work published in 1982.

Moridi, B. (2007), "Bar-rasī va moqāyese-ye gūyeš-e Lārī bā Daštestānī" [Analysis of the Lāri dialect and comparison with Dashtestāni dialect], *Orientalia*, vol. 4, pp. 64–72. In Persian.

Moridi, B. (2009), "The dialects of Lar (The state of research)", *Iran and the Caucasus,* vol. 13, no. 2, pp. 335–340.

Moridi, B. (2011), "Bar-rasī-ye čand važe-ye Lārī va moqāyese bā zabān-hā-ye Īrānī-ye bāstānī va mīyānī" [A study of some Lārī words in comparison with Old and Middle Iranian languages], *Farhang-e mardom* [Folklore], vol. 10/40, 212–218. In Persian.

Morris, M., Watson, J.C.E., and Eades, D. (2019), A *Comparative Cultural Glossary across the Modern South Arabian Language Family*, Oxford: Oxford University Press.

Nakano, A. (1986), *Comparative Vocabulary of Southern Arabic: Mahri, Gibbali, and Soqotri* (Studia Culturae Islamicae, no. 29), Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

O'Grady, W., and de Guzman, V. P. (1997), "Morphology: the analysis of word structure", *Contemporary Linguistics: An Introduction*, 3rd edition, ed. W. O'Grady, M. Dobrovolsky, and F. Katamba, London/New York: Longman, pp. 132-180.

Olsson, L. (2015), *Form and function of reduplicated nouns in Japanese* [Master's thesis], Stockholm University.

Ourang, M. (2022), *A reference grammar of Aheli: A dialect of the Lāri language, Iran* [Doctoral dissertation], University of New South Wales (UNSW).

Rasekh Mahand, M. (2009), "Bar-rasī-ye maʿnāyī tekrār dar zabān-e Fārsī" [Semantic analysis of reduplication in Persian], *Majalle-ye zabān-šenāsī* [Journal of Linguistics], vol. 23, no. 1, pp. 65–73. In Persian.

Rasekh Mahand, M., and Mohammadirad, M. (2014), "Bar-rasī-ye ṣūrī va maʿnāyī-ye tekrār dar gūyeš-e Kordī-ye Sōrānī" [Formal and semantic properties of reduplication in Sorani Kurdish], *Zabānhā va gūyešhā-ye Īrānī* [Iranian languages and dialects], vol. 3, pp. 133–146. In Persian.

Ratcliffe, R. (1992), *The Broken Plural Problem in Arabic, Semitic and Afro-asiatic: A Solution Based on the Diachronic Application of Prosodic Analysis*, Ph.D. dissertation, Yale University.

Rodgers, J. (1991), "The Sub-grouping of the South Semitic Languages", in: *Semitic Studies in Honor of Wolf Leslau on the occasion of his eighty-fifth birthday*, ed. Alan S. Kaye, Wiesbaden: Otto Harrassowitz, pp. 1323-1335.

Rubino, C. (2005), "Reduplication: Form, function and distribution", in: *Studies on Reduplication*, ed. B. Hurch with editorial assistance of V. Mattes, Berlin/New York: Mouton de Gruyter, pp. 11–30.

Salāmī, A. (2004–2011), *Ganjīne-ye gūyeš-šenāsī-ye Fārs* [The treasury of the dialectology of Fars], vols. 1–6, Tehran: Academy of Persian Language and Literature. In Persian.

Skjærvø, P. O. (1989), "Languages of Southeast Iran: Lārestānī, Kumzārī, Baškardī", in: *Compendium Linguarum Iranicarum*, ed. R. Schmitt, Wiesbaden: Ludwig Reichert Verlag, pp. 363–369.

Šeqāqī, V. (2000), "Farāyand-e tekrār dar zabān-e Fārsī" [The process of reduplication in Persian], in: *Majmūʿe-ye maqālāt-e čahāromīn konferāns-e zabān-šenāsī-ye naẓarī va kār-bordī* [Proceedings of the 4th Conference on Theoretical and Applied Linguistics], vol. 1, ed. S. A. Miremadi, Tehran: Allameh Tabataba'i University Press, pp. 519–534. In Persian.

Unseth, P. E. (2003), *Bi-consonantal reduplication in Amharic and Ethio-Semitic*. [Doctoral dissertation], University of Texas at Arlington.

Urbanczyk, S. (2017, March 29). “Phonological and morphological aspects of reduplication”, *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.80.

Vāḥedī Langarūdī, M. M., and Yūsefīrād, F. (2004), “Bar-rasī-ye farāyand-e tekrār-e nāqeṣ dar zabān-e Fārsī bar asās-e sāxt-vāže-ye navāyī” [A study of partial reduplication process in Persian based on prosodic morphology theory], in: *Majmūʿe-ye maqāle-hā-ye šešomīn konferāns-e zabān-šenāsī-ye mellī* [Proceedings of the 6th National Linguistics Conference], ed. E. Kāẓemī, Tehran: Allameh Tabataba’i University Press, pp. 403–420. In Persian.

Vos̲ūqī, M. B. (2001). *Dastūr-e zabān-e Lārestānī* [The grammar of the Lārestāni language], *Ketāb-e māh-e adabīyyāt va falsafe* [The Monthly Book on Literature and Philosophy], no. 44, pp. 62–65. In Persian.

Xonjī, L. (2009), *Negarešī tafṣīlī bar zabān-e Lārestānī va gūyeš-e Xonjī* [A detailed view on the Lārestānī language and the Xonjī dialect]. Shiraz: Īlāf. In Persian.

Xorram-Rūz, M. (2011), *Mīrās-e bāstān: Bar-rasī-ye riše-šenāsāne-ye āyīn-hā va bāvar-hā-ye farhang-e Lārestānī* [Ancient legacy: An analysis of rituals and beliefs in Larestani culture]. Shiraz: Īlāf. In Persian.

YEREVAN STATE UNIVERSITY PUBLISHING HOUSE

JIL 2/2 (2025): 66-86

Journal of Iranian Linguistics

Mapping the Languages of Kohgiluyeh va Boyer Ahmad Province, Iran: Is This Region Uniformly Lori Speaking?

Mortaza Taheri-Ardali*
Shahrekord University / Carleton University
Mansour Bozorgmehr
University of Ahvaz / UNESCO World Heritage Center, Susa, Iran
Erik Anonby
Carleton University

doi.org/10.46991/jil/2025.02.03

Abstract: Kohgiluyeh va Boyer Ahmad Province, which constitutes less than one percent of Iran's land area and a similar proportion of its population, is located in the mountainous south-west of the country. In the literature as well as in popular discourse in Iran, this region has been regarded as uniformly Lori-speaking. In this paper, we survey the distribution of languages in the province and investigate whether any languages other than Lori are spoken there as a mother tongue. Adopting the methodology outlined in the *Atlas of the Languages of Iran*, this article presents static and interactive maps showing each of the province's languages. The results of our survey indicate that in addition to seven distinct varieties of Southern Lori, Ghashghāi Turkic, Khuzestāni Arabic, and Bakhtiari are spoken as ancestral languages in the southern districts of Kohgiluyeh va Boyer Ahmad. Further, we signal the growing prevalence of Persian, the dominant prestige language across Iran, as a mother tongue.

Keywords: Kohgiluyeh va Boyer Ahmad Province, Linguistic geography, Southern Lori, Persian, Ghashghāi (Qashqāi) Turkic, Arabic

Mortaza Taheri-Ardali
E-mail: taheri@sku.ac.ir
mortazataheriardali@cunet.carleton.ca
ORCID: https://orcid.org/0000-0002-1260-6128
Mansour Bozorgmehr
E-mail: bozorgmehrmansour@gmail.com
ORCID: https://orcid.org/0009-0002-1420-383X
Erik Anonby
E-mail: erik.anonby@carleton.ca
ORCID: https://orcid.org/0000-0002-8719-4922
Received: 30.09.2025
Revised: 29.11.2025
Accepted: 05.12.2025

Conflict of Interest
The authors declare no conflicts of interest.

1. Introduction

Kohgiluyeh va Boyer Ahmad Province, with a population of approximately 700,000 people across over 2,000 settlements (ISC 2016), is located within the Zagros Mountains of south-western Iran. This tiny province makes up less than one per cent of Iran's land mass, but it shares borders with five other provinces: Chahar Mahal va Bakhtiari (C&B) to the north, Khuzestan to the west, Bushehr to the south-west, Esfahan to the east, and Fars to the south-east and east. Kohgiluyeh va Boyer Ahmad (hereafter K&B) is often regarded as a linguistically homogenous region where only Lori is spoken. Important studies on the region's Lori varieties have been conducted by scholars such as Lecoq (1989), Moqimi (1994), Loeffler and Windfuhr (1989), Anonby (2003a; 2003b; 2012), Nazari et al. (2015), and Taheri (2016). From the perspective of linguistic geography, Behnstedt (1990), Hourcade (2013), and Izady (2006–2025) have mapped the languages of this area in the wider context of Iran, and language distribution in most of the neighbouring provinces has been detailed (see *Section 2*). That said, it has not been clear, whether to experts or non-experts, how many different languages are spoken in K&B Province. For this reason, a more thorough investigation of linguistic composition in K&B is needed.

In this article, we seek to address the following three research questions:

1. Beyond Lori, are there any other languages spoken as a mother tongue in K&B?
2. How do speakers of Lori conceptualize and identify their language varieties, and what major varieties do they recognize?
3. To what degree has shift toward Persian, which is now impacting all parts of Iran, affected the transmission of ancestral languages as a mother tongue in the region?

The structure of the article is as follows: Section two provides a brief introduction to the *Atlas of the Languages of Iran* (ALI) project, a wide-ranging initiative in the field of Iranian linguistics. The third section describes the research process followed to investigate language distribution in K&B. After this, section four presents the key results of our research through province-level language maps, while the fifth part examines the individual languages spoken in K&B and their classification. The final section summarizes and reflects on our findings, and suggests avenues for further research.

2. The *Atlas of the Languages of Iran* (ALI) project

In Iranian linguistics, a large body of studies is devoted to the documentation, description and analysis of specific languages or language groups. However, it remains challenging to find scholarly publications that provide detailed

accounts of language distribution across particular regions. Two important studies in Iran are those of Bazin et al. (1982) and Papoli-Yazdi (1988), who produced maps of language distribution to the level of each settlement for parts of Gilan and Khorasan, respectively.

Recently, research on language distribution in Iran has gained further momentum. The initiation of the *Atlas of the Languages of Iran* (ALI) (Anonby and Taheri-Ardali et al. 2015-2025), a multi-institutional project carried out in collaboration with the Geomatics and Cartographic Research Centre (GCRC) at Carleton University in Ottawa, Canada, has progressively brought attention to this subject, engaging both scholars and the wider public. The ALI project was initiated in 2009 and launched online in 2015 through the website https://iranatlas.net. The project's structure and objectives are discussed, along with sample maps, in Anonby et al. (2019). Since its inception, the atlas has evolved in both technical and linguistic aspects. The website has been expanded, and various modules have been added in both Persian and English. In terms of the linguistic research that drives the atlas, several significant steps have been taken, including collection of language distribution and linguistic data in diverse geographical areas in Iran. Language distribution methodology is set out in Anonby, Mohammadirad and Sheyholislami (2019). Results of language distribution research for 10 of Iran's 31 provinces so far have been completed and published online: Hormozgan (Mohebbi Bahmani et al. 2015), Chahar Mahal va Bakhtiari (Taheri-Ardali and Anonby et al. 2015), Kordestan (Mohammadirad and Anonby et al. 2016), Bushehr (Nemati et al. 2017), Ilam (Gheitasi and Anonby et al. 2017), Hamadan (Izadi et al. 2021), Esfahan (Talebi-Dastenaei et al. 2022), Gilan (Poshtvan and Anonby et al. 2022), Lorestan (Taheri-Ardali and Anonby et al. 2023), and Khuzestan (Bozorgmehr et al. 2024). With regard to linguistic documentation, linguistic data questionnaires have been collected and oral texts have been recorded to date in the provinces of C&B, Zanjan, Hormozgan, Kermanshah, Ilam, Hamadan, Gilan and Esfahan. Proof-of-concept maps produced using linguistic data from C&B Province are available online (e.g., https://iranatlas.net/module/linguistic-data.cb-lexicon-leaf) and results of research have been shared and analyzed in other venues as well. For example, Anonby et al. (2021) reported findings on language distribution and carried out a lexical analysis of data from 31 language varieties in C&B.

The present study continues this line of research, drawing on a streamlined methodology and the cumulative knowledge and insights from earlier works. It is the most recent effort to identify the linguistic composition of a province as a piece in the larger puzzle of Iran's linguistic tapestry. The next section details the research process adopted for this study.

3. Research process

The present research on K&B Province followed that of Khuzestan (Bozorgmehr et al. 2024), where co-author Mansour Bozorgmehr gained substantial knowledge of pertinent research methodology and the neighbouring province's languages. The initial phase of work on K&B therefore drew on his own expertise as well as the available literature, together with observations from colleagues and acquaintances, to compile a provisional list of languages spoken in K&B. At the same time, ALI colleagues Hamideh Poshtvan and Mahnaz Talebi-Dastenaei compiled a spreadsheet containing demographic data from Iran's 2016 census (ISC 2016) – the most recent publicly-available dataset – and geographic coordinates for each of the province's 2257 listed populated places, using data from the National Cartographic Centre (NCC) of Iran and other sources.[1] These data, which are organized differently in the source documents, were manually aligned. Subsequently, Mansour Bozorgmehr conducted a two-week field survey across K&B in January 2024 to collect insights from local residents about the languages spoken in each city and village, and the estimated proportion of each language used as a mother tongue in that place. This survey resulted in a complete dataset on the geographic distribution of languages spoken in K&B. In tandem with this research, as a way of reinforcing local representations of language in the atlas, as well as familiarizing ourselves with linguistic forms characteristic in each region, local place names of each of the province's cities and villages were recorded and transcribed following the ALI transcription conventions[2]. Once data collection was complete, Mansour Bozorgmehr inserted the research findings into the spreadsheets containing the demographic and geographic data. To help ensure consistency, the entries were subsequently reviewed by Erik Anonby and Mortaza Taheri-Ardali. Hamideh Poshtvan and Mahnaz Talebi-Dastenaei conducted a final check of the spreadsheets to detect any remaining content and formatting issues.

4. Mapping language distribution

Having collected the data through fieldwork, filled in the ISC spreadsheet, checked and refined the data for language distribution as well as the Persian and local pronunciation of place names, the data were incorporated into the ALI website to map out the distribution of languages varieties with the support of technicians at GCRC.

---

[1] National Cartographic Centre, *National Database of Geographic Names of Iran*, Ministry of the Interior, National Cartographic Centre, 2025, https://gndb.ncc.gov.ir.

[2] For the ALI's transcription conventions: www.carleton.ca/iran/transcription/.

ALI makes use of two main types of maps to show language distribution and associated data: point-based interactive maps, and static polygon maps. Each plays an important but distinct role in representing language distribution. The interactive maps, which are constructed using the open-source Nunaliit Atlas Framework (https://nunaliit.org), provide detailed data at the settlement level, allowing users to explore each location individually. Static maps, built using the open-source mapping program QGIS (https://qgis.org), offer a broader overview that is useful for quick reference. The static maps have certain limitations, however: the boundaries they display do not clearly indicate how many settlements are included within each polygon-defined area, and they represent solely the dominant language in each community. Simply put, while interactive maps allow for in-depth exploration, static polygon maps provide clearer, but simplified, representations of complex patterns.

Figure 1 shows the interactive map generated for K&B Province on the ALI website. The map, constructed by a team of colleagues at GCRC, shows only languages spoken as a mother tongue, and only the main language in each place. However, clicking on any of the places brings up estimated proportions for all languages spoken as a mother tongue in that location. Places where no single language constitutes a majority, or portions of a community's population that includes people from all over Iran, with no further groups easily identifiable there, are designated as "mixed".

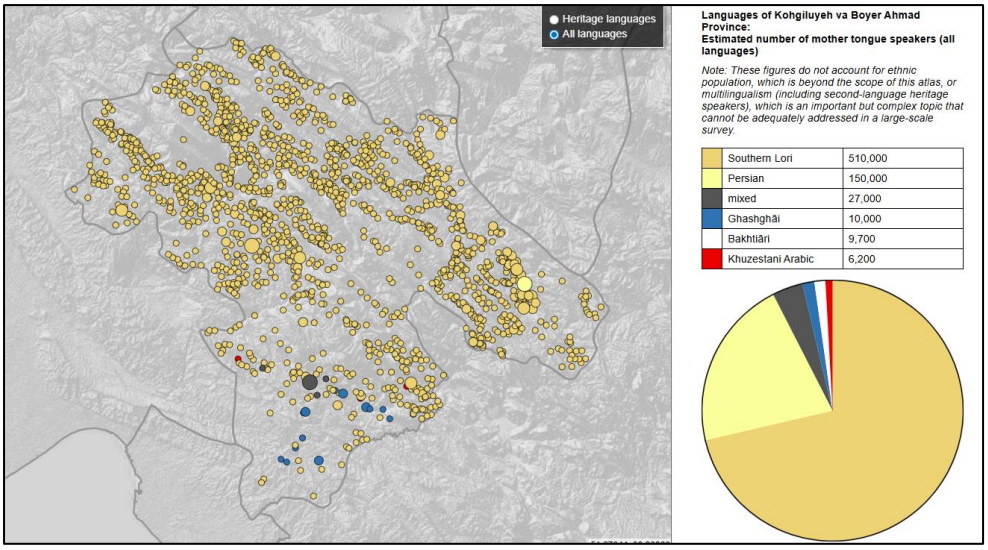


Figure 1. *Interactive map of language distribution in K&B Province at the settlement level. Source: https://iranatlas.net/module/languagedistribution.kohgiluyeh_va_boyer_ahmad.*

Figure 2 presents a static polygon map generated with QGIS, based on the same language distribution data. This map was constructed by GCRC colleague Adam Stone, along with Erik Anonby. In addition to marking the centre of each provincial sub-district (P. *shahrestān*), it situates italic labels for each Lori variety in the primary geographic zone where it is spoken.

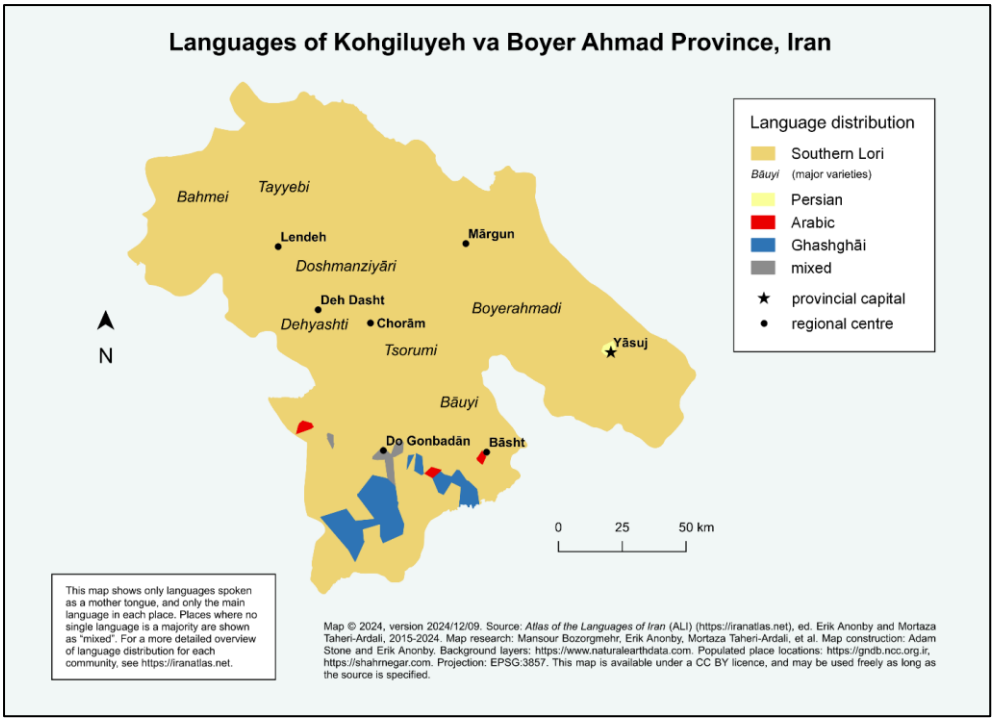


***Figure 2**. Polygon map of language distribution in K&B Province. Source: https://iranatlas.net/module/languagedistribution.kohgiluyeh_va_boyer_ahmad_static.*

Both maps show that there are several languages spoken as a mother tongue in K&B. As anticipated, Lori – and more specifically, Southern Lori (see 5.1) – is the dominant language of most of the province. However, Persian, Ghashghāi (Qashqāi) Turkic and Khuzestāni Arabic are also spoken here, and each of these is dominant in some locations. Bakhtiari speakers do not constitute a majority in any location, but their presence is signalled by the discussion and figures in the interactive module's side panel, and can be found by clicking on the dot for the city of Do Gonbadān (see Figure 10), which is the primary location where they are found.

5. Languages in K&B and their classification

Here, we provided a detailed description of these languages and their sub-varieties. For the purposes of discussion, we divide our account into

Southern Lori (5.1), other ancestral languages (5.2), and Persian (5.3). We follow this with a discussion of language distribution in Gachsārān in the southern part of the province (5.4), which is an area of higher linguistic diversity.

5.1. Southern Lori

Our findings confirm that Southern Lori, with over half a million mother-tongue speakers, is the dominant language of the province, and is spoken in all districts throughout the province. However, speakers distinguish a number of Lori sub-varieties, generally based on their association with different ethnic groupings.

Lori varieties belong to the Southwestern division of Iranian (Iranic) languages, and have been classified by Windfuhr and Perry (2009: 418) among what they refer to as "Perside" varieties of south-west Iran, as they share a range of features also seen in Early New Persian. The group of related varieties collectively known as Lori is spoken by the Lori and Bakhtiari peoples in the Zagros Mountains of western and south-western Iran and in surrounding lower-altitude regions (Anonby 2012). Within the Lori continuum, three major divisions (or languages) can be identified: Northern Lori, Bakhtiari, and Southern Lori (Anonby 2003b; Anonby 2012; Taheri-Ardali et al. 2025).

Due to historical movements of people and administrative actions, Lori speakers are now dispersed across twelve provinces of Iran. They constitute a significant portion of the population in four provinces: Khuzestan, Lorestan, C&B, and K&B. In Khuzestan, communities representing all three divisions (Northern Lori, Bakhtiari, and Southern Lori) are found; Northern Lori and Bakhtiari are spoken in Lorestan; and in C&B, Bakhtiari dominates. This study confirms that Southern Lori is the main language in K&B. Significant Lori-speaking minorities are also found in the provinces of Fars, Esfahan, Hamadan, Ilam, Bushehr, Markazi, and Gilan.

Although the Lori varieties documented in K&B – namely, Bahmei, Bāuyi (P. Bāvi), Boyerahmadi, Dehyashti (P. Dehdashti), Doshmanziyāri, Tayyebi, and Tsorumi (P. Chorāmi) – exhibit internal linguistic diversity, they share enough features to be classified collectively under the broader term Southern Lori.

Through the ALI website's navigation features, we can observe the geographic distribution of each language variety in the region. The following figure, a screenshot from the website, shows the distribution of Boyerahmadi, with specific data about the capital city of Yāsuj, as an example, displayed in the side panel. Consultants estimated that while just over half of the city's

population now speaks Persian as a mother tongue, approximately a third of the people in the city have learned Boyerahmadi as their first language.

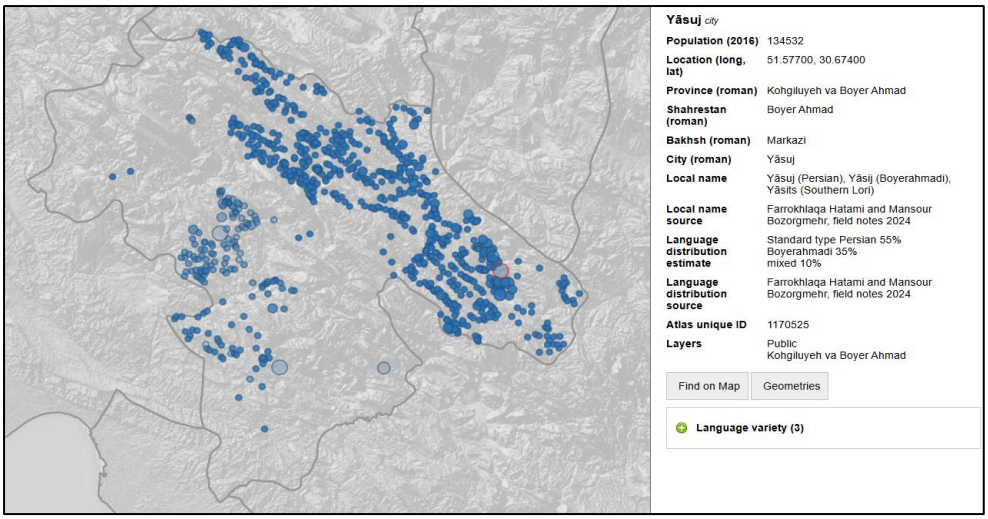


Figure 3. *Distribution of Boyerahmadi in K&B Province.*

As seen in the above figure, Boyerahmadi extends to the northern, eastern, and western parts of K&B. The side panel also provides information on population, coordinates, local place name transcription, estimated proportion of mother tongue speakers of each language community, and the unique Atlas ID for each settlement. The darker blue dots represent a higher proportion of speakers in each settlement.

The Bahmei grouping within Southern Lori is mainly distributed in the north-western part of the province (Figure 4). Data for the city of Dishmuk are illustrated in the right panel. Bahmei-speaking areas are located near the border of Khuzestan Province. The language distribution map of Khuzestan Province (see Bozorgmehr et al. 2024) reveals that the Bahmei speakers are also present in the south-eastern part of that province, adjacent to K&B.

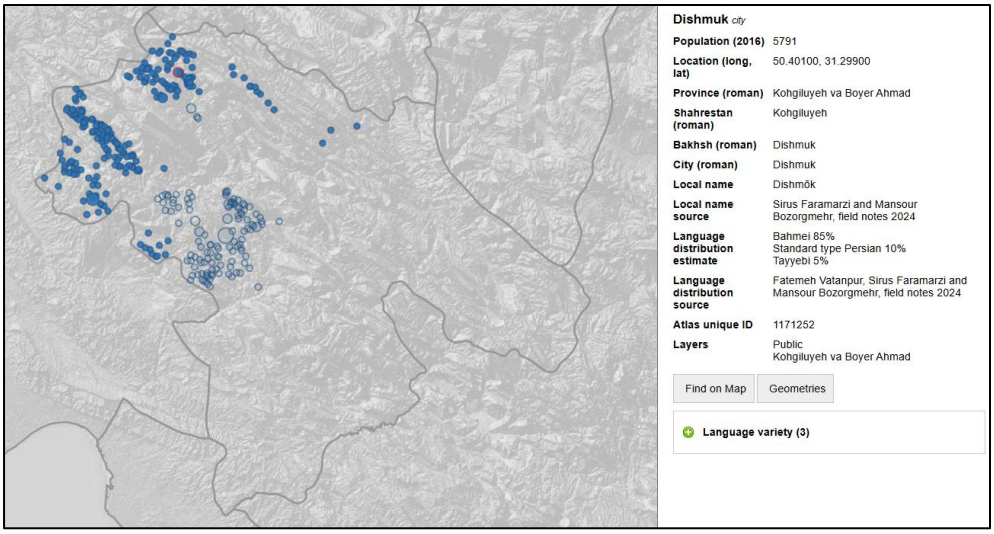


Figure 4. *Distribution of Bahmei in K&B Province.*

Bāuyi (P. Bāvi) is a Southern Lori variety spoken in the southern part of K&B near the border with Fars Province (Figure 5). Bāsht is one of the main centres with Bāuyi speakers, as shown in the side panel. In this city, about 70% of the population speaks Bāuyi.

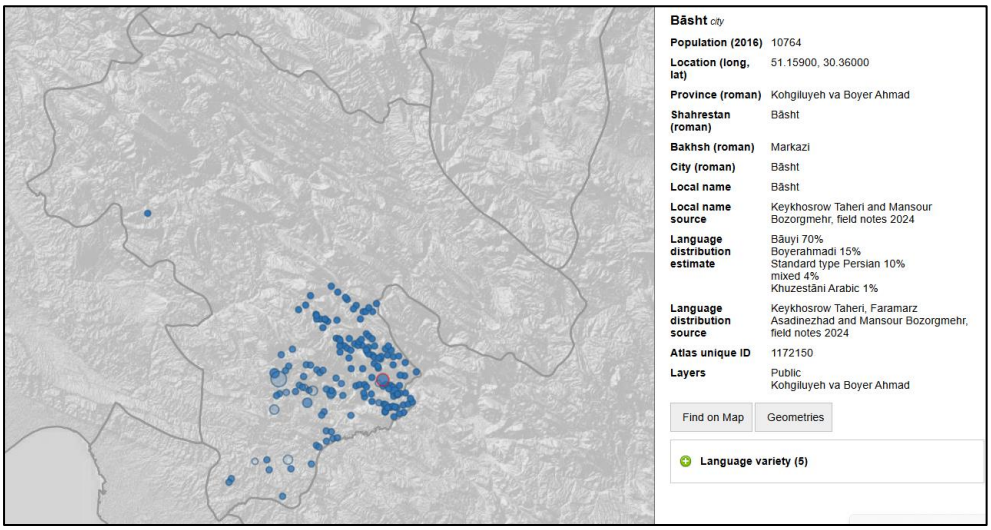


Figure 5. *Distribution of Bāuyi (Bāvi) in K&B Province.*

The next sub-variety of Southern Lori encountered in our research is Doshmanziyāri. The settlements where it is spoken as the mother tongue are shown in Figure 6. The main concentration of Doshmanziyāri speakers is located in the mid-western part of the province, extending southward toward the border with Khuzestan Province. The data for Dāryāb, as a small village in this part of the province, are shown in the following figure. The local pronunciation for this settlement is transcribed as Dāryow.

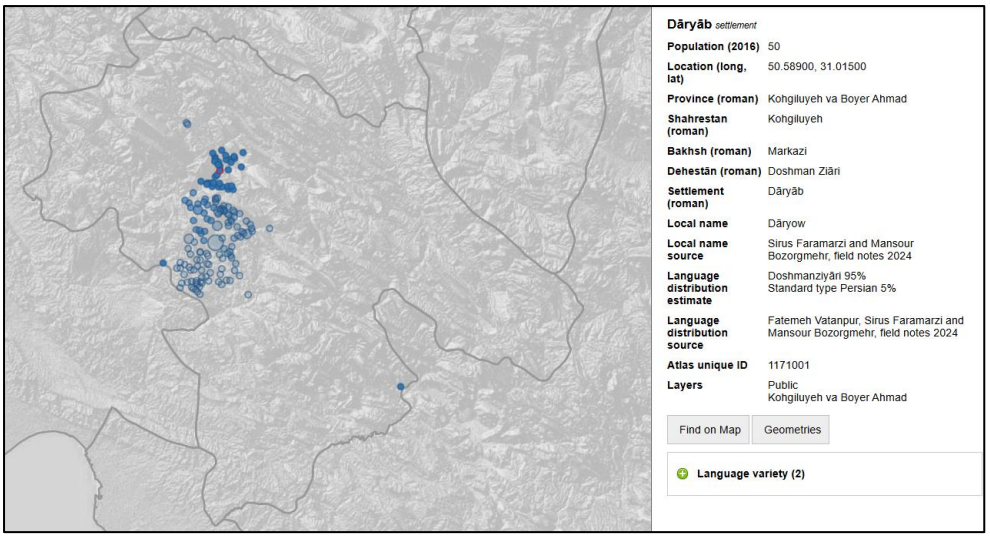


Figure 6. *Distribution of Doshmanziyāri in K&B Province.*

Tayyebi, another sub-variety of Southern Lori, is depicted in Figure 7. As evident in the map, certain dots are rendered in a darker blue, indicating a higher proportion of speakers in the given settlement. Jāvar Deh, with a population of 1276, and an estimated 95% of mother-tongue Tayyebi speakers, is shown as an example in the right panel. There are also a few Tayyebi speakers in Khuzestan Province, settled in pockets located near the provincial border with K&B (see Bozorgmehr et al. 2024).

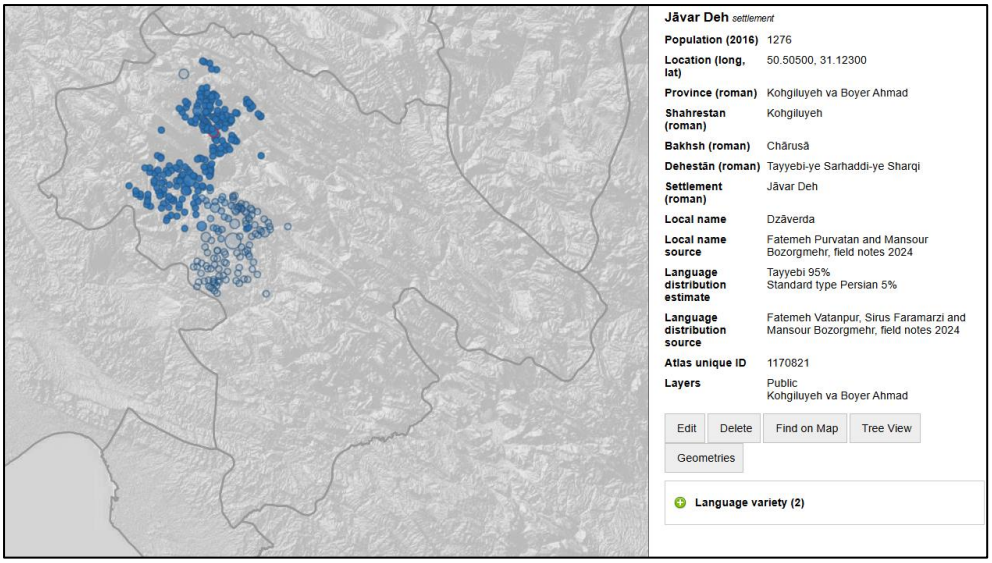


Figure 7. *Distribution of Tayyebi in K&B Province.*

Figure 8 shows the distribution of Tsorumi (P. Chorāmi) in the central part of the province. The city of Chorām is the largest community in that area, with an estimated 80% of the population speaking Tsorumi as a mother tongue.

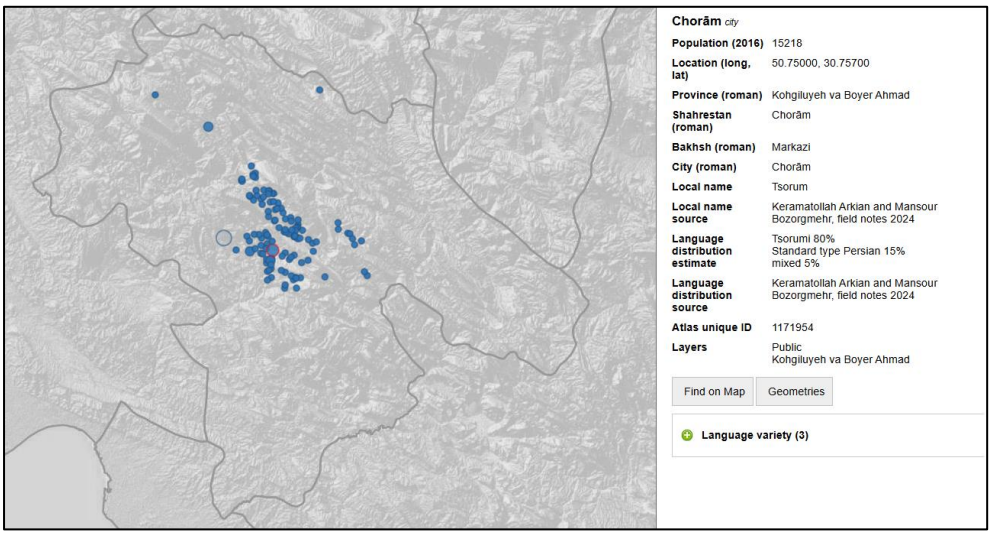


Figure 8. *Distribution of Tsorumi (P. Chorāmi) in K&B Province.*

A final sub-variety of Southern Lori, Dehyashti (P. Dehdashti), is the only variety reported by local residents as being spoken in a single location, namely, the city of Deh Dasht. As this sub-variety is attested in only one locality, a separate figure has not been provided. Consultants described Dehyashti as a fusion of other Southern Lori varieties, arising out of the mixture of people from many surrounding locations and resulting, over time, in a distinctive variety associated with the city itself rather than ethnic identification.

This brings the discussion back to the larger point that the other Southern Lori-speaking varieties documented in our research are identified by speakers using terms – Bahmei, Bāuyi (P. Bāvi), Boyerahmadi, Dehyashti (P. Dehdashti), Doshmanziyāri, Tayyebi, and Tsorumi (P. Chorāmi) – which reflect ethnic identities and conceptualizations rather than strictly linguistic criteria. Such categorizations, originating in the perspectives of the language communities themselves, are an important starting point for understanding language classification, and can be complemented by follow-up classificatory research based on comparison of linguistic structures from each of the areas (for further discussion, see Anonby and Sabethemmatabadi 2019; Anonby et al. 2020).

It is our impression that overall, from a linguistic perspective, the Southern Lori varieties of K&B fall within an intermediate position between Bakhtiari to the north, the Lirāvi varieties of Southern Lori to the west, and the Mamasani dialects of Southern Lori to the south-east.

During the transcription of place names, we observed three remarkable phonological patterns among Southern Lori varieties in K&B. First, in contrast to other Southern Lori varieties, where fronting occurs only before front vowels (for example, in Mamasani Southern Lori of Fars Province, cf. Anonby 2003a: 66-7), and in line with a similar sound change in Mukri Kurdish varieties much further to the north (Anonby and Öpengin 2026), in K&B there is sometimes a complete phonemic shift of the palato-alveolar affricates *č *ǰ to alveolar *ts* and *dz* respectively. The change in phonemic identity is confirmed by the alveolar pronunciation of *ts* before a vowel, for example, in the local names [t͡s]āl Gerru (cf. the official name, rendered Chāl Gerru in Persian) and Mela-y Ghu[t͡s]un (P. Tol-e Quchān). The voiced counterpart *dz* patterns similarly in the name [d͡z]āverda (P. Jāvar Deh). The palato-alveolar slot, for its part, is sometimes filled by a fronted allophone of *k* before front vowels, for example, in the local name Garda [t͡ʃ]alāt (P. Gard-e Kalāt).

Second, in a smaller subset of these varieties, a three-way contrast has emerged between *ts č k*. Contrast between *k* and *č* is found before the front vowel a. In contrast to the fronted stop in Garda Kalāt, mentioned just above,

both stops are velar in the first word of the local name [k]a[k]e-y Mobārak (P. Kākā Mobārak). It is unclear whether this pattern extends to the voiced counterparts; place names with forms like [d͡ʒ]owrensun (P. Gabrestān) do occur, apparently contrasting with *g* in [g]urow (P. Gurāb), but as the underlying first vowel in [d͡ʒ]owrensun may be a, as in some other Southern Lori varieties, contrast among the voiced stops *dz ǰ g* is less certain. To our knowledge, these three-way contrasts have not been reported in other Iranian (Iranic) languages.

Finally, as we previously observed in Southern Lori varieties in Khuzestan, and has been reported for Mamasani Southern Lori in Fars Province (Anonby 2003a: 83-84), in many varieties in K&B both *e* and *o* are realized as a mid-central [ə] in unstressed open syllables, and thus the distinction between them is neutralized (or perhaps **o* has shifted to *e*) in this position. Examples of place names where this neutralization occurs are:

S[ə]kandari	P. Eskandari
M[ə]la Owgir	P. Meleh-ye Ābgir
B[ə]nāri	P. Bonāri
H[ə]seyn Ābāy Mokhtār	P. Hoseyn Ābād-e Mokhtār

5.2. Other ancestral languages: Southern K&B

Although Southern Lori is the dominant language of K&B as a whole, the southern corner of the province shows significant linguistic diversity. Three other ancestral languages are spoken here: Turkic, Arabic, and Bakhtiari. Turkic and Arabic in particular are the main languages in a number of communities in this area.

The Turkic variety spoken in this region is Ghashghāi (also spelled Qashqāi, Qašqā'i, and Kashkāy). Ghashghāi belongs to the Southwestern (Oghuz) sub-branch of the Turkic language family together with, for instance, Azerbaijani, Turkmen, and Turkish of Turkey (Dolatkhah et al. 2016; Knüppel 2015). This language is spoken by traditionally nomadic members of the Ghashghāi tribal confederation in the southern provinces of Iran, centred in Fars Province. There are about 20 locations where Ghashghāi speakers maintain residences in southern K&B, which is a geographical continuation of the main Ghashghāi language area in Fars. In total, about 10,000 people speak the language in K&B. Bid Zard, with a population of just under 400, is one such place, and an estimated 90% of people speak Ghashghāi as a mother tongue in this locality.

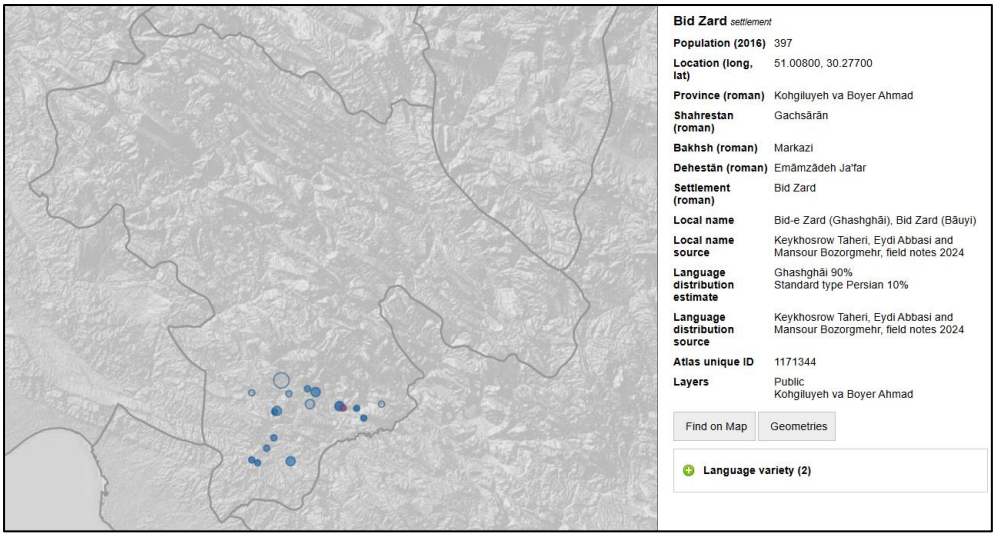


Figure 9. *Distribution of Ghashghāi Turkic in K&B Province.*

Arabic is another ancestral language spoken in Gachsārān District. The particular variety spoken here is Khuzestāni Arabic, which is centred in Khuzestan Province to the west. This variety has been classified among the Bedouin-type "gələt" dialects of southern Mesopotamia (Leitner 2020: 116). Originating in Khuzestan, speakers of Khuzestāni Arabic in southern K&B are scattered across nine locations, of which eight are in Gachsārān, and on (the village of Borghun) is in Bāsht District. In total, the results of our survey indicate that there are over 6000 mother-tongue speakers of this language in the province. Most live in the city of Do Gonbadān, where consultants estimated that approximately 5% of the city's population of over 97,000 speaks Khuzestāni Arabic as a mother tongue.

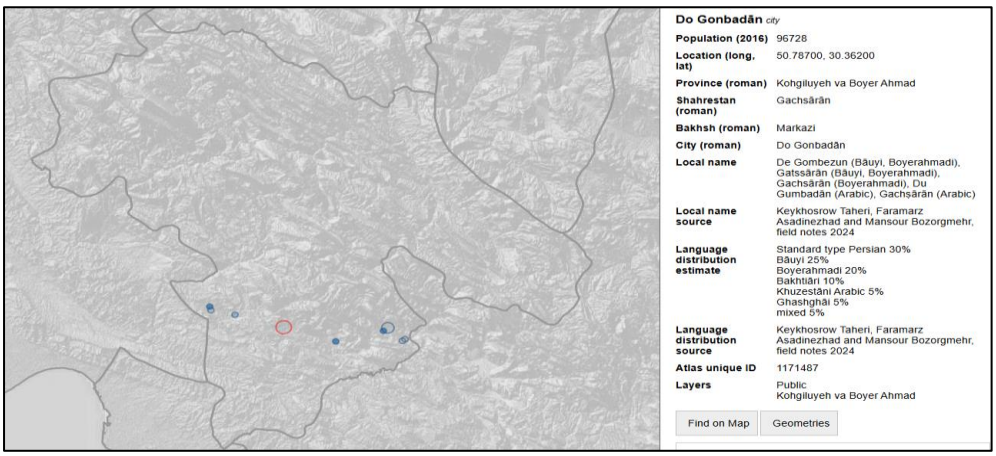


Figure 10. *Distribution of Arabic in K&B Province.*

Bakhtiari is the final ancestral language encountered during the course of our research in Gachsārān. Like Southern Lori (5.1), Bakhtiari is classified as Southwestern Iranian (Windfuhr 2009, McKinnon 2011, Anonby and Asadi 2014, Anonby 2025), and is a major group within the Lori continuum. Bakhtiari is spoken in the city of Do Gonbadān, which is the only location in K&B where consultants indicated that a significant Bakhtiari-speaking population is found. Consultants estimated that close to 10,000 people in the city speak Bakhtiari as a mother tongue. Similar to the Arabic speakers of this region, most of Bakhtiari speakers are originally from Khuzestan Province, and their presence in K&B is primarily attributable to the development of the oil industry.

5.3. Persian

Persian has emerged as a mother tongue in almost all locations throughout K&B, as many parents now speak Persian to their children at home. As a result, we estimate Persian is now the first language of 150,000 people, which is almost a quarter of the provincial population. In the capital city of Yāsuj in particular, consultants maintain that it is now the main mother tongue.

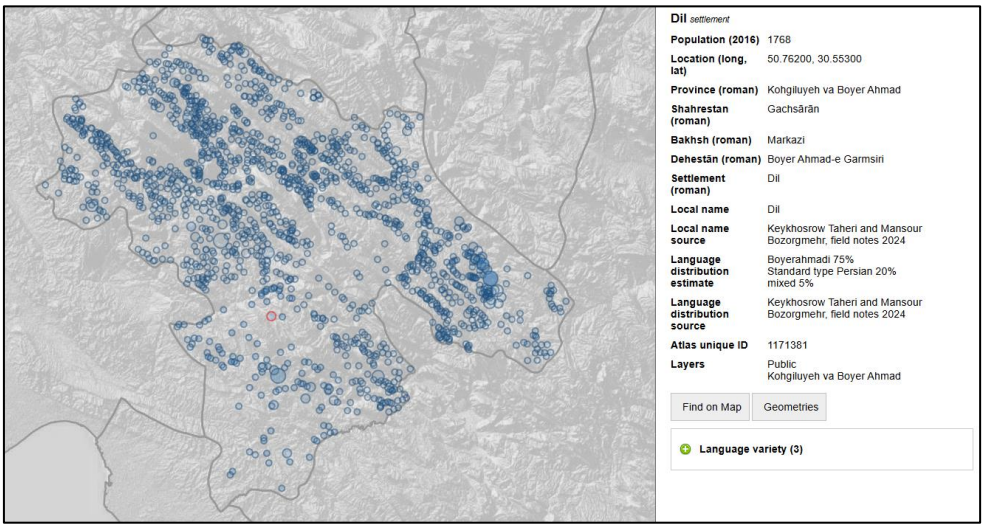


Figure 11. *Distribution of Persian in K&B Province.*

Figure 11 above shows an example of Persian spoken as a mother tongue by an estimated 20% of people in Dil, a town of about 1800 people in the west of the province.

5.4. Language distribution in Gachsārān

This Gachsārān provincial sub-district in southern K&B is one of the many hubs of the oil industry in Iran. Over the past century, many people have moved here from other areas of the province, and from other areas of Iran, particularly Khuzestan. In Gachsārān's capital city, Do Gonbadān, all of the province's ancestral languages come into contact with each other: Turkic, Arabic, and Bakhtiari, as well as the Bāuyi (Bāvi) and Boyerahmadi varieties of Southern Lori; and Persian is also spoken here. The interaction between languages in this city, and in southern K&B more general, is a worthwhile topic for further investigation.

Conclusion and future directions

This article presents the results of the first in-depth study of language distribution to be carried out in Kohgiluyeh va Boyer Ahmad Province. While this province has been viewed as uniformly Lori-speaking, our survey reveals significant linguistic diversity, with representatives of the Iranian (Iranic), Turkic, and Semitic language families.

For Southern Lori in particular, we found that ethnic identity, particularly affiliation with a given ethnic sub-grouping, is foundational in how speakers classify and label their varieties. The key Southern Lori varieties identified according to ethnic labels are Bahmei, Bāuyi (P. Bāvi), Boyerahmadi, Doshmanziyāri, Tayyebi, and Tsorumi (P. Chorāmi). In contrast, the Dehyashti (P. Dehdashti) variety is defined by its community of origin rather than ethnicity.

The other ancestral languages we encountered over the course of our research in K&B are Ghashghāi Turkic, Khuzestāni Arabic, and Bakhtiari. These three languages are found in various places in the southern district of Gachsārān, and they come into contact with one another and with Southern Lori in the district capital, Do Gonbadān.

In recent decades, Persian has emerged as a significant mother tongue, and its impact in displacing the ancestral languages of K&B Province cannot be underestimated. Based on discussions with consultants regarding the prevalence of parents introducing Persian as the first language of their children in the home in all parts of the province, we estimate that Persian is now the first language of nearly one quarter of the province's population. This phenomenon is happening in all of the language communities in K&B, indicating that intergenerational transmission of ancestral languages as a mother tongue is ceasing – a likely precursor to the eventual loss of these languages by the community as a whole.

While this study makes a foundational contribution by considering, for the first time, language distribution in each of K&B Province's 2257 listed populated places, because of the complexity of individual and community multilingualism (see Anonby and Yousefian 2011; Anonby et al. 2020), only the mother tongue is considered. The results of the study are also limited in that they are based on a survey rather than a census. It was neither possible nor practical to visit each populated place, and to inquire about language distribution from each person there, or even a representative cross-sampling of the population. Instead, we relied on consultants familiar with the language situation in each of the province's districts. The resulting language distribution figures are clearly indicated as estimates in the atlas, and need to be treated as such. This methodology has enabled a coherent picture of the language situation in a way that past censuses have not been able to achieve, but there is still room for improvement of language distribution estimates for each place. For this reason, the interactive atlas provides functionality for further, moderated contributions from people with additional local knowledge of the language situation.

In addition to addressing the research questions raised above, this study provides the necessary context for selecting field research locations for further linguistic documentation during the next phase of this project within the ALI framework. As is being done for other regions of Iran, the collection of linguistic data questionnaires and oral texts will help provide a complementary, and more detailed, understanding of the languages in this surprisingly varied province. The classification of Southern Lori sub-varieties can be re-examined, and the impact of contact among the various languages in the southern portion of the province can be explored. Along with the language distribution research presented here, this documentation will also contribute to a record of linguistic diversity in K&B at this final period in history where the ancestral languages of the province still resonate the streets and paths of each community.

Financial support

This work was partially supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) under Grant 435-2021-0794, *An Atlas of the Languages of Iran*, 2021–2026.

BIBLIOGRAPHY

Anonby, E. (2003a), *A phonology of Southern Luri*, Munich: Lincom Europa.

Anonby, E. (2003b), "Update on Luri: How many languages?", *Journal of the Royal Asiatic Society*, vol. 13, no. 2, pp. 171-197.

Anonby, E. (2012), "Sociolinguistic status of Lori", *Encyclopaedia Iranica*, New York: Center for Iranian Studies, Columbia University. Available at: http://www.iranicaonline.org/articles/lori-language-ii (accessed 2025-08-10).

Anonby, E. (2025), *Atlas of the Languages of Iran*: *A working classification*, Available at: https://iranatlas.net/index.html?module=module.classification (accessed 2025-08-10).

Anonby, E., and Asadi, A. (2014), *Bakhtiari studies: Phonology, text, lexicon*, Uppsala: Acta Universitatis Upsaliensis.

Anonby, E., Hayes, A., and Oikle, R. (2020), "A multi-dimensional approach to classification of Iran's languages", in: *Advances in Iranian linguistics* (Current Issues in Linguistic Theory 351), ed. by R. K. Larson, S. Moradi, and V. Samiian, Amsterdam: John Benjamins, pp. 29-56.

Anonby, E., Mohammadirad, M., and Sheyholislami, J. (2019), "Kordestan Province in the *Atlas of the Languages of Iran*: Research process, language distribution, and language classification", in: *Current Issues in Kurdish Linguistics* (Bamberg Studies in Kurdish Linguistics 1), ed. by S. Gündoğdu, E. Öpengin, G. Haig, and E. Anonby, Bamberg: University of Bamberg Press, pp. 9-38.

Anonby, E., and Öpengin, E. (in press), "Phonology of Kurdish: Comparative overview and theoretical issues" (= Chapter 8), in: *The Oxford Handbook of Kurdish Linguistics*, ed. by J. Sheyholislami, G. Haig, H. Khezri, S. Akin, and E. Öpengin, Oxford/New York: Oxford University Press.

Anonby, E., and Sabethemmatabadi, P. (2019), "Representing complementary user perspectives in a language atlas", in: *Further developments in the theory and practice of cybercartography: International dimensions and language mapping*, ed. by D. R. Fraser Taylor, E. Anonby, and K. Murasugi, Amsterdam: Elsevier, pp. 413-440.

Anonby, E., Schreiber, L., and Taheri-Ardali, M. (2020), "Balanced bilingualism: Patterns of contact influence in L1 and L2 Turkic and Bakhtiari speech in Juneqan, Iran", *Iranian Studies*, vol. 53, no. 3-4, pp. 589-622.

Anonby, E., Taheri-Ardali, M., et al. (eds.) (2015–2025), *Atlas of the Languages of Iran* (ALI), Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: https://iranatlas.net (accessed 2025-08-10).

Anonby, E., Taheri-Ardali, M., and Hayes, A. (2019), "The *Atlas of the Languages of Iran* (ALI): A research overview", *Iranian Studies*, vol. 52, no. 1-2, pp. 199–230.

Anonby, E., Taheri-Ardali, M., and Stone A. (2021), "Toward a picture of Chahar Mahal va Bakhtiari Province, Iran, as a linguistic area", *Journal of Linguistic Geography*, vol. 9, no. 2, pp. 106–141.

Anonby, E., and Yousefian, P. (2011), *Adaptive multilinguals: A survey of language on Larak Island*, Uppsala: Acta Universitatis Upsaliensis.

Bazin, M., Bromberger, Ch., Askari, A., and Karimi, A. (1982), *Gilân et Âzarbâyjân oriental: Cartes et documents ethnographiques*, Paris/Tehran: Institut Français d'Iranologie de Téhéran/Bibliothèque Iranienne.

Behnstedt, P. (1986–1990), "Vorderer Orient: Sprachen und Dialekte [Middle East: Languages and Dialects], plate A VIII 10", in: *Tübinger Atlas des Vorderen Orients (TAVO)*, ed. by E. Orywal, Wiesbaden: Reichert Verlag.

Bozorgmehr, M., Anonby, E., Bahrani, N., et al. (2024), "Language distribution in Khuzestan Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: http://iranatlas.net/module/language-distribution.khuzestan_ancestral (accessed 2025-08-10).

Dolatkhah, S., Csató, É. Á., and Karakoç, B. (2016), "On the marker -(y)akï in Kashkay", in: *Turks and Iranians: Interactions in Language and History*, ed. by É. Á. Csató, L. Johanson, A. Róna-Tas, B. Utas, Uppsala: Uppsala University.

Gheitasi, M., Anonby, E., et al. (2017), "Language distribution in Ilam Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: http://iranatlas.net/module/language-distribution.ilam (accessed 2025-08-15).

Hourcade, B. (2013), "Dominant languages and multilingualism", *Irancarto*. Available at: http://www.irancarto.cnrs.fr/record.php?q=AR-040503&l=en (accessed 2025-08-15).

Iran Statistics Center (2016), *Sarshomāri-ye omumi-ye nofus o maskan* [Public census of population and housing], Tehran: Ministry of the Interior, Iran Statistics Center. Available at: https://www.amar.org.ir (accessed 2025-08-10).

Izadi, E., Meshkinfam, M., Anonby, E., et al. (2021), "Language distribution in Hamadan Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: http://iranatlas.net/module/language-distribution.hamadan (accessed 2025-08-10).

Izady, M. M. (2006–2025), *Linguistic composition of Iran*. Available at: https://gulf2000.columbia.edu/images/maps/Iran_Languages_2000_sm.png (accessed 2025-08-10).

Knüppel, M. (2015), "Qašqāʾi tribal confederacy ii. Language", *Encyclopaedia Iranica*, New York: Center for Iranian Studies, Columbia University. Available at: https://www.iranicaonline.org/articles/qasqai-tribal-confederacy-ii-language (accessed 2025-08-10).

Lecoq, P. (1989), "Les dialectes du sud-ouest de l'Iran", in: *Compendium linguarum Iranicarum*, ed. by R. Schmitt, Wiesbaden: Reichert Verlag, pp. 341–349.

Leitner, B. (2020), "Khuzestan Arabic", in: *Arabic and contact-induced change*, ed. by Ch. Lucas, and S. Manfredi, Berlin: Language Science Press, pp. 115–134.

Loeffler, R., and Windfuhr G. (2016), "Boir Aḥmadī", *Encyclopaedia Iranica*, New York: Center for Iranian Studies, Columbia University. Available at: https://iranicaonline.org/articles/boir-ahmadi#pt2 (accessed 2025-08-10).

MacKinnon, C. (2011), "Lori dialects", *Encyclopaedia Iranica*, New York: Center for Iranian Studies, Columbia University. Available at: https://www.iranicaonline.org/articles/lori-dialects (accessed 2025-08-10).

Mohammadirad, M., Anonby, E. et al. (2016), "Language distribution in Kordestan Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: http://iranatlas.net/module/language-distribution.kordestan (accessed 2025-08-10).

Mohebbi Bahmani, H., Rashidi, A., Anonby, E., et al. (2015), "Language distribution in Hormozgan Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: http://iranatlas.net/module/language-distribution.hormozgan (accessed 2025-08-10).

Moqimi, A. (1994/1373), *Barresi-ye guyesh-e boyer ahmad va ...* [The dialect of Boyer Ahmad and, etc.], Shiraz: Navid-e Shiraz. In Persian.

Nazari, J., Khaleghzadeh, M. H., Moqimi, J., and Moqimi, A. (2015/1394), *Farhang-e vāzhehā-ye Lori-ye Boyer Ahmadi* [A Lori Dictionary of Boyer Ahmadi], Si Sakht: Farhang-e Mana. In Persian.

Nemati, F., Ghasemi, Sh., Anonby, E., et al. (2017), "Language distribution in Bushehr Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at:

http://iranatlas.net/module/language-distribution.bushehr (accessed 2025-08-10).

Papoli-Yazdi, M. H. (1988/1367), "Naqsheh-ye parākandegi-ye zabānhā dar rustāhā-ye shomāle-e Khorāsān" [Map of language distribution in villages of north of Khorasan], *Tahqiqāt-e Joghrāfiāi* [Geographical Research], no. 10, pp. 21–42. In Persian.

Poshtvan, H., Anonby, E., et al. (2020), "Language distribution in Gilan Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: https://iranatlas.net/module/language-distribution.gilan_heritage (accessed 2025-08-10).

Taheri, E. (2016/1395), *Guyesh-e Lori-ye Boyer Ahmad* [Lori dialect of Boyer Ahmad], Tehran: Institute for Humanities and Cultural Studies. In Persian.

Taheri-Ardali, M. (2021), "Definiteness marking in the languages of Chahar Mahal va Bakhtiari Province, Iran", *Paper presented at the 9th International Conference on Iranian Linguistics*, Vienna, 18–20 August 2021.

Taheri-Ardali, M., Anonby, E., et al. (2015), "Language distribution in Chahar Mahal va Bakhtiari Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: https://iranatlas.net/index.html?module=module.language-distribution.chahar_mahal_va_bakhtiari (accessed 2025-08-10).

Taheri-Ardali, M., Anonby, E., et al. (2023), "Language distribution in Lorestan Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: http://iranatlas.net/module/language-distribution.lorestan (accessed 2025-08-10).

Taheri-Ardali, M., Borjian H., and Anonby E. (2025), "The Bavadi and their Bakhtiari dialect", *Iranian Studies*, vol. 58, no. 1, pp. 1–47.

Talebi-Dastenaei, M., Borjian H., and Anonby, E. (2022), "Language distribution in Esfahan Province, Iran", in: *Atlas of the languages of Iran (ALI)*, ed. by E. Anonby, M. Taheri-Ardali, et al., Ottawa: Geomatics and Cartographic Research Centre (GCRC), Carleton University. Available at: http://iranatlas.net/module/language-distribution.esfahan (accessed 2025-08-10).

Windfuhr, G., and Perry, J. R. (2009), "Persian and Tajik", in: *The Iranian languages*, ed. by G. Windfuhr, London: Routledge, pp. 416–544.

YEREVAN STATE UNIVERSITY PUBLISHING HOUSE

JIL 2/2 (2025): 87-107

Journal of Iranian Linguistics

The Nonverbal Element in Persian Verbal Multiword Expressions: A Corpus Annotation Approach

Vahide Tajalli*
Shahid Beheshti University
Mehrnoush Shamsfard
Shahid Beheshti University
Yalda Yarandi
Shahid Beheshti University
Mahtab Sarlak
Shahid Beheshti University
Arezoo Haghbin
Shahid Beheshti University

doi.org/10.46991/jil/2025.02.04

Abstract: This article presents a linguistic framework for the identification and annotation of Persian (Farsi) Verbal Multiword Expressions (VMWEs), developed in alignment with the standards and methodologies set by the PARSEME Corpus—an international research network focused on the systematic analysis of multiword expressions across languages. The study aims to bridge the gap between universal annotation guidelines and language-specific grammatical features by tailoring the PARSEME framework to the structural and semantic properties of Persian. By extracting the characteristics of Persian VMWEs, particularly their nonverbal elements (preverbs) and their diverse syntactic and morphological patterns, this work contributes to a more refined understanding of Persian verbal idiomaticity and the advancement of natural language processing tasks. The article details the development of annotation guidelines that reflect both cross-linguistic categories and Persian-specific grammatical phenomena and the process of annotating a corpus of 5,617 sentences encompassing a wide range of Persian VMWEs including light verb constructions, verbal idioms, and prefix verbs. The practical applications of these guidelines in natural language processing are discussed, highlighting their potential to enhance machine understanding of complex verbal constructions, improve syntactic parsing accuracy, and support downstream tasks such as machine translation, information extraction, and semantic role labeling.

Keywords: Compound Verb; Nonverbal Element; Persian; Preverb; Text Corpus; Verbal Multiword Expression

Conflict of Interest
The authors declare no conflicts of interest.

Vahide Tajalli
E-mail: vtajalli@ut.ac.ir
ORCID: https://orcid.org/0000-0003-3118-2903
Mehrnoush Shamsfard
E-mail: m-shams@sbu.ac.ir
ORCID: https://orcid.org/0000-0002-7027-7529
Yalda Yarandi
E-mail: yalda.yarandi@gmail.com
ORCID: https://orcid.org/0009-0008-3088-8166
Mahtab Sarlak
E-mail: sarlak3@gmail.com
ORCID: https://orcid.org/0009-0002-0166-096X
Arezoo Haghbin
E-mail: haghbin33@gmail.com
ORCID: https://orcid.org/0009-0000-5285-1702

Received: 07.11.2025
Revised: 09.12.2025
Accepted: 25.12.2025

1. Introduction

Verbal Multiword Expressions (VMWEs), also known as complex predicates, complex verbs, or compound verbs, play a crucial role in natural language understanding, as they often convey meanings that go beyond the sum of their parts.

Persian is an Indo-European language with relatively few simple verbs and a large number of VMWEs. New verbs are typically formed by combining nouns or adjectives with a light verb (Karimi-Doostan, 2005). These constructions are referred to as "compound verbs" in Persian grammatical tradition. This language contains around 200 simple verbs (Mohammad and Karimi, 1992; Samvelian and Faghiri, 2013), most of which can function simultaneously as light and heavy verbs. Compound verbs have gradually replaced simple verbs since the thirteenth century. In some cases, the use of the simple verb is limited to written or elevated registers (Folli et al, 2005).
As mentioned, a Persian compound verb is formed by a light verb (LV) and a nonverbal element (NV). The NV precedes the LV (1-a) which has partly or entirely lost its original meaning. Different inflections including negation and future are normally prefixed or suffixed to the LV (1-b).

(1) a. *mæn* *dust-æm* *ra*[1] ***dæ'væt*** ***kærd-æm***.[2]
I friend-my ACC invite do.PST-1SG
'I invited my friend.'

b. *mæn* *dust-æm* *ra* ***dæ'væt*** ***næ-kærd-æm***.
I friend-my ACC invite NEG-do.PST-1SG
'I did not invite my friend.'

Identifying VMWEs in Persian poses several challenges, including a lack of recourses, morphological complexity, lexical variation, idiomatic expressions, and semantic ambiguity. Addressing these challenges requires linguistic resources along with innovative approaches that take into account the language's unique characteristics.

---

[1] - *ra* is Persian object marker.

[2] Persian examples are transliterated using a modified Latin-based system in which long vowels are represented as a, i, u; short vowels as æ, e, o; and the glottal stop is marked with an apostrophe (ʾ). In addittion, in the glossed examples, grammatical abbreviations are employed as follows: ACC (accusative), PRES (present tense), PST (past tense), SG (singular), PL (plural), NEG (negation), and PROG (progressive aspect).

In this article, we aim to present a framework for identifying Persian VMWEs in accordance with the guidelines of PARSEME Corpus. PARSEME (Parsing and Multiword Expressions) is a research network dedicated to studying multiword expressions. It aims to develop methods for their identification, analysis, and representation to enhance natural language processing tasks. PARSEME annotation largely relies on Universal Dependencies for the annotation of morpho-syntax, due to the shared objectives of universality. The annotation flow follows a decision diagram driven by linguistic tests. An expression is annotated as a VMWE if it can pass the tests on binary decision trees. By understanding VMWEs, which include idioms, collocations and phrasal verbs, PARSEME contributes to improving machine comprehension of language constructs. This corpus, in its version 1.3, contained the VMWEs of 26 different languages including Persian (Savary et al., 2023). The Persian part included 3617 sentences with VMWE tags which required revision based on both PARSEME guideline and Persian grammar, therefore, a framework was needed to enhance the accuracy of the previous annotations and add new sentences, thereby increasing the dataset size. In order to formulate the framework and annotate the Persian data, we studied the linguistic details of the Persian VMWEs, particularly the characteristics of the non-verbal elements, and identified the various forms they can take.

A Persian VMVE corpus, designed for a low-resource language, has significant implications for computational linguistics, natural language processing, and linguistic research. It provides valuable insights into Persian syntax, semantics, and pragmatics. As a foundational dataset, it assists in training algorithms for machine translation, enhancing their ability to handle idiomatic verbs and expressions to produce more accurate and natural translations. Moreover, it supports the development of tools for sentiment analysis, information extraction, and text summarization.

The present article is organized as follows: Section 2 provides an overview of previous research on processing Persian VWMEs. Section 3 outlines the steps involved in corpus development and explains the annotation process. section 4 introduces the annotation framework. Section 5 presents the findings and section 6 concludes the article.

2. Related Work

Studies on VMWEs comprise a significant part of Persian linguistic literature, among which some research focuses on the processing and detection of these expressions in Persian texts.

Iranpour Mobarakeh and Minaei (2009) focused on identifying verbs in Persian textual corpora. They mention that compound verbs consist of a light verb, which is morphologically similar to other verbs but differs from them

in terms of meaning, and a nonverbal element, which has no morphological suffixes—unlike the light verb. Hence, identifying the verbal part of compound verbs is relatively easy; however, the NV that precedes the verbal part complicates the identification of compound verbs. In the first step, the model developed in their research uses the structural information of Persian verbs to identify them. According to the derivational structure of Persian, its words can be stemmed systematically. The presented stem finder applies rules to extract verb stems and uses a vocabulary to improve the accuracy of the results. In the next step, using the n-gram approach, the disambiguation of co-occurring characters is discussed to enhance efficiency.

Rasooli et al. (2011) employ machine learning algorithms to identify Persian compound verbs. They highlight the diverse lexical semantics as the main challenging feature of the light verb constructions. They believe one of the difficulties of research in this field is the lack of reliable data sets for supervised and unsupervised learning. In their study, two unsupervised learning methods are applied to automatically identify Persian compound verbs. One is based on a modified version of the PMI concept, a method that calculates the probability of two words occurring together in the same corpus, known as PPMI**;** the other is based on the k-means clustering algorithm. They show that criteria such as PMI are insufficient for identifying Persian compound verbs, due to sparse data and the flexibility of word spacing in the Persian verbal construction. Using the k-means algorithm, they show that the average number of nouns between the NV and the LV significantly affects the identification efficiency of the compound verbs.

Mansoory et al. (2011) discuss the development of the Persian WordNet for verbs within FarsNet (Shamsfard, 2007), a semi-automated framework for building the Persian WordNet. They describe different types of Persian compound verbs, as well as the syntactic and semantic properties of each type. They then address the specific characteristics and behaviors of these verb types in order to develop a semantic lexicon. Finally, they present a method of using such linguistic properties in the automatic extraction of compound verbs and their relations from large text corpora and dictionaries with the aim of enriching the Persian WordNet of verbs. They treat compound verbs as lexical phenomena rather than syntactic ones; therefore, they can belong to synonym sets that include both compound and simple verbs.

Samvelian and Faghiri (2013) introduce PersPred as a manually annotated syntactic and semantic database for Persian compound verbs. They propose a framework for storing and describing Persian compound verbs. PersPred 1 contains more than 700 combinations of the verb *zædæn* (hit) with a noun, organized in a spreadsheet. Adopting a Construction-based approach, they study the way the productivity of Persian compound verbs can be explained despite their idiomaticity and the link generally established between compositionality and productivity.

In their study, Sarlak et al. (2023) propose both non-contextual and contextual methods to identify VMWEs. For the non-contextual strategy, they utilize a VMWE dataset based on Persian WordNet, created by collecting all compound verbs from FarsNet. They extract 21,462 VMWEs from FarsNet. To identify VMWEs in a sentence, they extract n-grams (n=2,3,4) and search for the presence of all components of a multi-word verb within the n-gram. However, not all cases found are VMWEs, particularly when intermediate words are present. For the contextual approach, their method addresses a sequence labeling problem aimed at recognizing the components thet comprise VMWEs. Through comparative analysis, the study evaluates two neural architectures, BiLSTM and ParsBERT (Farahani et al., 2021), for VMWE identification. Results indicate that a fine-tuned BERT model outperforms the BiLSTM model.

Eshaghi and Karimi-Doostan (2021) have developed a synchronic, monolingual corpus of 6000 Persian LVCs. They employ VBA macro codes to extract the LVCs consisting of 21 Persian LVs: *dashtæn*: have, *kærdæn*: do, *shodæn*: become, *gæshtæn*: turn, *gozashtæn*: put, *keshidæn*: pull, *didæn*: see, *dadæn*: give, *bakhshidæn*: give, grant, *gereftæn*: get, *yaftæn*: obtain, *amædæn*: come, *aværdæn*: bring, *residæn*: arrive, *ræftæn*: go, *oftadæn*: fall, *ændakhtæn*: throw, *bordæn*: take, *khordæn*: collide, *zædæn*: hit, and *bæstæn*: tie.

In the present paper, following a universal framework, we aim to investigate Persian VMWEs in a wider area comparing to the previous studies and focus on the nature of the non-verbal element in these expressions.

3. Corpus Development

The primary objective of this work was the *quantitative and qualitative expansion* of the PARSEME corpus. For the quantitative expansion, we *selected 2,000 sentences from PerUDT* (Safari et al., 2022) to add to the PARSEME Persian dataset, *based on their notable similarity in Universal Dependencies Treebank (UDT) structure and Part-of-Speech (POS) labels to the PARSEME corpus*. This deliberate expansion not only increased the corpus's volume but also enhanced its overall quality, enabling more comprehensive analysis and research.

The approach adopted in this study was grounded in Sarlak et al.'s (2023) model, specifically designed for the identification of Persian VMWEs within sentences. Utilizing their model, we accurately annotated PerUDT sentences to identify VMWEs. To ensure corpus balance, we included 1,500 sentences containing VMWEs and 500 sentences without any. Following the initial annotation process, we conducted a thorough review of each tag, making necessary corrections and categorizing each VMWE according to a defined set of guidelines. This included refining the initial PARSEME tags to

accurately designate the type of each VMWE, thereby ensuring the corpus's integrity for further linguistic research and applications.

The guidelines provided in this paper were instrumental, offering detailed descriptions and examples of VMWE types, which fostered a shared understanding among annotators and ensured consistency throughout the annotation process. To annotate the data, we employed FLAT2, a web-based annotation tool based on FoLiA format, which supports a wide range of linguistic annotations. This tool enabled us to efficiently revise and enrich the corpus with precise labels, ensuring the quality and accuracy of our linguistic data.

4. Guidelines

In this section, we provide a brief overview of the PARSEME guidelines, which form the basis for our categorization of Persian VMWEs.

PARSEME Categories of VMWEs:

1. **Universal categories**, valid across all languages in the corpus:
   a. Light verb construction (**LVC**)
      - The LV is semantically bleached (**LVC.full**) [to make a decision]
      - The LV adds a causative meaning (**LVC.cause**) [to cause a problem]
   b. Verbal Idioms (**VID**)
      - Verbal phrases with more than two components [to make ends meet]
      - Inflexible Verbal phrases with an adjective or an adverb as the NV [to come clean]
      - Verbal phrases with a non-eventive concrete noun as the NV [to give a hand]
      - Verbal phrases with a cranberry word as the NV [go astray]
      - Proverbs and conventionalized phrases [I beg your pardon]
2. **Quasi-universal categories**, valid for some languages in the corpus:
   a. inherently reflexive verbs (**IRV**):
      - Reflexives verbs in languages like French [s'évanouir]
      - Verbs including a reflexive pronoun [to help oneself]
   b. verb-particle constructions (**VPC**) with two subcategories:
      - the particle totally alters the verb's meaning (**VPC.full**) [to do in]
      - the particle adds a partly predictable meaning (**VPC.semi**) [to end up]

c. multi-verb constructions (**MVC**) [to let go]

Persian was one of the languages included in the PARSEME corpus, and based on its guidelines and Persian grammar, we aimed to categorize Persian VMWEs. As mentioned in Section 1, our primary dataset consisted of *3,617 sentences from PARSEME 1.2* and *2,000 sentences from PerUDT,* annotated according to separate instructions. Table 1 *presents the distribution of different VMWE types* found in this primary dataset.

Table 1: The number of different types of VMWEs in the primary dataset

	PARSEME	**PerUDT**	**Total**
Sentences	3617	2000	5617
VMWE	3450	1933	5383
LVC.full	3432	1933	5365
VID	17		17
IRV	1		1

As can be observed, the PerUDT dataset contained only LVCs, whereas PARSEME 1.2 included LVCs along with a small number of other VMWE categories. The following section outlines the challenges we encountered and the decisions we made in aligning the structural and semantic characteristics of Persian VMWEs with the categorization framework provided by the PARSEME guidelines.

4.1. Idiomaticity

Unpredictability or non-compositionality of the meaning is the most common criterion for defining idioms (Karimi, 1997). This concept, when applied to verbal expressions, exists on a continuum ranging from fully idiomatic to partially compositional verbs. In Persian, the lexical meaning of VMWEs is often not entirely predictable from the meaning of their individual component (Samvelian and Faghiri, 2013). Both the LV and the NV contribute distinct semantic roles**.** The pairing of a specific verb with a specific noun is frequently idiosyncratic, as there is often no clear semantic rationale for the selection of a particular verb (Samvelian and Faghiri, 2014).

(2) a. ***gush*** ***kærdæn*** (to listen)
ear doing
b. ***cheshm*** ***kærdæn*** (to hurt someone with an evil eye)
eye doing

As outlined in the PARSEME framework, the selection of NV plays a more decisive role in determining whether a construction qualifies as a VMWE.

Since semantic non-compositionality is difficult to assess directly, it is typically approximated through lexical and morpho-syntactic inflexibility. Accordingly, in example 3-a, the construction is classified as an LVC because the NV is an abstract noun. In contrast, example 3-b is categorized as a verbal idiom (VID) due to the presence of an adjective as the NV**.**

(3) a. ***hæds*** ***zædæn*** (to guess)
guess hitting

b. ***kutah*** ***amædæn*** (to give up)
short coming

Inflexibility refers to the inability to substitute the NV with similar lexical items while preserving the VMWE's meaning. For example, in 3-b, *bolænd amædæn* (lit. 'long coming')—as the opposite of *kutah amædæn* (lit. 'short coming')—does not yield a meaningful expression in Persian. This illustrates that the NV in *kutah amædæn* is not freely interchangeable, highlighting the idiomatic and fixed nature of the construction.

4.2. The Category of the NV

In Persian VMWEs, the NV is most often an abstract noun, as noted in the PARSEME guidelines for LVCs.

(4) ***pærtab*** ***kærdæn*** (to throw)
throw doing

Sometimes a preposition establishes the connection between the noun and the verb:

(5) ***be*** ***færamushi*** ***sepordæn*** (to forget)
to forgetfulness delivering

These verbal constructions are categorized as LVCs. However, other types of NVs are also used in Persian VMWEs. Constructions featuring these alternative NVs are typically labeled as verbal idioms in the PARSEME framework. In the following subsections, we examine the use of adjectives, adverbs, and concrete nouns as NVs in Persian VMWEs to reveal their structural and semantic properties.

4.2.1. Adjectives as the NV

Persian adjectives can be classified into two main types: simple adjectives and predicative adjectives.

4.2.1.1. Simple Adjectives

According to Persian grammar, adjectives can serve as the non-verbal element in Light Verb Constructions. For example:

(6) a. ***tæmiz*** ***kærdæn*** (to clean)
clean making
b. ***khamush*** ***kærdæn*** (to turn off)
off making

These types of verbal constructions are not classified as VMWEs in the PARSEME framework. Due to their flexible structure—where the NV can be substituted with similar adjectives while preserving the overall meaning—they do not meet the criteria for verbal idioms. Such constructions resemble expressions in English like *make somebody happy* **or** *make something clean***,** which denote a change of state rather than idiomatic meaning. Accordingly, we did not treat them as compound verbs in the Persian corpus either.

However, certain two-part verbal constructions consisting of an adjective and a verb exhibit inherently idiomatic meanings. In line with the PARSEME guidelines, these constructions are considered inflexible and are therefore categorized as VIDs. For example, in (7-a), *kæm aværdæn* (lit. 'little bringing') conveys an idiomatic meaning in Persian. Its hypothetical opposite, *ziyad aværdæn* (lit. 'much bringing'), does not form a meaningful expression, underscoring the inflexibility of the original construction. This lack of substitutability confirms its classification as a VID.

(7) a. ***kæm*** ***aværdæn*** (to give in)
little bringing

b. ***deraz*** ***keshidæn*** (to lie down)
long stretching

4.2.1.2. Predicative Adjectives

A subset of Persian adjectives, known as predicative adjectives, differs from standard adjectives in several key ways. These adjectives appear exclusively within the structure of VMWEs and are not used independently. They cannot be employed in comparative or superlative forms, do not function as noun modifiers, and cannot be intensified by degree modifiers (Karimi-Doostan, 2011). Below are two illustrative examples:

(8) a. ***færamush*** ***kærdæn*** (to forget)
forgotten making

b. ***mæhsub*** ***kærdæn*** (to consider)
considered making

The constructions **khaterat-e færamush* (lit. 'forgotten memories**'**) or **khaterat-e færamush-tær* (lit. 'more forgotten memories**'**) are ungrammatical in Persian.

Although these adjectives are not cranberry words in the strictest sense, given their morphological independence and clear semantic content, they are exclusively used within compound verb constructions. Therefore, we have classified them as cranberry words for the purposes of this study, and the corresponding verbal multiword expressions were counted as VIDs.

4.2.2. Concrete Nouns as the NV

Based on the guidelines, a non-eventive concrete noun cannot function as the NV in an LVC. However, if the concrete noun is eventive, i.e. if it denotes an event, it may be incorporated into an LVC. In 9-b the concrete noun ***jaru*** (broom) serves as the core of the construction by introducing the event as a whole.

(9) a. *maman* *otagh* *ra* ***jaru*** ***kærd-Ø.***
mom room ACC broom do.PST-3SG
'Mom swept the room.'

b. ***Jaru****-ye* *otaq* *do* *sa'æt* *tul* *keshid-Ø.*
broom-of room two hours length draw.PST-3SG
'It took two hours to sweep the room.'

In Persian, a light verb (LV) is frequently combined with the names of tools and objects to express the action performed using that tool.

(10) a. ***mesvak*** ***zædæn*** (to brush the teeth)
toothbrush hitting

b. ***telefon*** ***kærdæn*** (to make a phone call)
telephone doing

This is one of the productive mechanisms in Persian for forming multiword verbal expressions through the use of eventive concrete nouns. Consequently, such constructions are classified as light verb constructions.

Concrete nouns can be also used in the structure of VIDs. Our data indicated that 62% of concrete noun NVs in these constructions were body parts including eye, head, hand, foot, etc.

(11) a. ***æz*** ***pa*** ***dæramædæn*** (to be exhausted, to be killed)
of foot getting out

b. ***dæst*** ***shostæan*** (to be disappointed)
hand washing

In the same way as the adjectives, flexible cases with the sense of "making a change" including the ones in example 12 were excluded from the VMWE group, even though they are considered compound verbs in the traditional Persian grammar.

(12) a. ***sæng*** ***kærdæn*** (to change to a stone)
stone making

b. ***ard*** ***kærdæn*** (to change to flour)
flour making

4.2.3. Adverbs as the NV

In a small portion of the data, adverbs appeared as the non-verbal component of verbal multiword expressions. These constructions were also annotated as VIDs.

(13) a. ***pish*** ***amædæn*** (to happen)
forward coming

b. ***pæs*** ***oftadæn*** (to faint)
back falling

According to the PARSEME guidelines, verbal structures containing adverbs with directional meaning were not annotated as verbal multiword expressions.

(14) a. ***æghæb*** ***ræftæn*** (to go back)
back going

b. ***birun*** ***keshidæn*** (to pull out)
out pulling

4.3. Overlapping structures

In overlapping structures, a single LV or NV may occur in more than one verbal multiword expression. This may happen in the following cases.

4.3.1. Coordinate structures

There are instances in which an LV is implied rather than explicitly stated. In such cases, a single LV simultaneously serves multiple non-verbal elements, as illustrated in example 15.

(15) a. *Hæme-ye* *khane* *ra*
all-of house ACC
gærdgiri ***væ*** ***jaru*** ***kærd-æm.***
dusting and broom do.PST-1SG
'I dusted and swept the whole house.'

b. *dær* *anja* ***ærj*** ***væ*** ***mænzelæt*** ***dasht-Ø.***
in there esteem and dignity have.PST-3SG
'S/he had esteem and dignity there.'

4.3.2. An LV as a VMWE

There are Persian verbal multiword expressions in which the LV itself is a VMWE (Moloodi & Kouhestani, 2017). In such cases, the entire construction consists of three components, forming a layered multiword expression.

(16) $VMWE_1$ = NV + $VMWE_2$
kahesh ***[peyda*** ***kærdæn]*** (to decrease)
decrease found making

The tendency to form compound verbs in Persian has led to the coexistence of two sets of verbs, simple and complex, for a range of verbal concepts. This phenomenon is comparable to English, where both "decide" and "make a decision" convey the same meaning. For instance, the simple verb ***yaftæn*** (lit. 'to find**'**) functions as an LV in many verbal multiword expressions. Over time, however, it has been replaced by the compound verb ***peyda kærdæn*** (lit. 'found making**'**), which itself constitutes a VMWE.

yaftæn = **peyda kærdæn** (to find)

$[$kahesh yaftæn$]_{LVC}$ = [kahesh $[$peyda kærdæn$]_{VID}$ $]_{LVC}$ (to decrease)
decrease finding decrease found making

Peyda (found) functions as a predicative adjective. Therefore, in the construction ***kahesh peyda kærdæn*** (to experience a decrease), the expression contains a VID embedded within an LVC.

4.4. Agreement on the NV

As noted in Section 1, inflectional markers in Persian LVCs are typically prefixed or suffixed to the LV. However, there exists a unique type of verbal multiword expression in which an enclitic pronoun is attached to the non-verbal element. In these constructions, the subject agrees with the clitic in both person and number, while the verb consistently appears in the third person singular form (Rasekh-Mahand, 2014). The clitics involved are object clitics, distinct from the subject clitics found in standard simple or compound verbs. In this respect, the structure resembles reflexive verbs: it features an experiencer subject, lacks imperative forms, and cannot occur without the clitic. Moreover, these expressions rarely possess a usable infinitive form, or the infinitive fails to retain the same meaning. More precisely, they lack a prototypical form that would allow for straightforward classification as VMWEs.

Exceptional compound verb = NV + Object clitic + LV (3SG)

(17) a. *bæche* ***khab****-æsh* ***bord****-Ø*.
kid sleep-him take.PST-3SG
'The kid fell asleep.'

b. *mæn æz bagh* ***khosh****-æm* ***amæd****-Ø*.
I of garden good-me come.PST-3SG
'I liked the garden.'

c. *mæn* ***særd****-æm* ***æst***.
I cold-me be.PRES.3SG
'I am cold.'

d. *mærd* ***bavær****-æsh* ***shod****-Ø*.
man belief-him become.PST-3SG
'The man believed it.'

We decided to treat these constructions as standard VMWEs within the corpus. By omitting the clitic, we have an unusable combination of NV+LV as illustrated in the following examples, and we categorized them accordingly.

(18) a. ***khosh*** ***amædæn*** (to like) [VID]
good coming

b. ***bavær*** ***shodæn*** (to believe) [LVC]
belief becoming

4.5. Prefix verbs

In Persian, there are verbs that include prefixes. Their structure resembles that of particle verbs, except that the prefix is directly attached to the verb.

(19) a. ***bær-dashtæn*** (to pick up)
on-having

b. ***dær-gozæshtæn*** (to pass away)
in-passing

On the other hand, their behavior is more like LVCs in that other affixes, auxiliaries and clitics can appear between the prefix and the verb.

(20) a. ***bær****-mi-****dasht****-Ø*.
on-PROG-have.PST-3SG
'S/He was picking'

b. ***bær****-næ-****dasht****-Ø*.
on-NEG-have.PST-3SG
'S/He did not pick up'

These verbs are often regarded as a third category, distinct from both simple and compound verbs. In this corpus, they are classified as a subgroup of VMWEs, specifically termed verb-particle constructions (VPCs). This subgroup is further divided into two types:

1. VPC.full: the prefix totally changes the meaning of the verb.

(21) ***bær-dashtæn*** (to pick up)
on – having

2. VPC.semi: the prefix either does not add any meaning or adds a partly predictable meaning to the verb.

(22) ***bær-æfrukhtæn*** (to ignite)
on - igniting

4.6. Serial verb constructions

There are a few serial verb constructions in informal Persian, including:

(23) a. ***bezæn*** ***ber-im****!*
hit go-1PL
'Let's go!'

b. ***begir-æm*** ***bekhab-æm****.*

take-1SG sleep-1SG
'I'm gonna get some sleep.'

In these constructions, one verb functions as the main verb and carries the core semantic content of the combination (Anosheh, 2019). We chose to classify them as Multi-Verb Constructions (MVCs), However, no instances of serial verb constructions were found in our data, as this type of VMWE typically appears in informal contexts, whereas our corpus consisted of formal Persian sentences.

4.7. Reflexive verbs

Persian does not have reflexive clitics like those found in French. However, the pronoun 'oneself' can appear within verbal phrase structures. These instances were classified as Reflexive-Inflected Verbs (RIVs) under the VMWE framework.

(24) a. ***be*** ***khod*** ***amædæn*** (to come to one's senses)
to oneself coming

b. ***khod*** ***ra*** ***gereftæn*** (to be full of oneself)
oneself ACC taking

4.8. Passive voice

There are two mechanisms for forming passive constructions out of Persian transitive active VNWEs.

1. If the LV is ***kærdæn*** (to do, to make) or one of a few similar verbs, it is replaced by ***shodæn*** (to become) in the passive form and no auxiliary verb is used.

(25) active: ***dæ'væt*** ***kærdæn*** (to invite)
invite doing

passive: ***dæ'væt*** ***shodæn*** (to be invited)
invite becoming

In this group, the passive form was labeled as a VNWE if its corresponding active form met the criteria for a verbal multi-word expression, as illustrated in example (25).

2. In other cases, the verb ***shodæn*** is added to the verbal construction as a passive auxiliary and the LV changes to a past participle.

(26) active: ***ejare*** ***dadæn*** (to rent out)
rent giving

passive: ***ejare*** ***dad-e*** ***shodæn*** (to be rented out)
rent given becoming

In this group, only the NV and the past participle were tagged as a VMWE. Auxiliary verbs, like in other instances, were left unannotated.

Finally, we arrived at a classification scheme for Persian. In total, 5,617 VMWEs were tagged in our dataset. Of these, 83% were identified as LVCs, while the remaining 17% fell into other categories. The classification of verbal expressions, along with the percentage distribution of each type, is presented below.

1. **Light Verb Constructions** (LVC)
    - Verbal phrases with abstract nouns as the NV (98.4%)
    - Verbal phrases with eventive concrete nouns as the NV (1.6%)

2. **Verbal Idioms** (VID)
    - Verbal phrases with inflexible constructions and simple adjectives as the NV (13.18%)
    - Verbal phrases with inflexible constructions and adverbs as the NV (3.6%)
    - Verbal phrases with inflexible constructions and non-eventive concrete nouns as the NV (34.83%)
    - Verbal phrases with predicative adjectives as the NV (36.10%)
    - Verbal phrases with more than one NV (10.11%)
    - Proverbs and conventionalized phrases (2.17%)

3. **Verb-particle constructions** (VPC)
    - Prefix verbs in which the prefix totally changes the meaning of the verb (71.84%)
    - Prefix verbs in which the prefix adds a partly predictable meaning to the verb (28.16%)

4. **Reflexive verbs** (IRV)
    - Verbal phrases containing reflexive pronouns

5. **Multi-Verb Constructions** (MVC)

- Serial verb constructions

The following table shows the number of each type of VMWE and their subcategories in the final data.

Table 2: The number of different types of VMWEs in the final data

Category	**Subtype / Construction**	**Count (#)**	**Total**
LVC	Abstract N as NV	**4366**	**4434**
	Eventive concrete N as NV	**68**	
VID	Non-eventive concrete N as NV	**193**	**554**
	Simple Adj as NV	**73**	
	Adv as NV	**20**	
	Predicative Adj as NV	**200**	
	More than one NV	**56**	
	Proverbs (prov)	**12**	
VPC	Full	**227**	**316**
	Semi	**89**	
IRV	—	**10**	**10**
MVC	—	**0**	**0**

5. Analysis and Findings

This article outlined the development of a set of annotation guidelines to apply to a corpus of 5,617 sentences, encompassing various types of Persian

verbal multiword expression. Below, we present a summary of the key observations encountered during the annotation process.

- Persian contains a large number of verbal multiword expressions, the majority of which are light verb constructions.
- The verb ***kærdæn*** (to do, to make) is the most frequently used and the most productive Persian light verb.
- Most non-verbal elements used in forming Persian verbal multiword expressions are abstract nouns.
- After light verb constructions, verbal idioms are the second most frequent type of compound verbs in Persian, and their non-verbal components are typically adjectives or concrete nouns.
- The annotated data showed that 62% of concrete nouns in idiomatic verbal constructions were body parts like eye, head, hand and foot.
- Persian verbal multiword expressions exhibit different types of overlapping structures. In addition to coordinate structures (see Example 16), the data included three types of embedded structures:

  a. The light verb is an LVC

  (27) ***tæht-e*** ***tæ'sir*** ***[qærar*** ***dadæn]***$_{\textbf{LVC}}$ (to impress)
  under influence setup giving

  b. The light verb is a VPC

  (28) ***færyad*** ***[bær-aværdæn]***$_{\textbf{VPC}}$ (to yell)
  yell on – bringing

  c. The light verb is a VID

  (29) ***ronæq*** ***[peyda*** ***kærdæn]***$_{\textbf{VID}}$ (to prosper)
  prosperity found making

- Some flexible constructions, although considered VMWEs based on Persian grammatical criteria, were not annotated as such under the PARSEME framework:

  a. Adjective + verb

  (30) a. ***momken*** ***budæn*** (to be possible)
  possible being

b. ***khamush*** ***kærdæn*** (to turn off)
off making

b. Concrete noun + verb

(31) a. ***be*** ***gush*** ***residæn*** (to be heard)
to ear reaching

b. ***eynæk*** ***zædæn*** (to wear glasses)
glass hitting

c. Direction + verb

(32) a. ***æghæb*** ***ræftæn*** (to go back)
back going

b. ***birun*** ***keshidæn*** (to pull out)
out pulling

- Prefix verbs, which are typically treated as simple verbs in most corpora including the primary data of this study, were annotated as verbal multiword expressions due to their behavioral similarity to particle verbs.

- Persian has two mechanisms for forming the passive voice from compound verbs. If the light verb is *kærdæn* (to do, to make) or similar, it is replaced by *shodæn* (to become). If the light verb is something else, *shodæn* is added as an auxiliary verb.

6. Conclusion

In this study, we investigated Persian verbal multiword expressions (VMWEs) across a broad range. To this end, we followed the PARSEME corpus guidelines and adapted them to align with the grammatical features of Persian. Drawing on both language-specific properties and universal categories present in Persian, we developed annotation instructions and applied them to a corpus of 5,617 sentences. Additionally, we examined the characteristics of nonverbal elements across various types of Persian VMWEs.

7. Limitations & Future Work

We have obtained this instruction by studying articles on Persian VMWEs, PARSEME guidelines and examining the data of two corpora. There may still be cases not covered and be presented in future studies.

Acknowledgment

This work received advisory support from the CA21167 COST action UniDive, funded by European Cooperation in Science and Technology (COST).

BIBLIOGRAPHY

Anousheh, M. (2019), "Serial Verb Construction in Persian: A Minimalist Approach", *Journal of Researches in Linguistics*, vol. 11, no. 1, pp. 73–91. In Persian.

Eshaghi, M., and Karimi-Doostan, G. (2021), "The Productivity of Persian Light Verbs", *Journal of Language Researches*, vol. 12, no. 2, pp. 1–28. In Persian.

Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. (2021), "ParsBERT: Transformer-Based Model for Persian Language Understanding", *Neural Processing Letters*, vol. 53, pp. 3831–3847.

Folli, R., Harley, H., and Karimi, S. (2005), "Determinants of Event Type in Persian Complex Predicates", *Lingua*, vol. 115, no. 10, pp. 1365–1401.

Iranpour Mobarakeh, M., and Minaei-Bidgoli, B. (2009), "Verb Detection in Persian Corpus", *International Journal of Digital Content Technology and its Applications*, vol. 3, no. 1, pp. 58–65.

Karimi, S. 1997. "Persian Complex Verbs: Idiomatic or Compositional", *Lexicology*, vol. 3, pp. 273–318.

Karimi-Doostan, G. (2005), "Light Verbs and Structural Case", *Lingua*, vol. 115, no. 12, pp. 1737–1756.

Karimi-Doostan, G. (2011), "Separability of Light Verb Constructions in Persian", *Studia Linguistica*, vol. 65, no. 1, pp. 70–95.

Mansoory, N., Shamsfard, M., and Rouhizadeh, M. (2012), "Compound Verbs in Persian WordNet", *International Journal of Lexicography*, vol. 25, no. 1, pp. 50–67.

Mohammad, J., and Karimi, S. (1992), "Light Verbs Are Taking Over: Complex Verbs in Persian", in: *Proceedings of the Western Conference on Linguistics (WECOL)*, vol 5, ed. J. Nevis, and V. Samiian, Fresno: California State University, pp. 195–212.

Moloodi, A., and Kouhestani, M. (2017), "The Role of Metaphor and Metonymy in the Semantics of Persian Adjectival Preverbs: A Cognitive Linguistics Approach", *Language Art*, vol 2, no. 2, pp. 91–105.

Rasekh, M. (2014), "Persian Clitics: Doubling and Agreement." *Journal of Modern Languages*, vol. 24, no. 1, pp. 16–33.

Rasooli, M. S., Faili, H., and Minaei-Bidgoli, B. (2011), "Unsupervised Identification of Persian Compound Verbs", in: *Advances in Artificial Intelligence: 10th Mexican International Conference on Artificial Intelligence,*

MICAI 2011, Puebla, Mexico, November 26 - December 4, 2011, Proceedings, Part 1, ed. I. Batyrshin, and G. Sidorov, Heidelberg: Springer, pp. 394–406.

Safari, P., Rasooli, M. S., Moloodi, A., and Nourian A. (2022), "The Persian Dependency Treebank Made Universal", in: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, ed. N. Calzolari et al., Marseille: European Language Resources Association (ELRA), pp. 7078–7087.

Samvelian, P., and Faghiri, P. (2013), "Introducing PersPred, a Syntactic and Semantic Database for Persian Complex Predicates", in: *Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013),* ed. V. Kordoni, C. Ramisch, and A. Villavicencio, Stroudsburg: Association for Computational Linguistics, pp. 11–20.

Samvelian, P., and Faghiri, P. (2014), "Persian Complex Predicates: How Compositional Are They?", *Semantics-Syntax Interface*, vol 1, no. 1, pp. 43–74.

Sarlak, M., Yarandi, Y., and Shamsfard, M. (2023), "Predicting Compositionality of Verbal Multiword Expressions in Persian", in: *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, ed. A. Bhatia et al., Stroudsburg: Association for Computational Linguistics, pp. 14–23.

Savary, A., Khelil, C. B., Ramisch, C., Giouli, V., Mititelu, V. B., et al. (2023), "PARSEME Corpus Release 1.3", in: *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, ed. A. Bhatia et al., Stroudsburg: Association for Computational Linguistics, pp. 24–35.

Shamsfard, M. (2007), "Developing FarsNet: A Lexical Ontology for Persian", *Proceedings of the Fourth Global WordNet Conference / GWC (Szeged, Hungary, January 22-25, 2008)*, ed. A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, and P. Vossen, Szeged: University of Szeged, Department of Informatics, pp. 413-418

YEREVAN STATE UNIVERSITY PUBLISHING HOUSE

JIL 2/2 (2025): 108-116

Journal of Iranian Linguistics

An Examination of Two Proverbs in Khotanese and Their Equivalents in Certain New Western Iranian Languages

Majid Tame

The Acadmey of Persian Language and Literature

doi.org/ 10.46991/jil/2025.02.05

Abstract: Khotanese is a Middle Eastern Iranian language that was spoken in Khotan, located in the Xinjiang area of Turkestan, China, until the late 10th and early 11th centuries AD. The remaining Khotanese documents primarily consist of translations of Buddhist Sanskrit literature. One of the most significant poetic compositions in the Khotanese language is known as the *Book of Zambasta*. Unlike numerous other Khotanese texts that serve as translations of Buddhist literature, this particular work stands alone as an independent composition, although it is inspired by Buddhist texts. In this book, while explaining and interpreting religious teachings, references to stories and proverbs are occasionally made, the equivalents of which can be found in other languages, including Indian and Iranian languages. In the second chapter of the *Book of Zambasta*, two sentences address the themes of futile effort and exertion, which should be considered a type of proverb or maxim. In certain New Western Iranian languages, two proverbs akin to these Khotanese expressions remain prevalent, albeit with variations in vocabulary or usage. In this concise article, I will refer to these two proverbs in Khotanese along with their equivalents in certain New Western Iranian languages.

Keywords: Khotanese; Zambasta; proverb; Iranian languages

Majid Tame
E-mail: majidtameh@gmail.com
ORCID: https://orcid.org/0000-0002-1049-6607
Received: 30.08.2025
Revised: 14.11.2025
Accepted: 30.11.2025

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support

1. Introduction

Khotanese is classified as a Saka dialect and is regarded as a Middle Eastern Iranian language. Consequently, it shares a closer relationship with Middle and New Eastern Iranian languages rather than with Middle and New Western Iranian languages. The literary compositions in this language were primarily sourced from the Khotan of the Xinjiang Autonomous Region, located in Turkestan, China. Determining the exact period when the Khotanese Sakas settled in this region is challenging; nevertheless, it appears that the Khotanese language has been prevalent in this area since the third century AD (Emmerick 2009: 337). With the emergence and prevalence of the Muslim Turks in this region around the beginning of the 11th century, this language ceased to exist, and from this moment onward, there are no surviving remnants of this language (Ibid, 411). The corpus of texts that remain in Khotanese is considerably more extensive than that of numerous other Middle Iranian languages, with most of these texts being published by Bailey in a collection entitled 'Khotanese Texts'. Indeed, reliable and accurate translations are not available for all of these texts, and only a portion of them have been accurately edited and translated. Generally, these texts are categorized into two primary groups based on language: old and late, and into two principal groups based on style and method: literary and non-literary. The bulk of literary compositions are centered around the teachings of Buddha and the principles of the Mahayana school. Non-religious literary works cover a diverse array of topics. (Emmerick 1992: 4; Maggi 2009: 333, 404).

The *Book of Zambasta* represents the longest extant text in the Khotanese language. This literary piece is composed in Old Khotanese and consists of 24 chapters, some of which are incomplete, while one chapter has been completely lost. The probable timeframe for the composition of this book is approximately the mid-5th century AD. In contrast to other religious writings that serve as translations of Buddhist scriptures, this book is an original composition in Khotanese, composed with reference to Buddhist texts. This book serves as the primary reference for the meter used in Khotanese poetry. The manuscript of this document lacks a title; Bailey designated it as the *Book of Zambasta* because it references an individual named Zambasta, a Khotanese official, for whom the book was composed at his request (Emmerick 1990: 362; Maggi 2009: 348-350; Maggi and Martini 2014: 140-141, 157; Sims-Williams 2025: 15).

In both literary and educational writings, a prevalent approach to instructing and clarifying the content includes the use of similes and proverbs. Consequently, the use of similes and proverbs can be regarded as a defining feature of educational literature. In the field of rhetoric, an allegory

is regarded as an image or metaphor that conveys a message from the speaker, although the underlying intention may vary (Poornamdariyan 1988: 116). Furthermore, some allegories can communicate a tale, narrative, or proverb that holds a deeper significance hidden beneath its apparent meaning. The intent of recounting it is, in fact, to convey this this hidden meaning within the allegory (Shamisa 2002: 232). Allegory encompasses a wide range of interpretations and can include simile, metaphor, parable, proverb, and fable. Fundamentally, the main aim of allegory is to communicate an abstract and important idea or problem through a tangible and clear representation (Vafayi and Aqababayi 2013: 34). With respect to form and structure, parables can be classified into two categories: short and long. A short parable is generally expressed as a sentence, a verse, or a stanza, while a long parable is delivered in the form of a story or anecdote. In a short parable, the main topic or mental concept is first presented, succeeded by an objective and sensory depiction in the following sentence or verse, usually presented as a proverb (Mortezayi 2012: 34).

Proverbs, maxims, and idioms that are common across various cultures and nations play a crucial role in their language and literature, embodying the spiritual and moral traits, thoughts, ideas, customs, and practices of those cultures or nations. These proverbs have been developed and shared among various ethnic groups for millennia, gaining distinctive characteristics shaped by the ethnic and cultural features, customs, traditions, religions, and even the geographical and political contexts of each nation. Proverbs are generally concise expressions, but they successfully convey the intended meaning of the statement. A proverb can be generally described as follows: It is a brief, sometimes, poetic expression that incorporates a simile and imparts wisdom, having become well-known among the general public because of the fluidity of its wording, the lucidity of its message, and the gracefulness of its form. People often incorporate it into their conversations without any alterations (Zulfaqari 2007: 4). The main purpose of proverbs is to express intricate ideas in a concise and clear manner, as their straightforwardness and brevity enable them to summarize concepts that would otherwise require elaborate reasoning and justification for the audience to understand. Moreover, many proverbs trace their origins back to ancient times, thereby demonstrating the cultural continuity of a nation across its historical timeline. This continuity can be evidenced by proverbs and parables present in ancient texts that are still utilized in contemporary discourse.

Proverbs represent, on one hand, the distinct ideas and historical experiences of a specific people or nation, while on the other hand, they can emerge from the shared experiences of human societies. As a result, we

sometimes notice similar proverbs in different cultures. The proverbs that are comparable among different nations ought not to be regarded merely as adaptations; instead, they signify a type of unintentional coincidence. This indicates that they represent the shared experiences and concepts of various ethnic communities from distinct regions, which have been documented and communicated through language. In any event, sharing linguistic and cultural foundations may result in resemblances among the proverbs of groups that are part of the same language families. In this regard, we may refer to common origins or sources for these proverbs.

2. Discussion and review

In the *Book of Zambasta*, which was composed to explain the teachings of the Mahayana Buddhist sect, long parables are generally not presented as elaborate narratives or anecdotes. Rather, It is more common to employ short parables to clarify and explain Buddhist teachings. Instances of the short parables included in this book, which have been cited by scholars, include: "The Parable of the Elephant and the Dark House …" (Leumann 1933-1936: 24; Mirfakhrai 2005), and "This World Is a Mountain, and Our Actions the Shout" (Mirfakhrai 2007). The second chapter of the *Book of Zambasta* narrates the tale of a sorcerer named Bhadra, who attempts to deceive and surpass the Buddha through the use of magic; however, he ultimately fails and subsequently becomes a disciple of the Buddha. This chapter presents two statements about futile efforts that may be considered proverbs. The equivalents of these proverbs, though with slight differences, remain prevalent in certain New Western Iranian languages. In stanzas 15, 16, and 17 of this chapter, there are discussions about futile efforts, with stanza 16 notably containing two statements related to these futile efforts. Below, these three stanzas are provided along with their translation, which is based on the editing and translation of Emmerick (1968: 14-15) and Sims-Williams (2025: 32-33), albeit with slight modifications.

15 ttäte nä ttandrāma vicitra vrrata ttavaścaraṇa parāha
ku samu pharu karya u stāma [ne ju ye parśtä dukhyau jsa]

There are for them various such vows (vrata-), austerities (tapaścaraṇa-), restraints (śīlā-), in which there is much effort and exertion, [and yet one does not escape from sufferings].

16 kho ye siyato hvaittä bajsīha o ūtco maṃthäte kīśśa
kari ju vara rrūṇä ni byo[dä] [……..]

Just as when one pounds sand in a maortar or swirls water in a vessel, there is no oil (= butter) there at all …

17 ttrāmä ṣätä kūri parāhä biśśä karya stāma ttuśśīma cu ye ttarandari dukha tīndä [.....]

So this restraint is false when one makes sufferings for the body. All effort, exertion is empty …

The evidence provided above demonstrates that the Khotanese text utilizes two sentences to elucidate the concept of futile effort and exertion: *siyato hvaittä bajsīha* 'one who pounds sand in a mortar' and *ūtco maṃthäte kīśśa … rrūṇä ni byodä* 'one who swirls water in a vessel [to obtain butter], there is no oil (= butter)'. Considering the evidence of common proverbs in certain New Western Iranian languages that are similar to these two Khotanese sentences, we can view the proverbial expressions 'to pound sand in a mortar' and 'to make butter from water' as equivalents to these two Khotanese sentences.

As mentioned earlier, the *Book of Zambasta* is an independent literary work composed in the Khotanese language. Nevertheless, Sanskrit Buddhist literature was considered during its composition. Therefore, it is not unexpected that this Khotanese work includes proverbs and similes that are commonly found in Sanskrit. The proverbs referenced in stanza sixteen of the second chapter of the *Book of Zambasta* also have equivalents in Sanskrit. The proverb 'to make butter from water' is similarly present in Sanskrit as *jalamanthana*, which means 'churning the water'. It is acknowledged as *Jalamanthana-nyāya*, and this maxim can be translated broadly as 'churning the water does not yield butter'. Sims-Williams (2025: 57) proposes that the proverb 'to pound sand in a mortar' is similar to the Sanskrit maxim *sikatātaila-nyāya*, which means 'one cannot pound sand and get (sesame) oil'. Regardless of our viewpoint regarding the origins of these Khotanese proverbs, whether they stem from Sanskrit literature or not, the evidence found in certain New Western Iranian languages suggests that these concepts were indeed present among Iranian peoples. Therefore, it should not be presumed that they are simply adaptations from Indian culture.

3. Equivalents of Khotanese proverbs in certain New Western Iranian languages

As previously stated, both of the aforementioned Khotanese proverbs serve to convey the notion of futile exertion and effort. In Persian, however, the first proverb, 'to pound sand in a mortar', is used with a slight change to express the same notion of futile action and effort, taking the form of *âb dar hâvan kubidan* 'to pound water in a mortar'. It is important to note that, alongside

the expression *âb dar hâvan kubidan*, there are other phrases frequently used in Persian to express the idea of futile actions and efforts. These include *âb be rismân bastan* 'to tie water with a cord', and *âb be kolux bastan* 'to stop water with a lump of earth'. However, *âb dar hâvan kubidan* remains the most well-known proverb that expresses the concept of futile action and effort.[1] However, the second proverb, 'to make butter from water', has largely altered its usage in Persian. It is now utilized not to express the futility of effort or action, but rather to characterize a frugal and skilled individual who aims to gain from even the most minor things. Generally, this proverb refers to individuals who attain benefits and advantages in any circumstance through their intelligence and expertise. The following expressions serve as proof of the aforementioned two proverbs in Persian, as well as in several other New Western Iranian languages.

A) Equivalents of the proverb 'to pound sand in a mortar' in New Western Iranian languages

Persian: *âb dar hâvan kubidan* / *âb be hâvan kuftan* 'to pound water in a mortar', *âb dar hâvan sâyidan* 'to grind water in a mortar' (Azimi 2003: 3, 4).

Davâni: *ongori ow-i ke a-tu hovang hukoyen* 'As though water is pounded in a mortar' (Salami 1402: 403).

Qâ'eni: *ow hɛtu hava va mekuyɛde* 'One is pounding water in a mortar' (Meqdari 2012: 42).

Abuzeydâbâdi: *ow da rū yone-y-ä əkəri ho-w-äpəši* 'One pours water into the mortar and pounds (it)' (Razzaqi 2014: 124).

Gazi: *ow-â guye ru yâne de sef(t) keru* 'One wants to thicken water in the mortar' (Yazdani Gezi 2014: 19). The interpretation of this proverb

[1] It is important to highlight that within the folk literature of certain areas in Iran, a fabulous tale has been developed regarding 'to pound water in a mortar'. However, the meaning of this tale is not related to futile effort and exertion; rather, it aims to convey the psychological impact of negative inculcation on individuals. It is said that Plato and Aristotle had differing views on the purpose and effect of poisons, particularly the most lethal ones, leading each of them to attempt the formulation of a fatal poison. Plato formulated a lethal poison utilizing various substances, whereas Aristotle counteracted the effects of the poison by drinking milk and applying honey to his skin. Aristotle, without employing any particular substance, merely pounded water in a mortar for an extended period while feigning the making of poison. Conversely, Plato, who was unaware of the contents of the mortar, was perpetually anxious and fell ill after drinking the water that had been pounded in the mortar. Consequently, Aristotle proved that the effects of negative inculcation are significantly more perilous than any lethal poison (Beghaee 2002: 89; Darvishiyan and Khandan 2001: 445-446).

suggests that by pounding and stirring water in a mortar, it is impossible to produce a solid item.

Kermani: *âb tu juɢan kuftən* 'to pound water in a mortar' (Baghaee 2002: 89)

B) Equivalents of the proverb 'to make butter from water' in New Western Iranian languages

Persian: *az âb kare gereftan* 'to make butter from water'. Besides this idiomatic expression, another version of this proverb, *az âb roɢan gereftan* 'to make oil from water', is also prevalent in Persian. (Azimi 2003: 42). It is important to highlight that in the latter version of this Persian proverb, similar to Khotanese, the term 'oil' is used in place of 'butter'.

Targhi: *ov de kara a-gera* 'One makes butter from water' (Mohammadhosseini Targhi et al. 2023: 41).

Qâ'eni: *ɛz oʊ mɛskɛ megirɛde* 'One makes cream from water' (Meqdari 2012: 15).

4. Conclusion

Nevertheless, from an etymological standpoint, there are notable connections between Khotanese and Western Iranian languages regarding vocabulary, due to the conversion of the Khotan Sakas to Buddhism, there exists a reduced number of cultural themes in their remaining written works that are analogous and prevalent in the literature of Western Iranian languages. However, through a meticulous analysis and contemplation of the existing Khotanese texts, it is possible to discern common cultural and literary features. Within the current corpus of Khotanese literature, there are cultural and literary examples that are still employed by speakers of Iranian languages. This includes the allegories and proverbs present in Khotanese literature, which are still used in their original form or with modifications in New Iranian languages. The Khotanese proverb 'to pound sand in a mortar' and the Persian proverb 'to pound water in a mortar' both express the concept of a futile effort. Likewise, the Khotanese proverb 'to make butter from water' illustrates a futile exertion, while the Persian proverb *az âb kare gereftan* signifies cleverness and the ability to gain advantage from any situation. In the first proverb, the idea and significance of the Khotanese and Persian instances are identical, with the sole difference being the words 'sand' and 'water'. In the second proverb, although there is no significant difference in the meaning, the usage of the two varies; one signifies the futility of effort,

whereas the other indicates cleverness and profit-making. In any event, the key aspect regarding the two proverbs examined in Khotanese and New Western Iranian languages is their common origin. This common origin may arise from a shared cultural heritage among Iranian people, or it may be influenced by Indian culture and literature. Given that cultural interactions have been present between the Iranian and Indian people since ancient times, and Buddhism has also been prevalent among Iranians, it is probable that Indian culture has had an influence. Furthermore, given the extensive interactions between the cultures of Iranian people and other groups along the Silk Road, it is probable that these two Indian proverbs were introduced into Iranian languages through the literary and cultural exchanges that were prevalent along the Silk Road. However, it remains unclear when, how, and via which intermediaries Iranian people came to adopt these proverbs originating from Indian culture. In any case, additional examination of Khotanese texts, besides enhancing our comprehension of this language, can also offer more profound insights into the cultural and literary connections between the Khotan Sakas and the literary works of those who speak Western Iranian languages.

BIBLIOGRAPHY

Azimi, S. (1382/2003), *Farhang-e Bist Hezâr Maṭal va Ḥekmat va ʻeṣṭelâḥ* [*A Dictionary of Twenty Thousand Proverbs, Wisdom, and Idioms*], Tehran: Moʼasse-ye Moṭâleʻât ʻeslâmi-ye Dânešgâh-e Tehran. In Persian.

Baghaee, N. (1381/2002), *Amṯâl-e Farsi dar Guyeš-e Kermân* [*Persian Proverbs in Dialect of Kerman*], second edition, Kerman: Kermanology Center. In Persian.

Darvishiyan, A.A. and Khandan, R. (1380/2001), *Farhang-e Afsânehâ-ye Mardom-e Iran* [*The Folk Legends of Iran*], vol. 8, Tehran: Ketâb va Farhang.

Emmerick, R. E. (1968), *The Book of Zambasta: A Khotanese Poem on Buddhism*, London: Oxford University Press.

Emmerick, R.E. (1990). "Book of Zambasta", *Encyclopaedia Iranica*, ed. E. Yarshater, vol. IV, London: Routledge & Kegan Paul, pp. 361-363.

Emmerick, R. E. (1992), *A Guide to the Literature of Khotan*, Second Edition, Tokyo: The International Institute for Buddhist Studies.

Leumann, E. (1933-36), *Das nordarische (sakische) Lehrgedicht des Buddhismus: Text und Übersetzung (Abhandlungen für die Kunde des Morgenlandes 20)*, ed. M. Leumann, Leipzig.

Maggi, M. (2009), "Khotanese Literature", in: *A History of Persian Literature*: *The Literature of Pre-Islamic Iran:* ed. R. E. Emmerick and M. Macuch, London: I. B. Tauris, pp. 330-417.

Maggi, M. and Martini, G. (2014), "Annotations on the Book of Zambasta, III: Chapter 18 No More", *Scripta: An International Journal of Codicology and Palaeography* 7, pp. 139-158.

Meqdari, S. et al. (1391/2012). *Za'ferun Meṭqâl*: *Żarbolmaṭalhâ-ye Qâ'eni* [*Qa'eni Proverbs*], Qa'en: Akbarzade. In Persian.

Mirfakhrai, M. (1384/2005), "Tamṯil-e Pil va Xâne-ye Târik va Xerad-e Hameâgâh dar Še'ri be Zabân-e Xotani" ["The Parable of the Elephant and the Dark House and Omniscient Wisdom in a Poem in Khotanese Language"], *Nâme-ye Pârsi*, vol. 10, pp. 3-12. In Persian.

Mirfakhrai, M. (1386/2007), "In Jahân Kuh ast-o Fe'l-e Mâ Nedâ / Su-ye Mâ Aỷad Nedâhâ râ Sedâ" ["This World Is the Mountain, and Our Actions the Shout / the Echo of the Shouts Come back to Us"], *Farhang*, vol. 63-64, pp. 691-697. In Persian.

Mohammadhosseini Targhi, M. et al. (1402/2023). *Farhang-e Zabânzad-hâ-ye Targhi* [*Dictionary of Targhi Sayings*], Tehran: The Academy of Persian Language and Literature. In Persian.

Mortezayi, J. (1390/2012), "Tamṯil: Taṣvir yâ Ṣan'at-e Badi'i?" ["Allegory: Image or a Figurative Device?"], *Matnshenâsi-ye Adab-e Fârsi*, vol. 3(4), pp. 29-38. In Persian.

Poornamdariyan, T. (1367/1988). *Ramz va Dâstânhâ-ye Ramzi dar Adab-e Fârsi* [*Code and Secret Stories in Persian Literature*], Tehran: Elmi va Farhangi. In Persian.

Razzaqi, S.T. (1393/2014). *Frahang-e Amṯâl, Kenâyât va Eṣṭelâhât dar Guyeš-e Abuzeyedâbâdi* [*A Dictionary of Proverbs, Expressions, and Idioms in the Abuzeydabadi Dialect*], Tehran: Manshoor-Samir. In Persian.

Salami, A. (1402/2024). *Farhag-e Sine be Sine (Adabiyât-e Šafâhi-ye Mardom-e Davân)* [*Davâni Folk Literature*], Tehran: The Academy of Persian Language and Literature. In Persian.

Shamisa, S. (1381/2002). *Bayân* [*Rhetoric*], Tehran: Ferdows (in Persian).

Sims-Williams, N. (2025). *An Old Khotanese Reader: The Tale of Bhadra*, with contributions by J.A. Silk, Wiesbaden: Dr. Ludwig Reichert Verlag.

Vafayi, A. and Aqababayi, S. (1392/2013), "Barresi-ye Kârkerd-e Tamṯil dar Âṯâr-e Adabi-ye Ta'limi" ["The Study of Function of Allegory in the Didactic Literary Works"], *Pažuhešnāme-ye Adabiyāt-e Ta'limi*, no. 18, pp, 23-46. In Persian.

Yazdani Gazi, M.A. (1393/2014). *Eṣṭelâhât va Kenâyât va Ẓarbolmatal-hâ dar Zabân-e Gazi* [*Idioms, Expressions, and Proverbs in Gazi Language*], Isfahan: Kankash. In Persian.

Zulfaqari, H. (1386/2007). "Tafâvot-e Ẓarbol-maṯal bâ Barxi Gunehā-ye Zabâni va Adabi-ye Mošâbeh" ["The Difference between Proverbs and Some Similar Linguistic and Literary Forms"], *Najvâ-ye Farhang* 2(4), pp. 3-5. In Persian.