

Colloquial Persian has not incorporated broken plural morphology: Diglossia and code-mixing in Persian

Mohsen Mahdavi Mazdeh

This study examines an asymmetry between the status of broken plurals in Colloquial and Formal Persian. It is well known that Persian has borrowed many Arabic nouns along with their Arabic broken plurals, e.g., *kotob* is allowed as a plural form for *ketāb* 'book'. However, this study argues that once Colloquial Persian is clearly distinguished from Formal Persian, it can be seen that broken plurals are readily accepted only in Formal Persian while Colloquial Persian has largely resisted broken plurals, patterning with languages like Turkish rather than Formal Persian in this regard. To show this, this paper relies on several arguments including a statistical corpus analysis showing the rarity of true broken plurals in the regular speech of most speakers, the prevalence of broken plurals that are reanalyzed as singular (*aslehe* 'weapon'), the frequency of 'errors' wherein speakers pluralize broken plurals (e.g., *madārek-ā* 'document.PL-PL'), phonological factors, and comparison with other languages such as Turkish.

Keywords: Formal Persian, Spoken Persian, Informal Persian, Arabic, code-switching, morphology, bilingualism

1. Introduction

The formal and colloquial varieties of Persian coexist in Persian-speaking communities in a pattern of usage across social contexts that has often been described as diglossic.¹ In a linguistic community with diglossia, there are two varieties that differ not only in their lexicons but in various aspects of their grammars. This means that speakers have access to two (albeit largely overlapping) mental grammars. As with any bilingual speaker, a speaker with access to these two grammars can code-mix between them, creating a situation that, at least on the surface, blurs the boundaries between the two varieties. The relationship between code-mixing and diglossia has been studied in several linguistic communities, but no research has been conducted on the Persian case. Perhaps a major reason for this gap in research is the relative similarity of the formal and colloquial varieties of Persian, which makes cases of code-mixing too subtle and often easy to miss.

This paper focuses on the relationship between code-mixing and a specific morphological phenomenon: broken pluralization. Broken plurals are a special class of plurals in Arabic morphology that are created according to the root-and-pattern morphology of Arabic rather than affixation. They have been borrowed from Arabic into Persian since more than a millennium ago. I argue that while broken plurals have been properly adopted in Formal Persian and become part of the language, they have never been fully integrated into Colloquial Persian. In fact, even though there are words in Colloquial Persian that seem to have broken plural morphology on the surface, further scrutiny reveals that they are not used as true plurals in this language variety. This means that Colloquial Persian patterns more like languages such as Turkish rather than Formal Persian in this regard.

A core part of the argument presented in this paper relies on the observation that the use of broken plurals in the strict sense is in fact quite rare in Colloquial Persian conversations. The observation can be divided into the following three parts:

1. Words with broken plural morphology are relatively rare in the speech of people who have little familiarity with Formal Persian.
2. Of the instances where words with broken plural morphology do occur in these instances of speech, a large portion are morpho-syntactically not used as plurals.
3. Of the instances where these words do behave as plurals, a large portion do not have a corresponding singular form in the language.

Once we factor out the tokens that belong to the second and third items above, we realize that true broken plurals are quite rare in pure Colloquial Persian speech. We may then argue that given this rarity and other considerations such as phonological factors, the typology of broken plurals in languages that have been in contact with Arabic and Formal Persian, and the systematic repurposing of broken plurals as non-plural words in Colloquial Persian, it is more reasonable to assume the rare cases of true broken plurals in Colloquial Persian speech as involving code-mixing between Colloquial Persian and Formal Persian.

The remainder of this paper is organized as follows. Section 2 gives an overview of the diglossic situation in Persian and the social and linguistic differences between Colloquial Persian and Formal Persian. Section 3 lays out the theoretical foundation regarding borrowing, code-mixing, and the extent to which the two can be distinguished. After this setup, Section 4 discusses the case of broken plurals in Colloquial Persian in detail, partly relying on statistical evidence from a small corpus of Colloquial Persian speech. Finally, a concluding discussion is presented in Section 5.

2. Diglossia in Persian

In his seminal work on the phenomenon, Ferguson introduced diglossia as a situation where two varieties of a language are used in the same community in different social contexts, with one functioning as the higher register, reserved for more formal settings, and one functioning as the lower register used for everyday purposes.² Since then, a number of scholars have described the situation in Persian-speaking communities as diglossia.³ However, the gap between the two varieties in Persian is nowhere near as wide as the one in prototypical diglossic contexts such as the one seen in Arabic. This has led some scholars to deny the appropriateness of the term diglossia for the Persian case⁴ or to view Persian more or less as a marginal case.⁵

In treating the relation between Colloquial Persian and Formal Persian as diglossic, what matters for the purposes of the present paper is the fact that speakers have access to two different grammars and lexicons for Colloquial Persian and Formal Persian, regardless of how much overlap exists between them. In a sense, even if we treat the differences between Colloquial Persian and Formal Persian as a matter of stylistic choice rather than differences between two different language varieties, to the extent that we accept the existence of a meaningful distinctness in grammar and lexicon at the cognitive level, the core argument of the present study remains intact. What I wish to argue is that at the cognitive level, the grammar and lexicon of Colloquial Persian (whatever the sociolinguistic nature of Colloquial Persian might be) have not incorporated broken plurals.

In spite of what was said above, I do commit to the view that the Persian case counts as true diglossia. To justify this position, in addition to relying on existing literature, I present additional arguments in the remainder of this section. From this point on, I only focus on the linguistic situation in the Persian-speaking parts of Iran, although it seems that if anything, the gap between the colloquial and formal varieties is larger in Afghanistan and Tajikistan's Persian-speaking communities.

2.1. Usage of Formal Persian and Colloquial Persian

Historically, there has been a considerable degree of diversity among the Persian varieties used in everyday speech in different parts of Iran. In line with a common global trend, there has been a move towards convergence among these varieties in modern times as a result of the growth of literacy, transportation, and mass media. Nevertheless, the varieties of Persian spoken in different cities and villages are still distinct enough not only in their phonetics and phonology, but also in limited aspects of their lexicon and morpho-syntax. Among these varieties, the variety used by middle and upper class Tehranis holds a special position as it has gained standard status and is propagated through mass media. This standardized variety is no longer confined to Tehran and has become so prevalent across different regions of Iran in recent decades (not only in historically Persian-speaking areas, but also in other regions), especially among young urban

educated speakers, that I find it more appropriate to refer to it as “Standard Iranian Spoken Persian” rather than “Tehrani Persian”. For many young urban Iranians in different regions of the country, this is the only spoken language variety that comes naturally.

Within the context of Iran, the term “Colloquial Persian” (as well as “Spoken Persian” and “Informal Persian”) is sometimes used to refer to any or all of these varieties spoken in Iran, and sometimes exclusively to Standard Iranian Spoken Persian. In this paper, I use the term “Colloquial Persian” (henceforth CP) only in the broader sense and include data for it from speakers in two cities: Tehran and Shiraz. However, in giving examples and detailed descriptions of the linguistic phenomena, I take Standard Iranian Spoken Persian as my point of reference.

Compared to the formal variety, CP is by far the more common form of the language. It is used not only in all everyday conversations, but also in the vast majority of TV shows (including those involving academic and political discussions), and formal speeches. Even in written form, it has become more common than Formal Persian in certain domains such as social media posts and text messages. In recent decades, poetry in this language variety has also become increasingly popular.⁶ Thus, adjectives such as “colloquial”, “spoken”, and “informal” do not accurately capture the scope of usage of this variety, and must rather be seen merely as naming conventions.

Formal Persian (henceforth FP) is the prevalent written language variety in Iran. Virtually all Persian books, newspaper articles, school essays, and administrative documents are written in this variety. Even a simple written note to neighbors on the walls of an apartment complex is normally written in FP. The use of FP outside of writing is relatively limited. Hearing FP spoken does not happen every day for all speakers, but it is not a rare experience either. It happens regularly in various contexts such as adults reading story books to children, students or teachers reading for the class, and TV or radio presenters reading official announcements. Moreover, certain highly formal materials are produced in FP even when speech rather than text is the primary intended medium of presentation. The clearest examples of this are news broadcasts, historical dramas, and a small minority of formal speeches (although these usually involve plenty of code-switching).

Despite the clear differences between CP and FP, there is relatively little meta-linguistic awareness of the distinction among speakers. For a combination of linguistic and social reasons, many speakers tend to simply ignore the distinction or to view it as merely a matter of stylistic choice. This view is certainly not correct, not only because the differences are larger than what we normally see in the case of stylistic differences within a single language variety (see the next few subsections), but because the usage patterns of the two varieties in Persian-speaking communities cannot be accounted for through elements such as formality and prestige.

Regardless of whether we interpret ‘informality’ as familiarity, vulgarity, lack of sophistication, or any other related social concept, it is problematic to view an FP form such as *miravand* ‘they go’ as merely a more formal way of saying the same thing as the CP form *miran*. While levels of formality do largely overlap with the two varieties’ usage patterns (which is indeed what has led to the prevalence of the rather misleading term ‘Formal Persian’ in the first place), the choice between CP and FP often deviates from the division between ‘formal’ vs. ‘informal’ in its socio-linguistic sense. For instance, as mentioned earlier, it is possible to write informal sentences using FP. This is in fact what happens routinely in dialogues in many novels (both original Persian novels and those translated from other languages), where everything including conversations among peers, sentences involving vulgar jokes, and sentences attributed to children and uneducated people are written in Formal Persian.

The mismatch between language variety and level of formality occurs in the other direction too. Even in the most formal conversational settings such as academic discussions, official political debates, and contexts involving people speaking to someone of a higher rank in a social or organizational hierarchy, using CP is the norm. In such contexts, speakers may use educated vocabulary items as well as words and phrases that signal formality, e.g., *arz kardan* ‘to say (humbly)’, *be estelāh* ‘so to speak’, and *be lahāz-e* ‘in terms of’, yet consistently use CP forms, e.g., *miran* instead of *miravand* for ‘they go’.

The mismatches mentioned above, in both directions, clearly show that the difference between FP and CP is of a different nature than that of formal vs. informal styles within a single language variety, such as English “do not” vs. “don’t”. Of course, the relevance of formality in making a choice between CP and FP in Persian-speaking societies cannot be denied. However, the very fact that decoupling the two is possible shows that the difference between CP and FP cannot be reduced to levels of formality. This is most clear in the case of FP dialogue in novels. For instance, in a novel that is originally written in Persian and takes place in a Persian-speaking setting, when the utterances attributed to a child, an uneducated character, or in fact even a highly educated person, are written in FP, the Persian-speaking reader is not expected to imagine the character uttering those exact same FP sentence, because using FP in ordinary conversation would be highly unusual and in fact quite comical. Instead, these sentences are understood as FP translations of what the character might have said. This is a strong indicator of FP being a language variety that authors find suitable for books rather than a collection of stylistic choices within the same language variety.

2.2. Lexical differences

An important aspect of the discussion over the CP vs. FP distinction concerns the degree of difference between the two varieties. Mahmoodi-Bakhtiari (2018) lists many important differences, some of which are listed in what follows. The most obvious differences are lexical

differences in function words and bound morphemes. For instance, a large portion of the personal pronominal clitics (e.g., *-esh* vs. *-ash'* for 3SG), verbal agreement suffixes (e.g., *-in* vs. *-id* for 2PL), discourse particles (e.g., *mage* vs. *magar* for leading questions), and present stems of the most common verbs (e.g., *ār* vs. *āvar* for 'bring') differ between the two varieties. Similarly, there are lexical differences in content words (e.g., *kalle* 'head' is exclusive to CP). There are also important lexico-phonological differences between the two language. For instance, CP has a preference for *un* and *um* over *ān* and *ām*, e.g., *khune* vs. *khāne* 'house'. Another salient instance is that in CP, *-e* changes to *-a* before object markers, copulas, and personal pronominal clitics, e.g., *khune* 'house' vs. *khuna-ro* 'house-ACC'. Each of these differences may be small on its own, but they are numerous enough that in any given FP text, almost any sentence or phrase is different from its corresponding CP version.

2.3. Phonetic and phonological differences

Since we are particularly interested in differences between the grammars of the two language varieties, the more important differences are structural differences in the sound system, morphology, and syntax. In phonetics and phonology, perhaps the most important point of divergence between CP and FP is the status of vowel length. A vowel length distinction between the long vowels *ā*, *u*, and *i* is fully active in FP as evidenced by phonetic experimental work,⁸ but has largely disappeared in CP.⁹ At the phonological level, the difference in vowel length manifests itself most clearly in the versification systems of the two varieties. The two sets of vowels normally have different moraic lengths for metrical purposes in FP poetry. On the other hand, in CP poetry, the long vowels are treated as short in most contexts.¹⁰ Interestingly, there is a phenomenon in CP poetry that relates directly to this paper's argument regarding code-mixing. In words that inarguably belong to CP (e.g., function words, conjugated verbs, and CP-exclusive lexical items), keeping a length distinction between the two vowel types is rare. On the other end of the spectrum, when highly formal words are occasionally used in CP poetry, the length distinction is usually observed.¹¹ This suggests that in these cases, we are dealing with code-mixing between the two varieties.

2.4. Morphological differences

Most of the differences between the morphological structures of CP and FP seem to have been neglected in the literature. Mahmoodi-Bakhtiari mentions the adjective-producing suffix *-u* (e.g., *tars-u* 'coward' from *tars* 'fear') and the adverb-producing suffix *-aki* (e.g., *rāst-aki* 'really' from *rāst* 'true') as suffixes that are exclusive to CP.¹² There are also several suffixes that seem to be exclusive to FP and appear in CP only as part of FP loanwords, e.g., the agentive suffix *-ande* as in FP *āvar-ande* 'bringer' and *shenav-ande* 'listener' and the FP nominalizer *-esh* as in *khān-esh* 'interpretation' and *bin-esh* 'vision'. Crucially, exclusively CP forms cannot attach to these suffixes, e.g., **ār-ande*, **khun-esh*.

Many morphological differences are quite subtle and usually go unnoticed. One such example is how possessive compounds, especially those involving body parts, are made in the two varieties. The preferred construction is noun-adjective in CP, but adjective-noun in FP, e.g., FP *zibā-ru* (beautiful-face) ‘beautiful’, *qavi-del* (strong-heart) ‘brave’, *sakht-pust* (hard-skin) ‘crustacean’, *tang-dahān* (narrow-mouth) ‘small-mouthed’, *sorkh-mu* (red-hair) ‘red-haired’, and *siāh-chashm* (black-eye) ‘black-eyed’ vs. CP *mu-qashang* (hair-beautiful) ‘having beautiful hair’, *del-gonde* (heart-big) ‘brave’, *pust-koloft* (skin-thick) ‘thick-skinned’, *dahan-goshād* (mouth-wide) ‘wide-mouthed’, *mu-qermez* (hair-red) ‘red-haired’, and *chesh-sabz* (eye-green) ‘green-eyed’. The effect may not be obvious at first sight because some of the FP forms are used in CP too, especially in formal contexts. However, these should probably be seen as the result of borrowing or code-mixing. Nevertheless, there are even clear A-B vs. B-A pairs which make the difference easy to see, e.g., as *mu-sefid* (hair-white) vs. *sepid-mu* for ‘gray-haired’ and *qad-boland* (height-long) vs. *boland-qad* for ‘tall’.

A clear case of difference between CP and FP morphology can be seen in pluralization patterns. The default pluralization morpheme in FP is the suffix *-hā*, e.g., *ketāb* ‘book’ is pluralized as *ketāb-hā* ‘books’. The similar-sounding cognate of this suffix is used in CP too, where it normally appears as *-ā* after consonants and most vowels in CP, e.g., *ketāb-ā* ‘books’. However, FP has another native Persian pluralizing suffix too, i.e., *-(g)ān*, that is used almost exclusively for living things and body parts, e.g., *zan-ān* ‘women’, *nevisande-gān* ‘writers’, *dast-ān* ‘hands’. For living things, the use of this suffix is a completely productive process in FP. As new words are created for new concepts, especially if they end with native FP morphemes, *-ān* pluralization is completely natural, e.g., *veblāg-nevis-ān* (blog-write-PL) ‘bloggers’. This suffix is not an active pluralizing morpheme in CP. The relatively rare cases where it is used usually involve a limited set of words and must therefore be viewed as direct FP influence. For instance, the word *hezār-ān* ‘thousands’ is sometimes used in CP speech, but forms such as *zan-ān* ‘women’, *nevisande-gān* ‘writers’, and *dast-ān* ‘hands’ are completely unnatural as true plural forms in regular CP speech (see Section 4 for a discussion of what is meant by ‘true’ plurals in this context). Thus, *-ān* suffixation must not be seen as a productive process in CP. Instead, forms like *hezārān* must be seen as wholly taken from FP, perhaps even as an atomic unit.

There are three other pluralizing suffixes, all borrowed from Arabic, used in FP. Each of these is used for a fixed set of words, almost all of which are of Arabic origin. These are *-āt* (e.g., *moshkel-āt* ‘problems’), *-un* (e.g., *rowhāni-un* ‘clerics’), and *-in* (e.g., *mo’allem-in* ‘teachers’). These morphemes, especially *-āt*, appear occasionally in CP too. Discussing the extent to which they are incorporated into CP morphology is beyond the scope of this paper and not essential for our main goal.

Finally, FP has borrowed a large amount of plural forms with broken plural morphology from Arabic. For instance, it is perfectly normal to pluralize *jesm* ‘object’ as *ajsām* in FP. The main claim

of the present paper is that while these words are used as true plurals of their corresponding singular forms in FP, they have not been incorporated into CP in the same manner and their occasional appearance in CP speech must be viewed mostly or entirely as cases of code-mixing. I return to this issue in Sections 4 and 5.

2.5. Syntactic differences

We may move on to syntax now. The following list contains the most important syntactic differences mentioned by Mahmoodi-Bakhtiari.¹³ Note that the technical details of some of these phenomena are presented slightly differently here.

1. CP has two additional tense-aspect combinations; the present progressive and the past progressive, e.g., *dār-am mi-r-am* (have-1SG PROG-go-1SG) ‘I am going’.
2. CP has a uniqueness marker (stressed *-e*) for noun phrases, e.g., *ketāb-e* ‘the book’.¹⁴
3. CP usually places argument Prepositional Phrases after the verb, e.g., *oftād-am tu tale* (fall.PST-1SG in hole) ‘I fell into a trap’.
4. In CP, many Prepositional Phrases that appear with *dar* ‘in’ or *be* ‘to’ in FP appear without a preposition, e.g., *madrese na-raft-am* (school NEG-go.PST-1SG) ‘I didn’t go to school’ (cf. FP *be madrese na-raft-am*).
5. In CP, prepositions are allowed to attach to personal pronominal enclitics, e.g., *az-am* ‘from me’.
6. CP allows nouns and adjectives to attach (without an *ezafe*) and behave like a single noun receiving a definite or plural marker, e.g., *xodkār ābi-ā* (pen blue-PL) ‘blue pens’.

To this list we may add several other morpho-syntactic differences, as listed below.

7. CP uses the so-called object marker *-(r)o* extensively for topicalization, e.g., *ali-ro bā-hāsh harf zad-am* (Ali-ACC with-3SG speech hit.PST-1SG) ‘As for Ali, I talked to him’.
8. In CP, the indefinite marker *-i* cannot appear by itself in regular upward-entailing environments not followed in the absence of adjectives and relative clauses, e.g., **sib-i khord-am* (apple-INDF eat.pst-1SG) Intended: ‘I ate an apple’ (allowed in FP). In such environments, the numeral *ye* ‘one’ is used before the Noun Phrase instead.
9. CP allows clitic doubling in accusative sentences, e.g., *ali-ro did-am-esh* (Ali-ACC see.PST-1SG-3SG) ‘I saw Ali’.
10. FP has a separate future tense, e.g., *khāh-ad raft* (want-3SG go.PST) ‘will go’.

11. In FP, the anaphor morpheme *khod* is used without any attaching personal pronominal clitics (e.g., compare with CP *khod-am* ‘self-1SG’ for ‘myself’) and, crucially, has a different binding behavior in embedded environments.¹⁵
12. CP allows personal pronominal enclitics to attach to *wh*- words, e.g., *kojā-sh* ‘where in it’, *kodum-etun* ‘which of you’.
13. In CP, numerals preceding Noun Phrases always require a classifier (the default classifier is *tā*), e.g., *se tā ketāb* (three CLF book) ‘three books’ (cf. FP *se ketāb*).
14. In FP, the subjunctive prefix *be-* is omitted in a wide range of complex predicates (not limited to those involving the light verbs *kardan* ‘to do’ and *shodan* ‘to become’ as in CP), e.g., *edāme dah-im* (continuation give-1PL) ‘that we continue’ is at least as common as *edāme be-dah-im* in FP (the former pattern is disallowed in CP).
15. CP has an optional 3SG personal verbal agreement suffix, e.g., *raft-esh* (go.PST-3SG) ‘she/he/it went’.

The above list is certainly not comprehensive, but it is hopefully sufficient to give a sense of the type and diversity of the syntactic differences between the two varieties. The main argument I wish to make based on the above points is that we are indeed dealing with two language varieties that differ not only in the shapes of the morphemes but in their abstract structures including the role of vowel length, the ordering of morphemes within words, the ordering of words within sentences, definite and indefinite marking, tense-aspect combinations, agreement, noun-adjective interaction, and anaphor binding. Thus, it should come as no surprise that the two language varieties also have differences in the way they handle plurals.

3. Code-mixing and borrowing

In this section, I aim to clarify what I mean by ‘code-mixing’ and how it relates to diglossia in Persian. In this discussion, I use the term code-mixing in a broad sense, such that it encompasses what is traditionally referred to as ‘code-switching’ as well. In general, any instance of speech that uses the grammatical or lexical resources of more than one language variety (as they are stored and processed in the mind of the speaker) counts as code-mixing under this definition.

Code-mixing (or code-switching, for that matter) must be viewed separately from borrowing. In both code-mixing and borrowing, elements that originally belonged to one language appear in the context of another language. However, this surface phenomenon can be the result of either of two distinct processes.¹⁶ Focusing on the simple case of using a single word from one language in a sentence in another language, we may describe the difference as follows. In a prototypical case of borrowing, the word is fully incorporated into the lexicon of the host language. In code-

mixing, on the other hand, a form belonging to the lexicon and grammar of the donor language is used in a sentence otherwise created using the lexicon and grammar of the host language.¹⁷ Some scholars view the distinction primarily as a diachronic one, i.e., whether the host language has incorporated the new lexical item into its lexicon. From this perspective, code-mixing and borrowing form a spectrum. However, we are interested here in a categorical and synchronic conception of the distinction. As such, the question becomes whether the speaker uses an alien lexicon and grammar in the process of sentence formation or uses the lexicon and grammar of the host language exclusively.¹⁸ At least in its most extreme form, this view assumes that the two languages are meaningfully distinct in the mind of the speaker.¹⁹

Identifying cases of code-mixing evidently becomes more difficult when the two languages involved are very close. A valuable source for determining lexical items and structures that belong exclusively to one variety is to look at the speech of individuals who speak only one of the two language varieties. However, in diglossic situations, the issue becomes challenging since it is sometimes extremely hard, if not impossible, to find speakers who speak one of the two varieties only. Code-mixing between two varieties involved in diglossia has been studied in the case of several languages including Arabic,²⁰ French and French-influenced Creoles,²¹ and Italian,²² but no such work has been done on Persian.

Once we recognize CP and FP as two distinct language varieties, we must acknowledge the possibility of borrowing and code-mixing between the two. Cases of CP words used in FP are relatively rare and often noticeable, because using CP forms in FP, especially in writing, is generally frowned upon and seen as a sign of a low degree of literacy. Nevertheless, this does occasionally occur, especially with proper nouns that use CP morphemes. For instance, the fictional character ‘Little Prince’ from de Saint-Exupéry’s 1943 book is known as *shāzde-kuchulu* (prince-little) in FP even though both of the morphemes in the word can only appear in CP by themselves. Similarly, ‘big-foot’ is normally translated as *pā-gonde* (foot-big) in FP even though *gonde* ‘big’ by itself is normally allowed only in CP. As discussed earlier, when CP words are used in FP poetry, their long vowels often behave as short vowels, thus revealing their ‘alien’ origin. Mahdavi Mazdeh cites examples of this happening in FP poems with CP words such as *kuchulu* ‘small’, *tu* ‘in’, *bābā* ‘dad’, and *ākhund* ‘Shia clergyman’.²³

The use of FP words in CP is often harder to detect. The most obvious cases are conjugated verbs from FP that appear in frozen expressions in CP, e.g., *be-mān-ad* (SBJV-remain-3SG) is the regular 3SG subjunctive form of ‘to remain’ in FP, but is used as an interjection roughly meaning ‘let it remain unsaid’ in CP. The regular 3SG subjunctive form of ‘to remain’ in CP naturally uses the CP agreement suffix and follows CP phonology: *be-mun-e* (SBJV-remain-3SG).

Among other lexical categories, FP forms in CP are most reliably detectable in morphemes that preserve FP phonology in compounds. For instance, the word for ‘language’ is *zabān* in FP, but

zabun in CP (we should probably count the former form as a valid alternative in CP too). But the word for ‘linguist’ is *zabān-shenās* (language-discern) in both CP and FP. The fact that **zabun-shenās* is not acceptable in CP suggests that the compound is not built ‘natively’ in CP based on CP morphology and CP lexical items, but is taken as a whole from FP. Other examples from this large group of compounds include *mehmān-dār* (guest-have) ‘flight attendant’ (**mehmun-dār* is disallowed even though the CP word for guest is *mehmun*), *riāzi-dān* (math-know) ‘mathematician’ (**riāzi-dun* is disallowed even though the CP form of the stem for ‘to know’ is *dun*), and *ashk-āvar* (tear-bring) ‘tear gas’ (**ashk-ār* is disallowed even though in CP the stem for ‘to bring’ appears as *ār*).

Among the examples listed above, it is not immediately clear which ones count as borrowings and which ones as cases of code-mixing. The major criteria often used for distinguishing between borrowing and code-mixing, besides the socio-linguistic question of whether the word is in fact widely known by speakers of the host language, is the level of integration of the word into the host language’s phonetics, phonology, and morphology. For instance, the facts that *riāzi-dān* ‘mathematician’ has the sequence *ān* (rather than *un*) and is pronounced with a long *ā* in the first morpheme both count as evidence suggesting that this word is not integrated into CP and that therefore its occasional appearance in CP speech must be viewed as code-switching. However, admittedly, neither of these phonological arguments is definitive. The sequence *ān* (even within a syllable, which is the environment where the constraint seems to be stronger) does appear in words that are relatively common in CP too, e.g., *zamān* ‘time’. Vowel length is not an easily accessible diagnosis tool for such cases either, because except in the most extreme cases, CP allows both short and long realizations of the three traditionally ‘long’ vowels (although presumably with varying degrees of acceptability depending on the word), meaning that statistical information is needed to arrive at a reliable conclusion. We come back to the question of how borrowing can be distinguished from code-mixing in the following sections.

4. The adaptation of broken plurals

Broken plurals are plural forms in Arabic that are created through root-and-pattern morphology rather than affix concatenation. Arabic has several templates that create broken plurals. The template used for any given noun is fixed. The phonological shape of the singular form can to some extent predict the template that is used for broken pluralization, but it is not definitive. Formal Persian has adopted a large number of Arabic nouns along with their broken plural forms, and nearly all of the common pluralization templates of Arabic are represented in FP too. Examples of some of the most common patterns as used in FP are shown in (1). Note that in some of the examples, the words have consonants that are not part of the triconsonantal root. I skip

over these details of Arabic morphology as they are irrelevant to the main discussion of the paper.

(1)	Singular	Plural	Template	Gloss
	<i>fekr</i>	<i>afkār</i>	aCCāC	‘thought’
	<i>ketāb</i>	<i>kotob</i>	CoCoC	‘book’
	<i>vajh</i>	<i>vojuh</i>	CoCuC	‘aspect’
	<i>khādem</i>	<i>khadame</i>	CaCaCe	‘servant’
	<i>sānehe</i>	<i>savāneh</i>	CavāCeC	‘incident’
	<i>haqiqat</i>	<i>haqāyeq</i>	CaCāyeC	‘truth’
	<i>manzel</i>	<i>manāzel</i>	maCāCeC	‘abode’

In Arabic, broken pluralization is an active part of the morphology and even European loanwords often go through it, e.g., *aflām* ‘films’. In FP, however, this is for the most part not true. With only a few exceptions, broken plurals in FP are limited to the items that were borrowed from Arabic. Nevertheless, broken plurals are true plurals in FP; not only do they require plural agreement in the syntax, they seem to be in systematic relationship with their singular forms (more on this at the end of this section).

Among the non-Semitic languages that have been exposed to Arabic broken plurals, perhaps no other language has adopted broken plurals as much as Formal Persian has. A well-known example of a language influenced by Arabic but lacking broken plurals is Turkish. An enormous amount of Arabic nouns have entered Turkish. Among these, some are in fact Arabic words with broken plural morphology. However, they have been repurposed as singular nouns or even adjectives.²⁴ A few examples are shown in (2).

(2)	Word	Turkish Meaning	Arabic Singular	Meaning of Arabic singular
	<i>ahbap</i>	‘friend’	<i>ḥabīb</i>	‘beloved’
	<i>elbise</i>	‘clothing, dress’	<i>libās</i>	‘clothing’
	<i>eşya</i>	‘object’	<i>shay’</i>	‘object’
	<i>akraba</i>	‘relative (n)’	<i>qarīb</i>	‘close’
	<i>tuhaf</i>	‘strange’	<i>tuhfa</i>	‘rarity’
	<i>acayip</i>	‘strange’	<i>‘ajīb</i>	‘strange’

The Turkish case is interesting because it draws our attention to the fact that a language may categorically reject broken plurals as true plurals in spite of accepting many words with apparent broken plural morphology. In fact, Turkish is probably not alone in this respect. Most modern Iranian languages seem to be similar in this regard (although no systematic study of broken plurals in these languages exist). For instance, in Mazandarani, as far as the author can tell as a speaker (in the absence of a corpus study), in normal conversations that do not involve high degrees of code-mixing with Persian, broken plurals do not appear as true plurals. Another

example is standard Sorani Kurdish, where true broken plurals are virtually non-existent. Thackston's reference grammar does not refer to broken plurals in his description of Sorani Kurdish pluralization, and in the 60 pages of diverse authentic sample Kurdish texts he offers, there is not even a single true broken plural.²⁵

The analysis proposed in the present paper is that when it comes to broken plurals, CP is in the same group as Turkish, Sorani Kurdish, and Mazandarani rather than FP. Admittedly, true broken plurals do occur in CP speech, especially in the speech of educated speakers. However, I argue that these must be viewed as cases of code-mixing. To support this claim, I rely on six major observations: 1) Broken plurals are relatively uncommon in the CP speech of speakers with limited FP exposure, 2) as a method of pluralization, they are optional, 3) many broken plurals have been repurposed as singular, 4) speakers regularly make 'mistakes', using broken plurals that are standardly considered to be plural as singular, 5) broken plurals in CP that have plural meaning normally do not correspond to any singular forms, and 6) true broken plurals in CP follow FP phonology. These claims are discussed in detail in the following subsections.

4.1. Broken plurals in CP are uncommon and optional

To see how rare broken plurals are in CP, we need statistical data. The issue is complicated by the fact that code-mixing between CP and FP is quite common. Ideally, the speech of a CP speaker with no knowledge of FP could teach us a lot about CP. However, virtually all Iranian Persian speakers have some familiarity with FP. Adult literacy rate in Iran is reported to have been around 89% in 2023.²⁶ Crucially, even illiterate Persian speakers (both adults and children) normally have some degree of proficiency in FP. Materials in FP are often heard in media (see Section 2) and for centuries, there has been a rich tradition of reciting and memorizing Classical Persian poetry in Iran, even among uneducated people.

Given what was said above, it is almost impossible to find CP speakers who have no familiarity with FP. Our second best option is to look at the speech of Persian speakers who have limited or no formal education and are therefore less influenced by FP. For this, I was able to find four different sources that could be used for a corpus study, which are listed below. The statistical analyses throughout this section are based on these four sources.

1. Interviews with Shaban Jafari: Sha'ban Ja'fari (born in 1921 in Tehran, often known by the derogatory nickname *Sha'bān bimokh* 'brainless Sha'ban') was a practitioner of the Iranian traditional athletic and martial arts tradition (*zurkhāne* athletics or *varzesh-e bāstāni*) in Tehran who became famous for his leading role in a series of acts of political violence during the 1940s and 1950s. He was interviewed by Homā Sarshār around the year 2000 and a linguistically faithful presentation of the interviews was published by Sarshār.²⁷ Ja'fari's words in the first 223 pages of this book constitute one of the sources used in this paper. Ja'fari had only four years of schooling, but had some familiarity with

Persian poetry as evidenced by the poems he cites occasionally during the interviews. The interviews mostly involve politics, meaning that many high-register words appear in the data.

2. Mashq-e Shab: Abbās Kiārostami’s 1989 documentary Mashq-e Shab²⁸ (‘homework’ lit. ‘night practice’) contains interviews about homework with primary-school children in a school in Tehran. Since primary-school children do not have a large amount of experience in FP, I have used their speech as one of my sources for CP speech with minimal FP influence.

3. Interview with Abbās Barzegar: Abbās Barzegar (born in 1973 in Fars province) is an Iranian villager who became famous after establishing a successful tourist residence in the Fars region. He had two to three years of schooling. His speech in a 40-minute interview aired on the Iranian state TV in the 2010s²⁹ is used as a source of CP data in this research.

4. Bachoy Shahr: Iranian anthropologist Nasir Farrokhmehr has created a collection of oral history interviews with residents of the old neighborhoods of Shiraz and made them available on the YouTube channel *Bachoy Shahr* ‘Children of the city’. I use four of the interviews for this research.³⁰

To see how rare broken plurals are in Persian, we may take a quick look at their frequency in our four different sources, as shown in (3). The main numbers in the table correspond to token frequencies (counting every occurrence of a word separately) and the numbers in parentheses show type frequencies (i.e. number of distinct words). Quotations from Formal Persian (e.g., verses of poetry, names of institutions, and book titles) are not included in the counts.

(3)

Source	Nouns with plural morphology	Nouns with broken plural morphology	True broken plurals
Ja’fari	576 (232)	15 (13)	4 (2)
Mashq-e Shab	49 (27)	7 (4)	2 (1)
Barzegar	78 (51)	12 (6)	4 (2)
Bachoy Shahr	136 (97)	11 (9)	3 (1)

What we mean by ‘true broken plural’ in (3) is clarified in the next two subsections with examples and a thorough review of the data. For now, it suffices to say that we consider a word a true broken plural if it is clearly used as the plural form of another word in the lexicon. Thus, words with broken plural morphology that are used as singular or with meanings that do not directly

correspond to their morphologically expected singular forms do not count as true broken plurals.

The figures in (3) do show that true broken plurals are relatively uncommon in CP, but in the absence of additional evidence, they would still suggest that CP allows them. We see more evidence for our claim that they are not allowed in CP in the following subsections, but one piece of evidence that can be mentioned here comes from the optionality of broken plurals in CP. Any noun that is used in CP and can be pluralized has an acceptable regular plural form based on *-(h)ā*. This makes the status of broken plurals in CP different from irregular plurals in languages such as English, where the irregular plural is usually the only available plural form, e.g., ‘foot’ vs. ‘feet’, ‘stimulus’ vs. ‘stimuli’.

Crucially, unlike CP, broken plurals are not always optional in FP; there are words in FP whose only acceptable plural form is the broken plural. For instance, the FP word *fard* ‘person’ (non-existent in pure CP speech) is pluralized as *afrād*, almost never as *fard-hā* or *fard-ān*. Other words for which a plural other than the broken plural is unacceptable or highly marked in FP include *shakhs* ‘person’ (pl. *ashkhās*), *now* ‘type’ (pl. *anvā*), and *mowred* ‘item’ (pl. *mavāred*).

4.2. Broken plurals repurposed

A clear sign of resistance against broken plurals among Persian speakers is the extent to which these plurals are repurposed as singular nouns or even adjectives. When a speaker treats a broken plural as a singular noun, it suggests that their mental grammar has not incorporated words without overt affixal morphology as true plurals and thus has a hard time accepting these exceptional cases. Such individual ‘mistakes’ can collectively change the role of a lexical item in the language, turning it into a singular word in the host language. This has been an ongoing process for centuries, resulting in many words with broken plural morphology that are standardly used with a singular meaning not only in CP but even in FP. A few common examples of such cases in FP are given in (4). The meanings given in (4) are the Persian meanings, not the original Arabic ones.

(4)	Word	Meaning	Corresponding sg form	Meaning
	<i>advie</i>	‘spice’	<i>davā</i>	‘medicine’
	<i>aslahe</i>	‘weapon’	<i>selāh</i>	‘weapon’
	<i>asha’e</i>	‘ray’	<i>sho’ā</i>	‘ray’
	<i>hamāyel</i>	‘sash’	<i>hemāle</i> (not used)	—
	<i>shamāyel</i>	‘photo’	<i>shamile</i> (not used)	—
	<i>akhlāq</i>	‘behavioral trait’	<i>kholq</i>	‘behavioral trait’
	<i>talabe</i>	‘religious student’	<i>tāleb</i>	‘one who demands’
	<i>taba’e</i>	‘citizen’	<i>tābe’</i>	‘follower’
	<i>amale</i>	‘laborer’	<i>āmel</i>	‘agent’

<i>hoquq</i>	‘salary’	<i>haq</i>	‘right’
<i>arbāb</i>	‘master’	<i>rab</i>	‘Lord’

Some of these words are often treated as plural even in linguistically sophisticated FP texts and even pluralized again, e.g., *arbāb-ān* ‘masters’, *aslahe-hā* ‘weapons’. For others, knowledge of their plural origin exists either as an organic part of the lexical information associated with the word or through explicit education, resulting in a situation where the word is often used as a singular noun in FP texts but pluralizing through regular plural morphology is generally avoided, as is the case for *tabāʾe* ‘citizen’. There are also broken plurals that are used as singular in CP, even by educated speakers, but are not accepted as singular in FP, e.g., *aʿsāb* ‘temper’ (sg. *asab* ‘nerve’), *havās* ‘attention’ (sg. *hes* ‘feeling’), *ahvāl* ‘condition’ (sg. *hāl* ‘mood’), and two words meaning situation: *sharāyet* (sg. *sharite*; unused), and *owzā*’ (sg. *vaz*’).

The reuse of broken plurals as singular nouns in Persian is reminiscent of the Turkish case and suggests that the language resists this alien pluralization pattern. The driver for this change is presumably spoken language. After all, FP has historically been more closely associated with the educated elite, most of whom have had at least some knowledge of Arabic (even today, six years of morpho-syntax-heavy Arabic education is part of the official school curriculum in Iran).

What makes the situation even more similar to the Turkish case is the fact that Ottoman Turkish used to make extensive use of broken plurals. Note that in this context, by ‘Ottoman Turkish’ I refer to the written language of Turkish documents before the linguistic reforms of the early twentieth century. Even at the time, this written language was quite different from the spoken language of ordinary Turkish speakers, so much so that Kirchner describes the pre-reform linguistic situation in the Turkish-speaking world as diglossia.³¹ One of the goals of the linguistic reforms was in fact to close this gap between spoken and written language. It is therefore remarkable that the linguistic reforms resulted in a written language that is devoid of broken plurals. By analogy, it seems reasonable to argue that while FP makes extensive use of broken plurals (especially in certain genres of text), broken plurals occasionally appear in CP speech only as a result of bilingualism between CP and FP. Thus, we may speculate that in the hypothetical scenario that a development of the kind observed in the Turkish republic occurs in Iran and CP speakers lose contact with FP as we know it today, we may expect broken plurals to disappear from CP speech.

To give a sense of how pervasive the phenomenon of repurposing broken plurals is, we may look at our small corpus of CP speech. Of the 30 distinct words with broken plural morphology that appear in our corpus, 11 are used as singular at least once. Of these, four are cases that are not accepted as singular in standard educated CP speech (see the next subsection). The other seven are *aʿsāb* ‘temper’, *havās* ‘attention’, *owzā*’ ‘situation’, and *ahvāl* ‘mood’ (all four mentioned in the discussion above) and the last three rows in (4).

4.3. Persistent ‘errors’ in the use of broken plurals

The above discussion focused only on cases where words with broken plural morphology have gone through a shift in standard Persian, at least in CP. However, to argue that contemporary CP resists broken plurals, we need to be able to say that such repurposing does not merely reflect an earlier stage in the history of Persian, but is an active process. In other words, it must be the case that the type of ‘error’ that brought about these shifts is still regularly occurring in CP speech. Only in this case will be able to claim that today’s CP is still a language that has not incorporated broken plurals. This is indeed what we observe. Such ‘errors’ are quite frequent and commonly warned against in prescriptivist guides.³² There are many words that are occasionally used as singular and even pluralized with *-ā* by less educated speakers but not accepted as singular even in CP by educated speakers. A few notable examples are listed in (5).

(5)	Word	Meaning	Corresponding sg form
	<i>asbāb</i>	‘belonging(s)’	<i>sabab</i>
	<i>owlād</i>	‘child(ren)’	<i>valad</i>
	<i>vasāyēl</i>	‘tool(s)’	<i>vasile</i>
	<i>owbāsh</i>	‘thug(s)’	<i>vabsh</i> (not used)
	<i>arāzel</i>	‘thug(s)’	<i>razl</i>
	<i>sādāt</i>	‘descendant(s) of Mohammad’	<i>seyyed</i>
	<i>khoddām</i>	‘custodian(s) of a holy shrine’	<i>khādem</i>
	<i>atbā’</i>	‘citizen(s)’	<i>tābe’</i>

At the sociolinguistic level of analysis, we may think of these words as items that have the potential to enter standard CP in the future given the right social circumstances. However, at the individual level, treating broken plurals as singular is not limited to such words and can happen spontaneously for any word that has broken plural morphology on the surface. In our small corpus, non-standard singular use of broken plurals occurs four times. One case is the use of *sādāt* (listed above) by one of the speakers in the ‘Bachoy Shahr’ corpus. The other three words are not subject to any common ‘error’ pattern that the author is aware of. Ja’fari uses *avālem* ‘worlds’ (sg. *ālam*) as singular (p. 96), as shown in (6).

(6)	tu	in	avālem-ā	na-bud-am
	in	DEM.PROX	world.PL-PL	NEG-be.PST-1SG
	‘I was not in these worlds.’ (i.e., ‘I was not preoccupied with such matters.’)			

In another part of the interview, Ja’fari uses the high-register word *aqārīr* ‘confessions’ (sg. *eqrār*) twice. In one instance it is not clear if he uses the word as singular or plural, but in the other instance (p. 212), he uses it as a singular noun, saying *in aqārīr-e?* (this confession.PL-be.3SG) ‘Is this a confession?’.

Similarly, Barzegar uses *marāje* ‘high-ranking Shia clerics’ (sg. *marja*) as a singular word (00:28:14 in the YouTube file). While speaking about a particular Shia cleric, he refers to him as *in marāje* ‘this cleric (lit. ‘this clerics’). This is interesting because he uses this word three more times in his speech, and in one of them the structure of his sentence makes it clear that he is using it as plural. In particular, he uses the phrase *yek-i az marāje* (one-INDF of cleric.PL) ‘one of the clerics’. This is quite telling as it reveals that even when the speaker has at least some degree of knowledge of the plural nature of the word, he is still prone to pluralizing it again. This type of individual-level occasional error can feed sociolinguistic trends of the type described in the previous subsection.

4.4. Broken plurals without corresponding singular forms

We must make a distinction between different levels of incorporation of broken plurals. The most extreme case of incorporation would be to not only allow broken plurals into the language, but incorporate the morphological operation too for pluralizing new forms. FP does not seem to have done this with the exception of a few cases such as *farāmin* as the plural for *farmān* ‘command’. The second level of incorporation would be to accept broken plurals as true plurals, but not use the morphological patterns for any new forms. This seems to be what FP has done for the most part. The third degree of incorporation is to accept broken plurals as collective nouns, not as directly corresponding to any singular forms. This seems to be the case for many Arabic broken plurals in both CP and FP. I do not consider these forms true broken plurals (see below).

An example of this third case is the pair *adab* vs. *ādāb*. The singular form, *adab*, means ‘politeness’ (among other things). The plural form, *ādāb*, is indeed treated as a plural noun in the syntax, but it is not semantically the plural form of ‘politeness’. Instead, it means ‘social customs’. Crucially, it is not possible to refer to a single social custom as an *adab*. As this example demonstrates, this type of broken plural incorporation is quite subtle and often hard to detect because of the similarity between the meanings of the singular and plural forms and the complicated relationship between CP and FP.

The difference between this type of incorporation and the one in which the plural form has a corresponding singular form is theoretically significant. From a formal perspective, a true broken plural is composed of a main morpheme that may otherwise appear on its own to produce the singular version plus a pluralization morpheme whose phonological realization is irregular. For instance, FP *afrād* ‘persons’ is presumably composed of two morphemes, one of which is present in the singular form *fard* ‘person’ too and the other consists of the vowel and consonant pattern aCCāC and denotes pluralization. However, in a case such as *ādāb*, I argue that the word is composed of a single morpheme that is inherently plural, with no connection to *adab* in the mental CP lexicon and grammar of the speaker. Thus, in our synchronic analysis of CP, *ādāb* is no

different than the word *mardom* ‘people’ in its morpho-syntactic character; it is a mono-morphemic plural noun with no corresponding singular form.

Incorporating broken plurals at this minimal level occurs in Turkish too. For instance, the word *ahali* ‘residents’ is a plural noun in Turkish with no corresponding singular form in the language. However, the Turkish word that corresponds to the Arabic singular form of the word, *ehil*, is not used by itself as a noun to mean ‘resident’ in modern Turkish (although it is used as an adjective with other meanings). Thus, examples of this sort are in line with our broader claim that CP patterns with Turkish rather than FP when it comes to broken plurals. Incidentally, the particular example of *ahali* serves as an instance of this pattern in Persian too. In both CP and FP, *ahāli* is a plural noun meaning ‘residents’. However, the etymologically corresponding singular form *ahl* is used only as an adjective in Persian. Thus, even though it is possible to say, for instance, *ahl-e tehrān* to mean ‘Tehrani’, it is not possible to say *ye ahl* (one *ahl*) to mean ‘a resident’.

The words *ādāb* ‘social customs’ and *ahāli* ‘residents’ all appear in our small corpus of CP speech. Words that I argue belong to this category constitute 14 of the 30 distinct words with broken plural morphology that appear in our corpus. For some of them, listed in (7), the semantic disconnect between the singular and plural forms is relatively easier to see. Note that the meanings listed in (7) pertain to CP, and the contexts in which the words are used in our corpus. For instance, in FP, *akhbār* ‘pieces of news’ as a true plural for *khabar* ‘piece of news’ is completely acceptable. However, this sense of the word is normally not used in CP (except in frozen expressions) and where it does, I consider it a case of code-mixing (see the next subsection). In our corpus, one of the children in Mashq-e shab uses the word to refer to the news broadcast program on TV (rather than an actual collection of pieces of news), which is the most common way this word is used in CP.

(7)	Word	Meaning	sg form	Meaning
	<i>ādāb</i>	‘social customs’	<i>adab</i>	‘politeness’
	<i>ahāli</i>	‘residents’	<i>ahl</i>	‘hailing (from)’
	<i>rejāl</i>	‘political figures’	<i>rajol</i>	‘man’ (not in CP)
	<i>avārez</i>	‘tolls’	<i>āreze</i>	‘mishap’
	<i>ashāyer</i>	‘nomads’	<i>ashire</i>	‘clan’
	<i>akhbār</i>	‘news broadcast’	<i>khabar</i>	‘piece of news’

The other cases are less obvious. For instance, Ja’fari uses the word *qovā* ‘forces’ in the expression *qovā-ye komaki* ‘reinforcements’. This is one of the few frozen expressions in which *qovā* is used in pure CP. While the singular form *qovve* ‘force’ does exist in some frozen expressions and certain regional varieties of CP, it is not used by itself to mean ‘force’ in Tehrani Persian (which is Ja’fari’s variety). Moreover, the frozen expression *qovā-ye komaki* always appears in the plural form, i.e.,

qovve-ye komaki is not acceptable. This further strengthens the claim that *qovā* is part of a frozen expression here rather than a true plural for *qovve*. Finally, another piece of evidence showing that *qovā* is not the plural of *qovve* in the speakers' mental grammar is that *qovā-ye komaki* behaves like a singular noun phrase in most contexts. For instance, Ja'fari says *qovā-ye komaki ... umad* 'reinforcements arrived' with the 3sg verb *umad* rather than the 3pl form *umad-an*. To put this in perspective, we may compare it with another expression that also means 'reinforcements' and has a similar structure but uses the native Persian word *niru* 'force' with regular pluralization. For this, one has to say *niru-ā-ye komaki umad-an* 'reinforcements arrived' (with plural agreement) rather than **niru-ā-ye komaki umad* (with singular agreement). This shows that while *niru-ā* is a true plural in CP, *qovā* is not.

The rather detailed discussion above about *qovā*, tedious as it may be, demonstrates how subtle the situation may be with these words. An even more subtle (and in fact less certain) case showing up in our corpus is *omur* 'affairs, matters'. The common meaning for the singular form of the word, *amr*, is 'command'. In highly refined FP texts, it is possible to use the singular *amr* to mean 'affair' and 'matter' too, but in regular CP conversation, this almost never happens. For instance, it is totally acceptable in CP to say *in omur* 'such matters', but it is not quite normal (to say the least) to say *ye amr az in omur* 'one matter among these matters'. The use of *amr* with the meaning 'affair, matter' in CP is for the most part limited to a few frozen expressions such as *amr-e kheyr* 'auspicious affair' (referring to marriage). Thus, it seems fair to argue that *omur* is unrelated to *amr* in the mental CP lexicon of speakers.

A third interesting case in our corpus is *a'māq* 'depths' (sg. *omq*), which is used in CP in fixed expressions involving nouns such as the 'heart', e.g., *tu a'māq-e qalb-eshun* 'in the depths of their heart', as seen in the Barzegar interview in our corpus. However, in a regular context where the plural form for *omq* 'depth' is needed, *a'māq* is not used. For instance, to say that the different areas of a swimming pool have 'different depths', the accepted form is *omq-ā-ye mokhtalef*, not *a'māq-e mokhtalef*. Moreover, the regular plural cannot appear in the fixed expression either. Thus, saying *tu omq-ā-ye qalb-eshun* for 'in the depths of their heart' is not acceptable at all. These facts suggest that as far as a formal description of the mental CP grammar and lexicon of the speaker is concerned, *a'māq* is an independent lexical item rather than the plural form of *omq*.

The other words used in our corpus that may not be morphologically related in the synchronic grammar of CP to their etymological singular forms are *a'emme* 'imams' (sg. *emām*), *majāles* 'formal gatherings' (sg. *majles*), *madārek* 'documents' (sg. *madrak*), *rofaqā* 'friends' (sg. *rafiq*), *osarā* 'captives' (sg. *asir*) and *olamā* 'religious scholars' (sg. *ālem*). In all of these six cases, the singular form exists with a meaning similar to the one used for the plural, at least in FP, but the plural form is used naturally only in frozen expressions or special contexts. It is natural to talk about the twelve Shia Imams collectively as *a'emme*, but using this word as the plural form of *emām* is unnatural in most other contexts. For instance, when talking about two particular Shia Imams

(say the second and third Imams, who were in fact brothers), it is natural to say things like *harkodum az in emam-ā* (any from this Imam-PL) ‘any of these Imams’, but not *harkodum az in a’emme*.

One effect of acknowledging the fact that words of the type discussed above are not true broken plurals is that it shows that the number of broken plurals appearing in CP speech is less than one would otherwise assume, and in this sense strengthens the idea that broken plurals are not integral to CP. However, more importantly, this observation serves as another piece of evidence showing CP grammar’s active resistance against broken plurals. Even when these words are not treated as singular, they are treated as something other than plural nouns systematically corresponding to certain singular forms. This diachronic process is indirect evidence for the claim that synchronically, the CP speaker’s mental grammar does not accept broken plurals and their singular forms as systematically connected forms.

4.5. Broken plurals and code-mixing

We saw how many of the instances of broken plurals in CP may be misleading at first as they are not instances of true pluralization. However, even after we set aside all of the suspicious cases, there are still cases that seem to be true broken plurals in CP speech. In our corpus, there are four instances, listed in (8).

(8)	Plural	Singular	Meaning	Source
	<i>vokalā</i>	<i>vakil</i>	‘parliament member’	Ja’fari
	<i>takālif</i>	<i>taklif</i>	‘homework assignment’	Mashq-e Shab
	<i>marāje’</i>	<i>marja’</i>	‘high-ranking Shia cleric’	Barzegar
	<i>qavānin</i>	<i>qānun</i>	‘regulation’	Barzegar

Words of this category are probably true broken plurals, but it is not immediately clear whether they should be considered as true CP words or the result of code-mixing (with the definition established in the previous subsections). We already discussed several reasons pointing towards the idea that CP does not accept true broken plurals, including the facts that they are rare, they are optional, and they persistently get repurposed as singular or take on a different meaning. This leads us to suggest that cases like those in (8) must be viewed as cases of code-mixing. For some of these particular instances, we have additional evidence too. The word *takālif* ‘homework assignments’, for instance, appears in the speech of the children in Mashq-e Shab in response to prompts by the interviewer containing this word. Thus, the word is not a spontaneous choice by the children. Recall that the children were chosen in the first place because their familiarity with FP is limited. In the case of *marāje’* ‘high-ranking clerics’, even though Barzegar uses the word as a plural on a couple of occasions, he also uses it as a singular noun on one occasion, as mentioned earlier. This suggests that he is not fully comfortable with the morpho-syntactic behavior of the

word. This is understandable if he is accessing the word through his mental FP lexicon and grammar, given that he is probably not very comfortable with FP.

A different type of potential evidence for the FP origin of these words comes from phonology. As discussed earlier, the vowel length distinction observed in FP phonology is either non-existent or optional in pure CP. Thus, it is significant that the traditionally 'long' vowels in these words tend to be pronounced long in CP speech.³³ MacSwan and Colina see phonology at the core of language switching during speech.³⁴ This would suggest that the vowel facts by themselves serve as a powerful piece of evidence in favor of these words not being part of CP lexicon. However, admittedly, there are two factors that reduce the force of the argument. First, as noted earlier, the status of vowel length in CP is not straightforward and seems to be subject to a rather complicated pattern of distribution. Second, as others have noted,³⁵ the relationship between phonological incorporation and lexico-morphological incorporation is not always straightforward, and there seem to be important cases of mismatch. Thus, while the phonological facts do strengthen the position offered in this paper, they are not decisive by themselves.

5. Conclusion

This paper argued for the position that broken plurals have not been truly incorporated into Colloquial Persian. To show this, first, the distinction between the lexicons and grammars of Colloquial Persian and Formal Persian was highlighted and it was shown that code-mixing and borrowing between the two language varieties is quite common, even though it usually goes unnoticed because of the high degree of similarity between the two varieties and speakers' meta-linguistic assumptions about the relationship between CP and FP.

After the distinction between CP and FP was established, several arguments were offered in favor of broken plurals not having been fully incorporated into CP. These arguments involved their rarity and optionality in CP, the common process of repurposing words with broken plural morphology as singular nouns, and the frequent semantic disconnect between the singular and broken plural forms. Based on these observations as well as phonological factors, it was argued that the cases that do feature true broken plurals in CP (which are common in the speech of highly educated speakers but rare otherwise) must be viewed as cases of code-mixing.

In parallel with the language-internal arguments mentioned above, cross-linguistic comparisons are helpful in both clarifying and strengthening the main argument of the paper. There are other languages that have borrowed large amounts of Arabic words but have resisted broken plural morphology, including both Iranian languages (e.g., Kurdish varieties) and Turkish. The Turkish case is particularly eye-opening when its history is considered. The gap between standard written Ottoman Turkish (in which broken plurals were prevalent) and spoken Turkish before the linguistic reforms of the early 20th century may serve as a convenient example

for the kind of analysis offered here regarding the gap between FP and CP. The fact that broken plurals disappeared from standard Turkish once the spoken language of the uneducated masses was given priority during the linguistic reforms makes it easier to imagine how the continuation of the use of broken plurals in a language may be the result of continued prevalence of diglossia in a community rather than true incorporation of the broken plurals.

Bibliography

- Alexiadou, Artemis, and Terje Lohndal. "Units of Language Mixing: A Cross-Linguistic Perspective." *Frontiers in Psychology* 9 (September 27, 2018). doi:10.3389/fpsyg.2018.01719.
- Bokamba, Eyamba G. "Code-Mixing, Language Variation, and Linguistic Theory." *Lingua* 76, no. 1 (September 1988): 21–62. doi:10.1016/0024-3841(88)90017-4.
- Carnie, Andrew. *Syntax: A Generative Introduction*. Wiley-Blackwell, 2013.
- Ferguson, Charles A. "Diglossia." *Word* 15, no. 2 (1959): 325–40. doi:10.1080/00437956.1959.11659702.
- Gaetano, Berruto. "Code-Switching and Code-Mixing." In *The Dialects of Italy*, edited by Dr Martin Maiden. Routledge, 1997. doi:10.4324/9780203993880.
- Gharavian, D., H. Sheikhzadeh, and S. M. Ahadi. "An Experimental Multi-Speaker Study on Farsi Phoneme Duration Rules Using Automatic Alignment." In *8th Australian International Conference on Speech Science and Technology*, 186–91, 2000.
- Gumperz, John J. *Discourse Strategies*. Cambridge University Press, 1982. doi:10.1017/CBO9780511611834.
- Henderson, Michael M. T. "Diglossia in Kabul Persian Phonology." *Journal of the American Oriental Society* 95, no. 4 (1975): 651–54. doi:10.2307/601021.
- Jeremiás, Éva. "Diglossia in Persian." *Acta Linguistica Academiae Scientiarum Hungaricae* 34, no. 3/4 (1984): 271–87.
- Jones, Taylor. "A Corpus Phonetic Study of Contemporary Persian Vowels in Casual Speech" 25, no. 1 (2019): 129–37.
- Jasbi, Masoud. "The suffix that makes Persian nouns unique." In *Advances in Iranian Linguistics*, edited by Richard Larson, Sedigheh Moradi, and Vida Samiiian. John Benjamins Publishing Company, 2020. doi:https://doi.org/10.1075/cilt.351.06jas.
- Kiarostami, Abbās. *Masgh-e Shab*, 1989.

- Kirchner, Mark. "Turkish." In *Encyclopedia of Arabic Language and Linguistics*, edited by Kees Versteegh, 4:583–89. Leiden: Brill, 2009.
- MacSwan, Jeff, and Sonia Colina. "Some Consequences of Language Design: Codeswitching and the PF Interface." In *Grammatical Theory and Bilingual Codeswitching*, 185–210. The MIT Press, 2014. doi:10.7551/mitpress/8338.003.0011.
- Mahdavi Mazdeh, Mohsen. "Linguistic Change and the Future of Metrical Persian Poetry." *Iranian Studies* 54, nos. 5–6 (2021). doi:10.1080/00210862.2020.1777391.
- . *Nezām-e Vazni-e She'r-e Fārsi [The Metrical Structure of Persian Poetry]*. Tehran: Nashr-e Now, 2024.
- . "Quantitative Meter in Persian Folk Songs and Pop Lyrics." In *Advances in Iranian Linguistics*, edited by Richard Larson, Sedigheh Moradi, and Vida Samiian. John Benjamins Publishing Company, 2020. doi:https://doi.org/10.1075/cilt.351.11mah.
- Mahmoodi-Bakhtiari, Behrooz. "Spoken vs. Written Persian: Is Persian Diglossic?" In *Trends in Iranian and Persian Linguistics*, 183–212. De Gruyter, 2018. doi:10.1515/9783110455793-011.
- Managan, Kathe. "Diglossia Reconsidered: Language Choice and Code-Switching in Guadeloupean Voluntary Organizations." In *Proceedings of the Eleventh Annual Symposium about Language and Society*, 251–61. Austin, 2003.
- Mejdell, Gunvor. "Diglossia." In *The Routledge Handbook of Arabic Linguistics*, 332–44. Milton Park, Abingdon, Oxon ; New York, NY : Routledge, 2017.: Routledge, 2017. doi:10.4324/9781315147062-18.
- Muysken, Pieter. *Bilingual Speech : A Typology of Code-Mixing*. Cambridge University Press, 2005.
- Najafi, Abolhasan. *Ghalat Nanevisim [Let Us Not Write Incorrectly]*. Tehran: Markaz-e Nashr-e Dāneshgāhi, 2012.
- Perry, John. "Persian as a Homoglossic Language." In *Iran: Questions et Connaissances. Actes Du IVe Congrès Européen Des Études Iraniennes, Vol. III: Cultures et Sociétés Contemporaines*, edited by Bernard Hourcade, 11–28. Peeters, 2003.
- Poplack, Shana, and Marjory Meechan. "Introduction: How Languages Fit Together in Codemixing." *International Journal of Bilingualism* 2, no. 2 (1998): 127–38. doi:10.1177/136700699800200201.
- Rossi, Adriano V, and others. "Diglossia in Persian." *From Aşl to Zā'id: Essays in Honour of Éva M. Jeremiás. Píliscsaba: The Avicenna Institute of Middle Eastern Studies*, 2015, 211–19.

Sang-Tajan, Shahin, and Mahmood Bijankhan. “Barresi-e Mâhiat-e Sowtshenâkhti-e Vâkehâ-Ye Fârsi Dar Hejâhâ-Ye CV-e Goftâr-e Ezhâri [A Study of the Phonetic Nature of Persian Vowels in CV Syllables in Citation Form]” 4, no. 1 (2013): 97–116.

Sarshâr, Homâ. *Sha’bân-e Ja’fari*. Los Angeles: Naab, 2002.

Sobrero, Alberto. “Code Switching in Dialectal Communities in Italy.” *Rivista Di Linguistica* 6, no. 1 (1994): 39–55.

Thackston, W. M. “Sorani Kurdish: A Reference Grammar with Selected Readings,” 2006.

Vahidiân Kâmyâr, Taqi. *Barresi-e Vazn-e She’r-e Âmiâne [A Review of Meter in Folk Poetry]*. Tehran: Āgah, 1978.

World Bank. “Literacy Rate, Adult Total - Iran, Islamic Rep.,” March 2026.
<https://data.worldbank.org/indicator/SE.ADT.LITR.ZS?locations=IR>.

¹ Behrooz Mahmoodi-Bakhtiari, “Spoken vs. Written Persian: Is Persian Diglossic?,” in *Trends in Iranian and Persian Linguistics* (De Gruyter, 2018), 183–212, doi:10.1515/9783110455793-011; Éva Jeremiás, “Diglossia in Persian,” *Acta Linguistica Academiae Scientiarum Hungaricae* 34, no. 3/4 (1984): 271–87; Michael M. T. Henderson, “Diglossia in Kabul Persian Phonology,” *Journal of the American Oriental Society* 95, no. 4 (1975): 651–54, doi:10.2307/601021.

² Charles A. Ferguson, “Diglossia,” *Word* 15, no. 2 (1959): 325–40, doi:10.1080/00437956.1959.11659702.

³ Mahmoodi-Bakhtiari, “Spoken vs. Written Persian: Is Persian Diglossic?”; Jeremiás, “Diglossia in Persian”; Henderson, “Diglossia in Kabul Persian Phonology.”

⁴ John Perry, “Persian as a Homoglossic Language,” in *Iran: Questions et Connaissances. Actes Du IVe Congrès Européen Des Études Iraniennes, Vol. III: Cultures et Sociétés Contemporaines*, ed. Bernard Hourcade (Peeters, 2003), 11–28.

⁵ Adriano V Rossi and others, “Diglossia in Persian,” *From Aşl to Zā’id: Essays in Honour of Éva M. Jeremiás. Piliscsaba: The Avicenna Institute of Middle Eastern Studies*, 2015, 211–19.

⁶ Mohsen Mahdavi Mazdeh, “Linguistic Change and the Future of Metrical Persian Poetry,” *Iranian Studies* 54, nos. 5–6 (2021), doi:10.1080/00210862.2020.1777391, 29–30.

⁷ In this and the rest of the examples, the CP version precedes the FP version.

⁸ D. Gharavian, H. Sheikhzadeh, and S. M. Ahadi, “An Experimental Multi-Speaker Study on Farsi Phoneme Duration Rules Using Automatic Alignment,” in *8th Australian International Conference on Speech Science and Technology*, 2000, 186–91.

⁹ Shahin Sang-Tajan and Mahmood Bijankhan, “Barresi-e Mâhiat-e Sowtshenâkhti-e Vâkehâ-Ye Fârsi Dar Hejâhâ-Ye CV-e Goftâr-e Ezhâri [A Study of the Phonetic Nature of Persian Vowels in CV Syllables in Citation Form]” 4, no. 1 (2013): 97–116; Taylor Jones, “A Corpus Phonetic Study of Contemporary Persian Vowels in Casual Speech” 25, no. 1 (2019): 129–37.

¹⁰ Taqi Vahidiân Kâmyâr, *Barresi-e Vazn-e She’r-e Âmiâne [A Review of Meter in Folk Poetry]* (Tehran: Āgah, 1978); Mohsen Mahdavi Mazdeh, “Quantitative Meter in Persian Folk Songs and Pop Lyrics,” in *Advances in Iranian Linguistics*, ed. Richard Larson, Sedigheh Moradi, and Vida Samiian (John Benjamins Publishing Company, 2020), doi:<https://doi.org/10.1075/cilt.351.11mah>.

¹¹ Mohsen Mahdavi Mazdeh, *Nezām-e Vazni-e She’r-e Fârsi [The Metrical Structure of Persian Poetry]* (Tehran: Nashr-e Now, 2024), 252–260.

¹² Mahmoodi-Bakhtiari, “Spoken vs. Written Persian: Is Persian Diglossic?”

¹³ Ibid.

¹⁴ Masoud Jasbi, “The Suffix That Makes Persian Nouns Unique,” *Advances in Iranian Linguistics* 351 (2020): 107.

-
- ¹⁵ Andrew Carnie, *Syntax : A Generative Introduction* (Wiley-Blackwell, 2013), 159–160.
- ¹⁶ Shana Poplack and Marjory Meechan, “Introduction: How Languages Fit Together in Codemixing,” *International Journal of Bilingualism* 2, no. 2 (1998): 127–38, doi:10.1177/136700699800200201.
- ¹⁷ Eyamba G. Bokamba, “Code-Mixing, Language Variation, and Linguistic Theory:,” *Lingua* 76, no. 1 (September 1988): 21–62, doi:10.1016/0024-3841(88)90017-4; John J. Gumperz, *Discourse Strategies* (Cambridge University Press, 1982), doi:10.1017/CBO9780511611834.
- ¹⁸ Artemis Alexiadou and Terje Lohndal, “Units of Language Mixing: A Cross-Linguistic Perspective,” *Frontiers in Psychology* 9 (September 27, 2018), doi:10.3389/fpsyg.2018.01719.
- ¹⁹ Pieter Muysken, *Bilingual Speech : A Typology of Code-Mixing* (Cambridge University Press, 2005).
- ²⁰ Gunvor Mejdell, “Diglossia,” in *The Routledge Handbook of Arabic Linguistics* (Milton Park, Abingdon, Oxon ; New York, NY : Routledge, 2017.: Routledge, 2017), 332–44, doi:10.4324/9781315147062-18.
- ²¹ Kathe Managan, “Diglossia Reconsidered: Language Choice and Code-Switching in Guadeloupean Voluntary Organizations,” in *Proceedings of the Eleventh Annual Symposium about Language and Society* (Austin, 2003), 251–61.
- ²² Berruto Gaetano, “Code-Switching and Code-Mixing,” in *The Dialects of Italy*, ed. Dr Martin Maiden (Routledge, 1997), doi:10.4324/9780203993880; Alberto Sobrero, “Code Switching in Dialectal Communities in Italy,” *Rivista Di Linguistica* 6, no. 1 (1994): 39–55.
- ²³ Mahdavi Mazdeh, *Nezām-e Vazni-e She’r-e Fārsi* [The Metrical Structure of Persian Poetry], 255–256.
- ²⁴ For a brief account of the fate of broken plurals in Turkish, see Mark Kirchner, “Turkish,” in *Encyclopedia of Arabic Language and Linguistics*, ed. Kees Versteegh (Leiden: Brill, 2009).
- ²⁵ W. M. Thackston, “Sorani Kurdish: A Reference Grammar with Selected Readings,” 2006.
- ²⁶ World Bank, “Literacy Rate, Adult Total - Iran, Islamic Rep.,” March 2026, <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS?locations=IR>.
- ²⁷ Homā Sarshār, *Sha’bān-e Ja’fari* (Los Angeles: Naab, 2002).
- ²⁸ Abbās Kiarostami, *Masghq-e Shab*, 1989.
- ²⁹ <https://youtu.be/g9g5ixwWkHA>
- ³⁰ Interviews with Ahmad Mowlā, Ebrāhim biābāni, Shirzād Chahābdār, and Rasul Moqaddam.
- ³¹ Kirchner, “Turkish.”
- ³² e.g., Abolhasan Najafi, *Ghalat Nanevisim* (Tehran: Markaz-e Nashr-e Dāneshgāhi, 2012), 152.
- ³³ Confirming this last claim decisively requires phonetic experiments, but the author’s intuitions strongly point in this direction.
- ³⁴ Jeff MacSwan and Sonia Colina, “Some Consequences of Language Design: Codeswitching and the PF Interface,” in *Grammatical Theory and Bilingual Codeswitching* (The MIT Press, 2014), 185–210, doi:10.7551/mitpress/8338.003.0011.
- ³⁵ Alexiadou and Lohndal, “Units of Language Mixing: A Cross-Linguistic Perspective.”