Persian Influence on Classical Sarikoli in a Diglossic Context

Junmin Zeng

School of Foreign Studies, Guangzhou University of Chinese Medicine

510006 Guangzhou, China

email: tsang1983@gzucm.edu.cn

Keywords: Classical Sarikoli, Persian, Language contact, Diglossia, Chinese Tajiks.

Abstract. Classical Sarikoli, the literary language used by the Chinese Tajiks in Xinjiang, remains an academic lacuna to be bridged. The present study demonstrates obvious evidence of Persian influence on Classical Sarikoli in both lexical and syntactic levels, and proposes that this situation probably results from the pre-1930 Persian-Sarikoli diglossia practiced by the then Tajik community in westernmost China. Data reveal that Classical Sarikoli has far more Persian loanwords and Persianized lexical forms than contemporary Colloquial Sarikoli does, and the *izofat* construction and *wa*-coordination typical in Persian are ubiquitous in Classical Sarikoli, despite their absence from contemporary Sarikoli vernacular. The abundant Persian loan words and expressions in Classical Sarikoli have granted it a higher status, turning it into a 'high' variety in the 'Classical–Colloquial' sub-diglossia. As evidenced by that, the pre-1930 situation in the Tajik-populated Tashkorgan is better described as 'embedded diglossia'.

1. Classical Sarikoli and the Pre-1930 Persian-Sarikoli diglossia in Tashkorgan

Sarikoli is an Eastern Iranian language exclusively spoken by the Tajik ethnic group in China [1]. Previous studies of this language are very scanty [1, 2, 3, 4], and hardly any existing research has paid attention to its historical varieties for lack of data [5]. In 2016, when probing into the *General Catalogue and Abstracts of China's Ancient Ethnic Tajik Literature* (GCAT henceforth) [6], the present author identified a literary variety of Sarikoli, and termed this 'fused lect' as 'classical Sarikoli literary language' [5], which will be referred to as Classical Sarikoli hereafter. Up till now, Classical Sarikoli has escaped the research lens of the linguists interested in Eastern Iranian languages, and remains a gap to be bridged in this field.

Classical Sarikoli, or Literary Sarikoli, as the author has presented in the 2016 article, differs distinctly from Colloquial Sarikoli. In contemporary Tashkorgan (the county with the largest population of Chinese Tajiks), people still use Colloquial Sarikoli for daily communication, but Classical Sarikoli is only known to and mastered by a very rare group of Chinese Tajik poets and writers. Moreover, the lack of an official writing system adds to the difficulty of collecting historical documents in Classical Sarikoli. Fortunately, in the investigation of Chinese Tajik traditional medicine, an obscure subbranch of the Perso-Arabic traditional medicine, the present author has accumulated sufficient data for further analysis, and this is part of the outcome.

Classical Sarikoli as a literary language developed from the close contact of Persian and Colloquial Sarikoli in a typical context of interlingual diglossia before the 1930s [5]. In a diglossic situation, two or more languages (or varieties of the same language) serve differentiated functions, conventionally labeled as 'high' and 'low' [7, 8]. The high language or variety has more prestige and is used in formal contexts such as literature, and the non-prestigious low variety can only be applied to informal situations like everyday conversation.

However, in a more complicated multilingual community, there often exists the so-called 'embedded diglossia' (calque of the French term '*diglossie enchâssée*') as introduced by Calvet [9]. In such a situation, one diglossia is embedded in another. For instance, in Singapore, Singapore English and the local languages (Mandarin, Hokkien, Cantonese, Malay, Tamil etc.) are in a diglossic relationship at the primary level, with Singapore English as the high language and other languages as low languages, in which embedded are the diglossia between Singapore Standard English (as the high

variety) and the Singapore Colloquial English (as the low variety), and the diglossia between Mandarin (as the high variety) and other non-Mandarin Chinese dialects (as low varieties). Analogously, in Imperial China, Classical Chinese and other Chinese dialects in China have a first-level diglossia, while within most of the 'low' Sinitic varieties there are, more often than not, embedded diglossic situations, e.g. between Literary Teochew and Colloquial Teochew. The literary variety of a Chinese dialect, such as Literary Suzhounese or Literary Taiwanese, is the high variety, and the colloquial dialect, the low variety. Even the same Chinese character often has different pronunciations in these two varieties, with the high variety using the 'literary reading' or *wendu*, and the low variety the 'colloquial reading' or *baidu*.

Such 'embedded diglossia' is also found in the pre-1930 Chinese Tajik community. At the first level, Persian, as the high language, served most formal written functions while the local Eastern Iranian language, Sarikoli, was often used in informal oral contexts. However, embedded in the Persian-Sarikoli diglossia was the secondary relationship between Classical Sarikoli as the high variety and Colloquial Sarikoli as the low variety. Classical Sarikoli was used in poetry (e.g. *Rubāʿī*) and other elegant, prestigious literary forms, but Colloquial Sarikoli served merely daily purposes.

Classical Sarikoli can in general be regarded as a literary Sarikoli variety heavily influenced by the then high language Persian [5], somewhat comparable to Literary Chagatai. Zeng's study defines Classical Sarikoli as a 'fused lect' resulting from the contact between Persian and Sarikoli [5], but no specific investigation has been conducted to find out the detailed categorical distinction between Classical and Colloquial Sarikoli in terms of Persian influence.

The present study is an attempt to address this gap via examination of the Classical Sarikoli data sampled from GCAT in contrast to the Colloquial Sarikoli. The 1,625 entries of GCAT consist of 52 books and 1,573 texts of oral literature in various genres. The books are all in Persian, the high language then, while the oral literature texts largely represent Classical Sarikoli. The sample includes the entry titles from the first six volumes of the 10-volume oral literature section, with outliers such as proper names of gods and historical figures removed. Note that the all the data, including the two varieties of Sarikoli, are re-transliterated in line with Neikramon Ibrukhim's Chinese Tajik Romanization system [10], with some modifications based on the transcription convention of Tajik Persian for the loan words and expressions [11].

2.   The Persian influence on Classical Sarikoli: Lexical level

In this section we examine the Perso-Arabic loanwords in Classical Sarikoli. Sarikoli people did not have direct contact with Arabic community, and their Arabic loanwords were all imported via Persian. Thus, Perso-Arabic words can be jointly seen as evidence of Persian influence on Classical Sarikoli. In this paper, Perso-Arabic loanwords are further divided into 'naturalized' and 'exotic' ones, with the first category having adapted to the phonological rules of Sarikoli (i.e. with 'colloquial reading'), and the second one still preserving the foreign flavor in pronunciation (i.e. with 'literary reading'). For instance, the lexical duo for the official ethnonym includes two loanwords, *Tojik* (exotic/literary reading, L hereafter) and *Tujik* (naturalized/colloquial reading, C hereafter), with former showing more Persian influence. The naturalized loanwords, additionally, consist of active loanwords and obsolete loanwords, and the diminishing of the latter is a sign of the fading Persian influence. In our survey, we pay special attention to three parameters at the lexical level: A) exotic loanwords; B) obsolete loanwords; and C) native words and active naturalized loanwords that are readily used in contemporary Colloquial Sarikoli. The percentage of A and B is supposed to reflect the Persian influence in the pre-1930 diglossic literary Sarikoli.

In total we collected 420 entry titles from GCAT, and, as the data revealed, 170 entries contain exotic Perso-Arabic loanwords (accounting for 40.5%), 75 entries have obsolete Perso-Arabic loanwords (17.9%), and 56 entries are found to be with both (13.3%), making a joint percentage of 71.7%. In other words, only 28.3% of the total data are lexically in line with contemporary Colloquial Sarikoli. This drastic contrast demonstrates the once-tremendous Persian influence on the Classical Sarikoli lexicon.

Exotic Perso-Arabic loanwords are found in altogether 226 entries, making the largest category. Noticeably, each exotic loanword in Classical Sarikoli usually has a naturalized counterpart, and the two were used in different contexts for different functions. For example, 'myth' is *afsona* in GCAT, in the exotic/literary form, while it is *afsuno* in the colloquial variety. Such L/C lexical duos include *rivoyat/reewoyat* ('legend'), *vafodor/wafudur* ('loyal, faithful'), *pakhlavon/palwun* ('warrior'), *ishq/usheq* ('love'), *umed/eemeith* ('hope'), *yor/yur* ('lover'), etc. In contemporary Tajik community, the exotic/literary form is largely replaced by the naturalized form, even in literary creations.

Obsolete Perso-Arabic loanwords account for the second largest category, found in 131 entries (with 56 also containing exotic loanwords). For instance, 'water' is often lexicalized as *ob* in GCAT, but in contemporary Sarikoli, *k'ac* is the prevailing word. In this duo, *ob* is a Persian loanword now fading out of daily use, whereas *k'ac* is the native Eastern Iranian word. Such cases of Persian loanwords replaced by native Sarikoli words or loanwords from Uyghur and Chinese mirror the diglossic shift taking place in the Tajik community in 20th century. To exemplify that, *otash* 'fire' is replaced by *yuc*, *boron* 'rain' by *wareij*, *sang* 'stone' by *zher*, *kuh* 'mountain' by *qir* (Turkic loanwords), *gurg* 'wolf' by *k'ith'p*, etc.

3. The Persian influence on Classical Sarikoli: Syntactic level

At the phrasal level, focuses are placed on coordination and *izofat* constructions.

Typical coordination in contemporary Sarikoli is often formed by applying the Eastern Iranian conjunction *at*, such as *waz at teau* 'I and you'; in Persian, *wa* is used in the place of *at*. In our GCAT sample, Classical Sarikoli used three conjunctions: *wa*, *at* and *kham*. *Kham* is only found in one verse and thus in our survey we only contrast the *wa-* and *at-* constructions. In all 45 coordination phrases sampled, *wa* is found in 33 (73.3%) and *at* in 12 (26.7%). The nouns linked by *wa* can be exotic loanwords as in *odam wa gurg* ('the man and the wolf'), but they can also be pure native Sarikoli vocabulary like *rapc wa yeerk'* ('the fox and the bear'). However, in contemporary Sarikoli, *wa* has totally died out.

The other peculiarity we examined is the *izofat* or *ezāfe* construction, a 'noun-modifier' construction with the *izofat* marker *-i* added to the noun modified [12]. *Izofat* constructions are ubiquitous in modern Persian and many Iranian languages, but in contemporary Sarikoli such constructions are no longer productive and can only be found in fossilized expressions like *Sadichin* (lit. 'Wall of China', i.e. 'the Great Wall'). Instead, Chinese Tajiks today use only the native 'modifier-noun' construction devoid of the linking *-i*. To our surprise, *izofat* constructions were never a rarity in GCAT. In all 281 entries that contain certain forms of NPs with a modifier, *izofat* constructions account for 72.6% (204 entries in total), while the native non-*izofat* modifier-noun constructions amount to only 77 entries (27.4%). A recurrent example of the *izofat* construction is *otash-i ishq* (lit. 'fire-of love'), in contrast to pre-1930 Colloquial Sarikoli *ishq yuc* (lit. 'love fire'). Apparently, *izofat* constructions were of paramount dominance in Classical Sarikoli.

When it comes to the clausal level, we have sorted out 25 sentences from the 420-entry sample, which consist of 12 sentences composed of native lexicon and naturalized loanwords, 8 sentences with purely Persian vocabulary, and 5 sentences that are of mixed elements. 'Loan sentences' entirely borrowed from Persian, such as *Yor biyo!* ('My love, come!') and *Khol-i shumo chitavr ast?* (lit. 'condition-of you how is', i.e. 'How are you?'), account for 32% of all sentences sampled, which is an interesting phenomenon. Such cases are often regarded as code-switching rather than borrowing, but it also suggests a strong influence of Persian. In contemporary Sarikoli, such 'loan sentences' are confined to only a very few formulaic expressions, like *Tashakeer* ('Thank you!'), *Hudo khofiz* ('Goodbye!') and *Heesh omadi* ('Welcome!').

4. Conclusion

Classical Sarikoli, the high variety once forming an embedded diglossic situation with Colloquial Sarikoli within the macro-diglossia between Persian and Sarikoli before the 1930s, has dwindled away on the verge of extinction since Uyghur and Chinese replaced Persian as the high language after

1934 [5]. In present-day Tashkorgan, few Tajiks can speak or understand Persian as their grandparents did, and most of them even do not know that there has existed such a complex diglossic period in history. Such embedded diglossia can also find analogy in ancient Vietnam, Korea, Japan and Ryukyu, where the once East Asian lingua franca, Classical Chinese, used to play the role of high language, whilst the national languages of these countries became the low languages. From these 'low' languages in turn developed 'embedded diglossia', with a literary high variety heavily influenced by Classical Chinese in stark contrast to the colloquial variety of ordinary people. For example, in Meiji-era Japan, *kanbun* (Classical Chinese) and Japanese formed the primary diglossia, while *bungo* (Classical Japanese) and colloquial Modern Japanese constituted a secondary diglossia. Similar to Classical Sarikoli, Classical Japanese fell out of use in the early Shōwa period, partly due to the drastic decrease in the influence of Classical Chinese in modern East Asia.

Since China proposed the 'Belt and Road' Initiative, Tashkorgan, home to most Chinese Tajiks, has been redefined as an essential hub on the Silk Road, and Persian, official language of Iran, Tajikistan and Afghanistan, once a lingua franca across Inner Asia and a high language in Tashkorgan, has redrawn increasing attention of Chinese Tajiks. There is a chance that contemporary Sarikoli will re-absorb more Persian influence, and Classical Sarikoli be reincarnated in some brand-new form in future.

Acknowledgements

References

[1] J. Zeng, Determination of genetic distance between Khotanese and two Chinese Pamir languages: A lexicostatistic approach, *Acta Linguistica et Litteraturaria Sinica Occidentalia*, vol. 13, pp. 350-373, 2016.

[2] T. N. Pakhalina, *Sarykol'skij Jazyk*. Moskva: Izdatel'stvo Nauka, 1966.

[3] E. Gao, *A Grammatical Sketch of the Chinese Tajik Language*. Beijing: The Ethnic Publishing House, 1985.

[4] D. Kim, *Topics in the Syntax of Sarikoli*, Ph.D. thesis. Leiden University, 2017.

[5] J. Zeng, The 'Chinese Tajik' script and the pre-1930 Persian-Sarikoli diglossia in Tashkorgan, *The Journal of Northern Cultures Studies*, vol. 7, pp. 69-78, 2016.

[6] S. Zamir *et al*. *General Catalogue and Abstracts of China's Ancient Ethnic Tajik Literature*. Beijing: Encyclopedia of China Publishing House, 2011.

[7] C. Ferguson, Diglossia, *Word*, vol. 15, pp. 325–340, 1959.

[8] J. A. Fishman, *The Sociology of Language*. Rowley, MA: Newbury House, 1972.

[9] L.-J. Calvet, *La Guerre des Langues et les Politiques Linguistiques*. Paris: Payot, 1987.

[10] N. Ibrukhim, *Chinese Tajik Alphabet*. Ürümchi: Xinjiang University, 2012.

[11] J. Jilani, Tajik practical dictionary. New York: Hippocrene Books, 2009.

[12] J. R. Perry, *A Tajik Persian Reference Grammar*. Leiden: Brill, 2005.