Report on Kashmiri Yeh

To: Unicode Technical Committee
Author: Manish Goregaokar (मनीष गोरेगांवकर) <manish@unicode.org>
Date: 18 Nov 2023

This document outlines the situation around the Kashmiri Yeh, building on work done in [L2/23-241](), and talks about some potential solutions we can pursue.

Background

Kashmiri is an Indo-Aryan language that has traditional orthographies in multiple scripts, including Perso-Arabic. The Perso-Arabic orthography for Kashmiri is structured similarly to that of Urdu with some additional characters, including an additional "letter" to mark palatalization.

This character was proposed in [L2/09-215](), subsequently encoded as [U+0620]() ARABIC LETTER KASHMIRI YEH in Unicode 6.0.

Kashmiri uses a number of code points called YEH, all considered distinct with distinct purposes:

	Final	Medial	Initial	Isolated
[U+06CC]() ARABIC LETTER FARSI YEH	ی	ﻴ	ﻳ	ی
[U+0620]() ARABIC LETTER KASHMIRI YEH	ٯ	ﻬ	ﺋ	ﯼ
[U+06D2]() ARABIC LETTER YEH BARREE	ے			ے

Note that the final and isolated forms of FARSI YEH and KASHMIRI YEH are similar but KASHMIRI YEH has a much shorter swoosh.

ARABIC LETTER KASHMIRI YEH currently looks like this in various fonts:

U+0620	Final	Medial	Initial	Isolated
Noto Nastaliq (Google)	ڕ	ؠ	ؠ	ی
Scheherazade New (SIL)	ی	�ی	ﯨ	ی
Lateef (SIL)	ی	ﯨ	ﯨ	ی
Harmattan (SIL)	ی	ﯨ	ﯨ	ی
Noto Sans (Google)	ﯨ	ﯨ	ﯨ	ﯨ
Amiri (Khaled Hosny)	ﯨ	ﯨ	ﯨ	ﯨ
SF Arabic (Apple)	ﯨ	ﯨ	ﯨ	ﯨ
Microsoft Sans Serif (Microsoft)	ﯨ	ﯨ	ﯨ	ﯨ
Simplified Arabic (Microsoft)	ﯨ	ﯨ	ﯨ	ﯨ

The form of KASHMIRI YEH with the ring is usually called "Gol Ye" ("round ye"), whilst the form with the shorter swoosh and no ring is called "Taler Ye" ("half ye").

Most Naskh fonts use the Gol Ye form in final position ([all of the ones on Google Fonts](#) except the first four above). Kashmiri is typically written in the Nastaliq style where it seems like Taler Ye is consistently used at the end of the word. This is in line with the proposal in [L2/09-215](#), which suggests that Naskh and Nastaliq differ on how they ought to render this character in final form. This recommendation has made it into [Chapter 9 of the Unicode Standard](#):

Table 9-10. Forms of the Arabic Letter *yeh* (Continued)

Character	Joining Group	X_n	X_r	X_m	X_l
U+0620 KASHMIRI YEH	YEH	ئ	ي	ڊ	ڊ

Unicode also has a character [U+06CD](#) ARABIC LETTER YEH WITH TAIL for use in Pashto (since 1.1), only found at the end of words for the feminine marker. This was not encoded for Kashmiri.

	Final	Isolated
Noto Nastaliq	ٮ	ئ
Noto Sans	ے	ی

The current situation is that a lot of extant digital Kashmiri text uses the Pashto YEH WITH TAIL to represent Taler Ye at the end of words, because users expect the palatalization mark to take the Taler Ye form at the end of a word (i.e. the spec recommendation on rendering word-final KASHMIRI YEH as Gol Ye is incorrect). This Pashto letter is present in Kashmiri text in CLDR, in the output of Bing Translate, and on the Kashmiri GBoard keyboard available by default on Android.

This is a problem: it is not recognized by native speakers as the correct character, though the intent is understood.

This report documents findings on the situation so far; to figure out what needs to be done to improve the situation for Kashmiri speakers: whether we need to disunify Gol and Taler Ye, whether we need to recommend fonts and keyboards handle KASHMIRI YEH differently, or whether we need to do something else entirely.

Research on word-final Gol Ye

(See Appendix 1.)

Most sources seem to consistently use Taler Ye at the end of words, regardless of whether they are Naskh or Nastaliq.

There is an example of Gol Ye at the end of the word in Appendix 1A (in a context where the user has some control over presentation), and as noted previously [most fonts](#) render KASHMIRI YEH as Gol Ye at the end of words. However, the creator of that content felt that that rendering was incorrect.

When U+0620 was encoded, there was feedback from the Indian IT Ministry in [L2/10-169](#) that, based on expert opinion, stated that the final and isolated forms of Gol Ye are never seen, rather the Taler Ye form is always used in those positions.

It seems to be that word-final Gol Ye is not an expectation of Kashmiri users; but rather an artifice introduced by fonts incorrectly choosing to render KASHMIRI YEH as Gol Ye in the word-final form, based on Unicode's incorrect advice.

Research on word-medial Taler Ye

(See Appendix 2)

The person who had filed [CLDR-16595](#) brought up the possibility of word-medial Taler Ye, which would either require disunifying the Taler Ye form from KASHMIRI YEH, or potentially recommending use of a ZWNJ, since KASHMIRI YEH is supposed to dual-join and appear as a Gol Ye form in regular circumstances.

The vast majority of Taler Ye found in "word-medial" position is at a morpheme boundary, similar to how [Persian sometimes has joining breaks within words](#). Appendix 2D also contains examples of words with a joining break for letters other than Taler Ye, showing that this is a more general pattern. These cases are likely best served with a ZWNJ or even just a space if native users are amenable to that.

However, there are some examples of words where Taler Ye is found inside a word at a non-morpheme boundary, which can be found in Appendix 2A. These seem to invariably have the Taler Ye right before a regular Ye. An alternative orthography choice for these words seems to be to replace the Taler Ye with a kasra, as shown in Appendix 2E.

It appears to be that Gol Ye is not preferred in this position because Taler Ye is a pure palatalization mark, whereas Gol Ye is considered to have a vowel value as well, which sounds strange before another vowel/semivowel.

Research on isolated forms

(See Appendix 3)

Isolated forms of both Gol and Taler Ye are attested, sometimes in the same document, in a nominative/pedagogical context (Appendix 3A). Gol Ye is typically shown after an alef, using a modified YEH BARREE form.

The letters are often thought of as separate in pedagogical contexts. A solution to this issue will likely have to maintain a way to render both, and may even need to maintain compatibility with old documents.

Not all sources treat them as separate (Appendix 3B). These appear to be the sources consulted by the original proposal, which unfortunately also treats the two forms as variants of each other.

Isolated Gol Ye is not attested in contexts that are not pedagogical or nominative: Triggering an isolated form within a word would require it to be succeeded by a non-right-joining letter (such as U+0621 ARABIC LETTER HAMZA); which is not found in Kashmiri. As established in the previous section, Gol Ye is not really attested at the end of words so an isolated form has not been found triggered there either.

Isolated Taler Ye in non-pedagogical contexts is relatively common (Appendix 3C): Triggering an isolated form within a word would just require it to be preceded by a non-left-joining letter, which happens e.g. in the word

شُری.

Native user outreach

This situation brings up a fair number of questions. The author spoke to various literate native speakers to get an idea of the situation, with major discussions with three people but some supplemental data-gathering from others.

Could there ever be a Gol Ye at the end of a word?

On speaking to people, nobody believed that the Gol Ye at the end of the word is ever correct; however they felt that people tend to understand what was intended when they see a Gol Ye at the end of the word. Users seem to consider fonts with a Gol Ye final form to be a bug that they need to work around.

L2/09-215 claimed that Gol Ye at the end of the word is in Naskh, but native speakers who were aware of Naskh vs Nastaliq did not seem to concur.

Is the Pashto YEH WITH TAIL form acceptable?

No user considered the Pashto YEH WITH TAIL acceptable. They recognized that it was quite common as a workaround, but the user mental model around this character seems to be that it's better than having a Gol Ye, but it still looks wrong.

Is Taler Ye acceptable within a word?

Opinions were divided on this one. Two people felt that yes, there are cases where Taler Ye is found within a word, another person did not think it to be the case but when shown an example felt that it made sense. Everyone agreed that the non-morpheme examples from Appendix 2A were, in fact, "not a space". Many felt that the morpheme examples from Appendix 2C were just "separate" (though some felt that they should not have a space).

Both people who felt that Taler Ye is allowed within a word had examples I had not found myself that they shared. They explained that the pronunciation of Gol and Taler Ye are different; Gol Ye is [ʲa] (palatalization of a consonant and a vowel), whereas Taler Ye is [ʲ] (just palatalization of a consonant). IPA transcriptions are based on the author's best attempt, which seem to match the IPA found for Kashmiri elsewhere.

Persian Half Yeh

Roozbeh Pournader has been doing some research on a similar "Half Yeh" found in early and classical New Persian. We may end up needing to encode a separate character for this, which is relevant to the situation for KASHMIRI YEH if we decide to disunify it.

Appendix 4 contains some examples we have found so far.

Potential paths forward

The current situation is clearly suboptimal for Kashmiri users. Most fonts show an incorrect form of the palatalization mark at the end of words, and Kashmiri users have been reaching for a different also-incorrect-but-not-as-bad encoding to remedy this.

In general there seem to be two potential paths forward, not entirely mutually exclusive. It's important for these paths to not just consider Unicode but also figure out how this will work holistically for fonts, keyboards, people producing Kashmiri text, and systems containing Kashmiri text.

Stronger recommendations

One solution is to release some stronger recommendations about Kashmiri Yeh building on the work done in [L2/23-241](). This would involve updating annotations and the Unicode standard, potentially with some information about the history here, and introducing a new Joining_Group for KASHMIRI YEH.

For fonts, we would likely recommend that they render final and isolated KASHMIRI YEH as a Taler Ye, and declare the Gol Ye incorrect in word-final form. The isolated form must be a Taler Ye (since it can occur after a non-right-joining letter).

For keyboards, we should recommend they remove YEH WITH TAIL and instead maintain a single KASHMIRI YEH key, ideally rendered in a font that shows it as Taler Ye. Showing Gol Ye leads to people reaching for other characters to render Taler Ye. Other potential UI improvements would be to show both forms on the key (ی/ؠ), or have a separate key (with identical input) that has a Gol form.

Systems containing Kashmiri text should be asked to replace YEH WITH TAIL with KASHMIRI YEH.

People wishing to use word-medial Taler Ye may use ZWNJ to force the form.

This solution is relatively quick from a Unicode perspective, and once Unicode releases this work can be done in parallel.

Potential downsides are that this is somewhat suboptimal when it comes to word-medial Taler Ye, and if Persian Half Yeh is later encoded, we may have a potential confusable with people reaching for Persian Half Yeh instead of KASHMIRI YEH.

Furthermore, educational material may not be able to distinguish between the two forms. This may theoretically be ameliorated with some combining sequence, but we would need to design something that works, ZWJ or ZWNJ can't be used the way they're used for e.g. eyelash reph since they have an existing impact on Arabic

letters. Suggesting that ALEF + KASHMIRI YEH render as the attested pedagogical form for Gol Ye may help solve the issue. This does not help the nominative use case of referring to the character specifically rather than as a letter in a pedagogical context.

Even with such a change, font changes may negatively impact the rendering of existing pedagogical documents. This may disincentivize font vendors from quickly implementing this change.

Disunify KASHMIRI YEH

A more drastic measure is to disunify KASHMIRI YEH into KASHMIRI YEH (Gol Ye) and HALF YEH (Taler Ye). HALF YEH would be a new right-joining code point.

In such a case we can recommend keyboards expose both letters as keys, and recommend that fonts intending to support Kashmiri include the new letter.

There is an open question about compatibility: given that word-final Gol Ye seems to not be kosher, should we recommend fonts render final and isolated KASHMIRI YEH as Taler Ye for compatibility? This can lead to better text but it would somewhat persist the confusing case of character identity here, and may also impact pedagogical documents that care about distinguishing the forms.

On the plus side, this works well with any plans to encode a Persian Half Yeh (ideally we have one proposal talking about both).

Further reading

- [CLDR-16595](#) Multiple issues in CLDR ks.xml
- [L2/23-241](#) Annotations for Kashmiri Yeh
- [L2/09-215](#) Proposal to add two Kashmiri characters

- [L2/10-169](#) Comments on the Proposed ى Arabic Letter Kashmiri Yeh

- [Kashmiri orthography notes](#)
- [W3C alreq #249](#) Lack of font support for Kashmiri characters makes text deviate from true semantics

Appendices

Appendix 1: Word-final Gol/Taler Ye

A. Nastaliq

[हिंदी-कश्मीरी-अंग्रेज़ी त्रिभाषा कोश (Hindi-Kashmiri-English Trilingual Dictionary)](#):

[Kashmiri English Dictionary For Second Language Learners by Omkar Koul](#):

[Kashmiri Dictionary) - JK Cultural Academy](#) كأشُر لفظٕ راش:

[Kashmiri Romanization Table - Library of Congress](#):

[Kashir Dictionary) Vol 1](#) كأشُر ڈکشنَری (by Toushikhani S.k, Koul J.lal, Hajni Mohi-ud-din, Pushp P.n , Akhter Mohi-ud-din)

B. Naskh

These examples are from Instagram account @kaeshirlugat. One can see both Taler and Gol Ye forms. On speaking with the account author, they felt that the Gol Ye at the end of the word is incorrect, despite having used it.

(Instagram is a good source for this because it has a bunch of modern text written in a context where you do have control over presentation; the choice of rendering is relevant for understanding what people consider "acceptable")

أچھ كأنٕ جان تہ وَتھ كأنٕ نٕ

achh káni jaan tah wath káni nah

Better that the eye be blind than
that the way be blind

@kaeshirlugat

amaar chàne phir mae waraq lõlè kitãban
barè barè mae thawem seenè chaé almãre gulabo

امارِ چانے پھِرٕ مٕے وَرق لولٕ کِتابَن
بُرٕ بُرٕ مٕے تھاوٕم سینٕ چِے المارِ گولابو

Mahjoor

@proverbsinkashmiri

كأکی

käkⁱ

وَن تٕہ بے وِیسی بے ووفأیی شیوَے دِلدار چھا
نازنِیَن، ماہ جُبِیَن قتل و غارت کار چھا

Mahjoor

This example is from the @museumofkashmir account, using the Pashto YEH WITH TAIL:

وَلہ ویسی گژہ وَے بَبرے

Come, dear friend,
let's go and gather basil

museumofkashmir
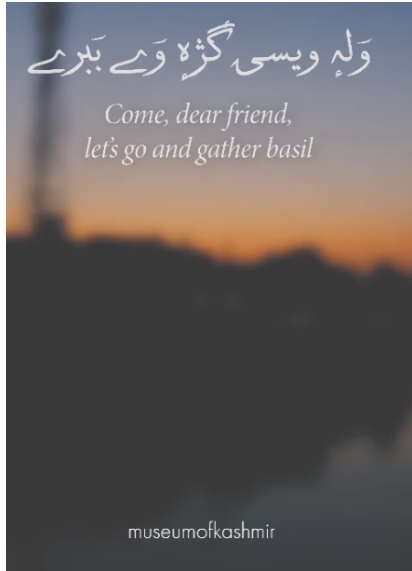
Appendix 2: Word-medial Taler Yes

A. At Non-Morpheme boundaries

بَنۍیان

[Kashmiri English Dictionary For Second Language Learners by Omkar Koul](#)

bAn'ya:n (f.s.) مَنۍیان   sweater
bAn'ya:nɪ (p)

كأشُر ڈکشنَری [(Kashir Dictionary) Vol 1](#)

ذَریا
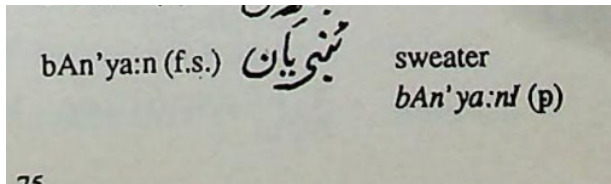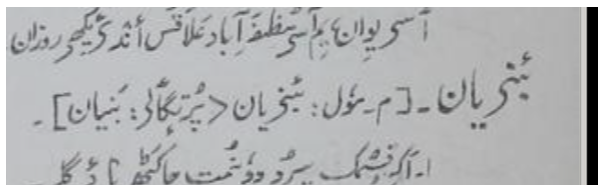
كأشُر لفظٕ راش [(Kashmiri Dictionary) - JK Cultural Academy](#):
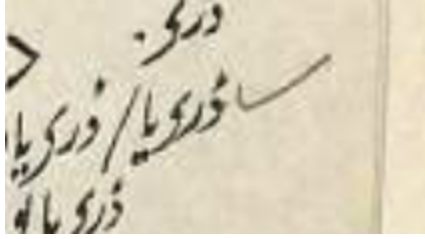
کٲشُر ڈکشنَری (Kashir Dictionary) Vol 3:
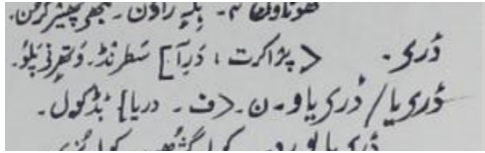
دُنْیا

हिंदी-कश्मीरी-अंग्रेजी त्रिभाषा कोश (Hindi-Kashmiri-English Trilingual Dictionary):

| दुनिया | स्त्री. 1. संसार, जगत; | دُنِیا دۅنْیِا; | the world; |

کٲشُر ڈکشنَری (Kashir Dictionary) Vol 3:

وَسِیتھ
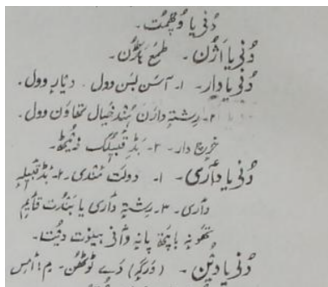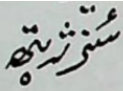
(Note: There is no morpheme boundary here: the word is a loan from Arabic وصیة)
हिंदी-कश्मीरी-अंग्रेजी त्रिभाषा कोश (Hindi-Kashmiri-English Trilingual Dictionary):

वृस्यहथ, वस्ययथः   will;

B. Unclear if at morpheme boundary

(sources same as previous)
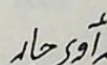
sAn'ts'ath (m.)	مُنْثَ ثَہ بِتھ	saving

C. At Morpheme boundaries
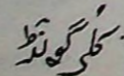
(sources same as previous)

(m.) gul'm'u:Th (m.)		wedding present
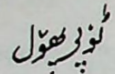
bil'bicur (m.)		bulbul, nightingale be

zA:v'ja:r (m.)		fineness

Kashmiri-English Dictionary

Du:n'kul (n)		walnut tree

kul' g~D (m.)		

gu:r'ba:y (f.s.)		milkwoman *gu:r'ba:yi* (p)

Tu:p' (f.s.)		a cap *To:pi* (p)
Tu:p'phol (m.s.)		button *Tu:p'phAl'* (p)

TAT' (f.)		lavatory *Taci* (p)
TAT'b'o:l (m.s.)		cotton seeds
TAT'vo:l (m.s.)		pony man

pAz' kin' (ad)		really, in fact
pAz'yo:r (a)		a person who speaks truth
pAz' pA:Th' (ad)		truly, actually, really

garrulous.

कथ् कुतॉल्यगी   talkativeness, loquaciousness, garrulity.
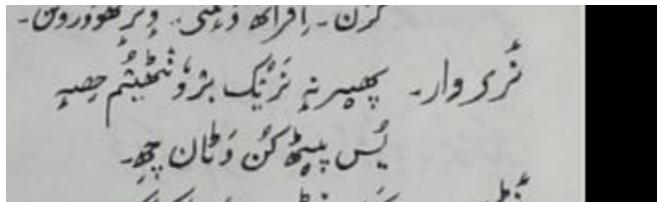
खुर्पदार   complicated intricate.

खॉर्यजियथ   objectivism, realism.

शुर्यबान;   womb;

کأشِر ڈکشنَری (Kashir Dictionary) Vol 6:

D. Non-Ye breaks at morpheme boundaries

(sources same as previous)

cha:nIba:y (f.s.) چھائنہ باے a carpenter's wife
cha:nIba:yi (p)

'ajA:yibgar عجائیب گھر museum
(m.)

فرقہ پرستی फिरकुपरस्ती communalism, sectarianism.
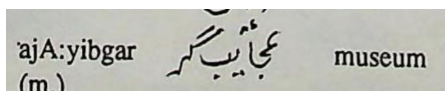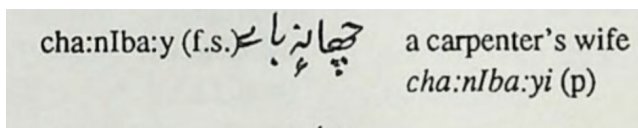
E: Alternative orthography choice for medial Taler Ye + Regular Ye

دریاو दरियाब river (table)
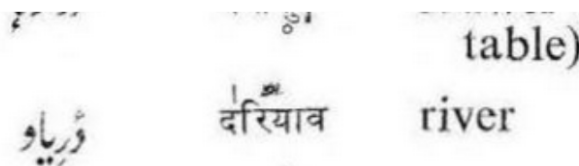
Kashmiri Romanization Table - Library of Congress:

				ý (see Note 1)
-	ٮٕ	ٛى	لٖ	

1. This sign palatalizes the preceding consonant and in the medial position is at times
    followed by another vowel, e.g. *a* (ٓ◌) or *e* (◌ٔ).
    apẏar آپیَر
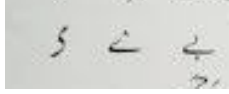   When the short vowel *e* (◌ٔ) follows a palatalized consonant, replace it by ◌.
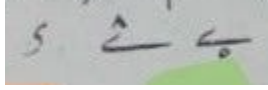   khẏen کِھین

Appendix 3: Isolated forms

A: Sources showing both isolated forms as separate concepts

كأثُر ڈکشنَری [(Kashir Dictionary) Vol 1](#)

(Source unknown, found on [CLDR-16595](#), possibly same as previous)

[Kashmiridictionary.org](#)

[Kashmiri Class III textbook](#)

B: Sources that merge the two or only talk of one

-	ـہؕ	ؠ	لٟ	ẏ (see Note 1)

(This document predates Unicode's encoding, which is why it doesn't list codepoints for the Kashmiri yeh)
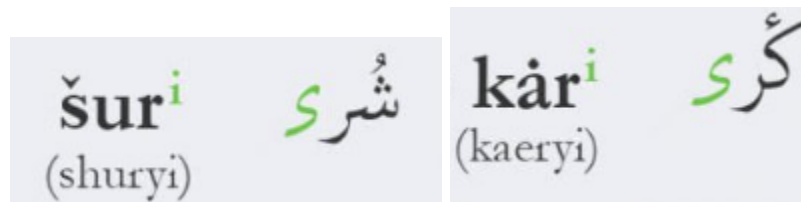
3.6. Kashmiri

Kashmiri employs the Urdu alphabet with a few additions to represent its specific vowels. Kashmiri has two additional Yehs (ی), one with an oval below (ؠ) and the other with a 'v' mark above (ٟ). It also has two additional Waws (و), one with a circle at the ending tail (ۄ) and the other with a 'v' mark above (ۆ). In diacritical marks, it adds two diacritical marks, a slightly modified Hamza (ٔ) written above and below the character. The extra characters of Kashmiri are shown in Table 10. It is also traditionally written in Nasta'leeq style.

Sr.	Symbol	Unicode	Sr.	Symbol	Unicode
1	ؠ []	-	4	ٟ [e]	06CE
2	ۄ [ɔ]	06C4	5	ٛ [ə]	-
3	ۆ [o:]	06C6		ٟ []	-

Table 10: Kashmiri characters

C: Isolated forms of Taler Ye

šurⁱ (shuryı)	شُری	kảrⁱ (kaeryı)	کٛری

ثُری

[हिंदी-कश्मीरी-अंग्रेजी त्रिभाषा कोश (Hindi-Kashmiri-English Trilingual Dictionary)](#)
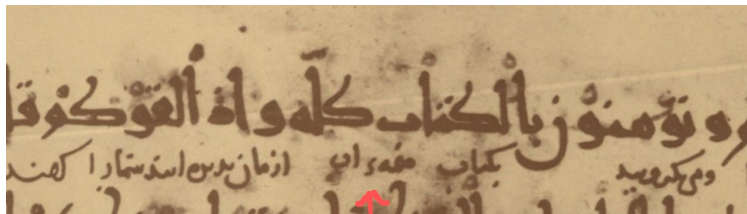
واسکٹ، سٔدری   वासकठ, संदुर्य   waist-coat.

سوأری   सवॉर्य

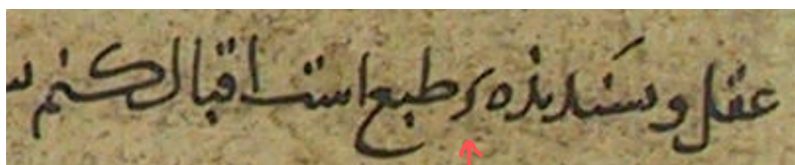[Kashmiri English Dictionary For Second Language Learners](#)

tapA:r' (ad)	تیٲری	in that direction
tapA:r (f.s.)	تیٲری	a kind of abuse

Appendix 4: Persian Half Yehs
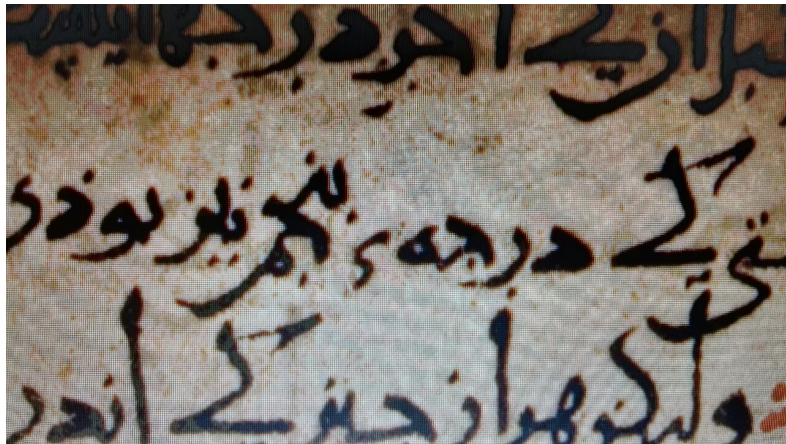
p. 49: ویژگیهای املائی دستنویس تفسیر قرآن پاک

و قصّه چِ و آن' ، ۳:۱۰

و قصّه ۽ این ، ۴۲:۱۱ ؛ ۳۳:۲۰

و قصّه چِ این ، ۱۲:۱۰

قصّه ۽ انکشتیره چِ سلیمن ، ۳۴:۱۰

اما قصّه چِ کشتن ، ۹:۱۸

The manuscript version of the above, from ТОЛКОВАНИЕ КОРАНА (ЛАХОРСКИЙ ТАФСИР), p. 291