Title: Revised Proposal to Encode Gujarati Signs for the Transliteration of Arabic
Source: Script Encoding Initiative (SEI)
Author: Anshuman Pandey (anshuman.pandey@berkeley.edu)
Date: 2015-03-30

1 Introduction

This is a proposal to encode six additional characters in the 'Gujarati' block of the Unicode standard:

GLYPH	CODE	CHARACTER NAME
◌ઁ	0AFA	GUJARATI SIGN SUKUN
◌ં	0AFB	GUJARATI SIGN SHADDA
◌ઃ	0AFC	GUJARATI SIGN MADDAH
◌઄	0AFD	GUJARATI SIGN THREE-DOT NUKTA ABOVE
◌અ	0AFE	GUJARATI SIGN CIRCLE NUKTA ABOVE
◌આ	0AFF	GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE

The characters are proposed for contiguous allocation in the six empty code points at the end of the 'Gujarati' block after the approved, but not yet encoded character ૹ U+0AF9 GUJARATI LETTER ZHA. The location is suitable because GUJARATI LETTER ZHA, like the six characters proposed here, is also used for transliteration.

The characters were previously discussed in the "Proposal to Encode Gujarati Signs for the Transliteration of Arabic in ISO/IEC 10646" (L2/14-131). The major difference between the present proposal and the previous document is the inclusion of additional information for characters that are not presently proposed for inclusion (see section 6).

2 Background

The proposed signs are used for the transliteration of the Arabic script into Gujarati by Ismaili Khoja communities. They are used for representing Arabic letters and signs for which correspondences do not exist in Gujarati. They were devised in the late 19th century and are standard elements of the Gujarati orthography used by the Ithnashari Khoja ("Twelver Shia") and the Agakhani Khoja communities. The creation of the full set and the first documented printing of these signs was undertaken by the Ithnashari Khoja publisher Gulāmalī Ismāʾil of Bhavnagar, Gujarat in 1901. The signs are used in manuscripts and in printed materials, predominantly in Khoja texts such as the "Agakhani Dua" and Gujarati-script versions of the *Qurʾān*.

3 Encoding Model

The ◌ઁ GUJARATI SIGN SUKUN, ◌ઁ GUJARATI SIGN SHADDA, and ◌ઁ GUJARATI SIGN MADDAH occur with several base letters and combining signs; however, in the available Ithnashari sources the distribution of the three *nukta* signs is limited to certain base letters. The ◌ઁ GUJARATI SIGN THREE-DOT NUKTA ABOVE is used with several base letters, but ◌ઁ GUJARATI SIGN CIRCLE NUKTA ABOVE occurs with only two consonants and ◌ઁ GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE occurs with only one letter. There are two approaches to encoding these latter two *nukta* signs: 1) as combining signs, or 2) as atomic characters consisting of each sign combined with a respective base letter. This proposal recommends the first option in order to provide flexibility in the usage of these signs.

The ◌ઁ GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE is used in Ithnashari orthography only in combination with ઝ GUJARATI LETTER JHA, in the form ઝઁ, for the transliteration of ظ U+0638 ARABIC LETTER ZAH, the pharyngealized voiced dental fricative [ðˤ]. Given this limited usage it may be practical to encode ઝઁ as an atomic character, ie. *GUJARATI LETTER ZAH. The case is identical to that of ૹ GUJARATI LETTER ZHA, which was proposed by Vinodh Rajan in May 2013 and approved by the Unicode Technical Committee (UTC) for future encoding at U+0AF9 in the Gujarati block (see L2/13-143). The letter ૹ is used for transliterating 𐬚 U+10B32 AVESTAN LETTER ZHE. In principle, ૹ GUJARATI LETTER ZHA is a precomposed character consisting of the base જ U+0A9C GUJARATI LETTER JA combined with the sign ◌઼, which is a three dot variation of ◌઼ U+0ABC GUJARATI SIGN NUKTA. Ideally, the UTC would have considered encoding this ◌઼ three-dot *nukta* as a combining mark instead of encoding ૹ GUJARATI LETTER ZHA as an atomic character. This would have made it possible to use ◌઼ with other base letters, if needed. Indeed, Rajan had initially proposed the encoding of ◌઼ as *GUJARATI SIGN TRIPLE NUKTA so that ૹ might be represented as sequence of a base letter and combining mark (see L2/13-066). But, it seems that ૹ was encoded as the atomic GUJARATI LETTER ZHA on account of its correspondence to ज़ U+0979 DEVANAGARI LETTER ZHA, which is also used for transliterating Avestan 𐬚 ZHE. The encoding of ૹ as an atomic letter, however, does not account for the possibility that ◌઼ may occur in some source with another letter, eg. ઝ઼ <GUJARATI LETTER JA, *three-dot nukta below*>. The discovery of such a case would necessitate either the encoding of that combination as a separate atomic character or, ultimately, the encoding of ◌઼ as a combining mark. For this reason, although ◌ઁ GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE occurs in Ithnashari orthography with only ઝ GUJARATI LETTER JHA, it is recommended that it, along with ◌ઁ GUJARATI SIGN CIRCLE NUKTA ABOVE, be encoded as a combining sign instead of as atomic compositions with base letters in order to account for the possibility that they may be used with other base letters in other sources.

4 Characters Proposed

4.1 GUJARATI SIGN SUKUN

The ◌ઁ GUJARATI SIGN SUKUN is used for indicating a pause during recitation. It occurs, for example, in the word وَلٰكِنْ / વલાકિંન્, highlighted in dark green in figures 1 and 2. The sign may be used with both vowel and consonant letters, but not with vowel signs. It may represent a bare consonant when it occurs with a consonant letter. When used for the latter, it behaves similar to ◌્ U+0ACD GUJARATI SIGN VIRAMA in its function of silencing the inherent vowel of a consonant, but it does not possess the control properties of VIRAMA. The Gujarati SUKUN corresponds to ◌ْ U+0652 ARABIC SUKUN. It is used in encoded text as follows:

અ્ <અ GUJARATI LETTER A, ◌ઁ *sukun*>

સ્ <સ GUJARATI LETTER SA, ◌ઁ *sukun*>

4.2 GUJARATI SIGN SHADDA

The ◌ૐ GUJARATI SIGN SHADDA is used for marking consonant gemination. It occurs, for example, in the word اِنّ / ઈૐન, highlighted in cyan in figures 1 and 2. It is identical to ◌𑍍 U+11237 KHOJKI SIGN SHADDA, and both are based upon ◌ّ U+0651 ARABIC SHADDA. Although such usage is redundant, it is written above doubled consonants represented as conjuncts, where it is placed above both the half-letter and the regular letter.

સ્સ <સ GUJARATI LETTER SA, ◌ૐ *shadda*>

સ્સ્સ <સ GUJARATI LETTER SA, ◌ૐ *shadda*, ◌્ VIRAMA, સ GUJARATI LETTER SA, ◌ૐ *shadda*>

4.3 GUJARATI SIGN MADDAH

The ◌ૅ GUJARATI SIGN MADDAH represents the elongation of a vowel during recitation of a text. It occurs, for example, in the word أُولَٰئِكَ / ઓલૅાએક, highlighted in red in figures 1 and 2. The sign may be used with vowel and consonant letters and vowel signs. It is modeled after ◌ٓ U+0653 ARABIC MADDAH ABOVE.

કૅ <ક GUJARATI LETTER KA, ◌ૅ *maddah*>

કૅા <ક GUJARATI LETTER KA, ◌ા GUJARATI VOWEL SIGN AA, ◌ૅ *maddah*>

કૅુ <ક GUJARATI LETTER KA, ◌ુ GUJARATI VOWEL SIGN U, ◌ૅ *maddah*>

4.4 GUJARATI SIGN THREE-DOT NUKTA ABOVE

The ◌૮ GUJARATI SIGN THREE-DOT NUKTA ABOVE is used for representing the Arabic letters shown below. It occurs, for example, in the word مَرَضٌ / મરઝૂનૅ, highlighted in purple in figures 1 and 2. It may be used with vowel and consonant letters. The sign corresponds to ◌𑍎 U+11236 KHOJKI SIGN NUKTA.

અ૮	<અ GUJARATI LETTER A, ◌૮ *three-dot nukta above*>	=	ع ARABIC LETTER AIN
ઈ૮	<ઈ GUJARATI LETTER I, ◌૮ *three-dot nukta above*>	=	ع ARABIC LETTER AIN
ક૮	<ક GUJARATI LETTER KA, ◌૮ *three-dot nukta above*>	=	ق ARABIC LETTER QAF
ખ૮	<ખ GUJARATI LETTER KHA, ◌૮ *three-dot nukta above*>	=	خ ARABIC LETTER KHAH
ગ૮	<ગ GUJARATI LETTER GA, ◌૮ *three-dot nukta above*>	=	غ ARABIC LETTER GHAIN
ત૮	<ત GUJARATI LETTER TA, ◌૮ *three-dot nukta above*>	=	ط ARABIC LETTER TAH
સ૮	<સ GUJARATI LETTER SA, ◌૮ *three-dot nukta above*>	=	ص ARABIC LETTER SAD
ઝ૮	<ઝ GUJARATI LETTER JHA, ◌૮ *three-dot nukta above*>	=	ض ARABIC LETTER DAD

When both *three-dot nukta above* and a dependent vowel sign occur together with a base letter, the *three-dot nukta above* is placed before the vowel sign in the encoded sequence. The rationale for the encoding order

is that, identical to the regular GUJARATI SIGN NUKTA, the *three-dot nukta above* is a consonant modifier and its usage affects the value of the consonant independently of any accompanying vowel sign.

કુ̣̈ <ક GUJARATI LETTER KA, ◌̣̈ *three-dot nukta above*, ◌ુ GUJARATI VOWEL SIGN U>

4.5 GUJARATI SIGN CIRCLE NUKTA ABOVE

The ◌̊ GUJARATI SIGN CIRCLE NUKTA ABOVE is used for representing the two Arabic letters shown below. It occurs, for example, in the word وَاِذَا (وَإِذَا) / વએઝ઼ા, highlighted in blue in figures 1 and 2.

સ઼ <સ GUJARATI LETTER SA, ◌̊ *circle nukta above*> = ث ARABIC LETTER THEH

ઝ઼ <ઝ GUJARATI LETTER JHA, ◌̊ *circle nukta above*> = ذ ARABIC LETTER THAL

4.6 GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE

The ◌̊̊ GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE is used for representing the Arabic letter shown below. It occurs, for example, in the word عَظِيمْ / અ઼ઝ઼ીમ, highlighted in light green in figures 1 and 2.

ઝ઼ <ઝ GUJARATI LETTER JHA, ◌̊̊ *two-circle nukta above*> = ظ ARABIC LETTER ZAH

5 Considerations for Rendering

More than one of the proposed marks may occur with a single base letter. In such cases it is necessary to adjust the placement of the signs in order to prevent clashing. Generally, the signs are placed horizontally.

Glyph ordering When *sukun* occurs together with one of the three above-base *nukta* signs or with *shadda*, it is positioned to the right of these signs in the output, but is placed before them in the encoded sequence:

અ઼̣̈ <અ GUJARATI LETTER A, ◌̣̈ *three-dot nukta above*, ◌̣̇ *sukun*>

ઝ઼̣̇ <ઝ GUJARATI LETTER JHA, ◌̊̊ *two-circle nukta above*, ◌̣̇ *sukun*>

સ઼̣̇ <સ GUJARATI LETTER SA, ◌̣̇ *shadda*, ◌̣̇ *sukun*>

The same principle applies to the co-occurrence of *shadda* and one of the three above-base *nukta* signs. The *shadda* is positioned to the right of these *nukta*-s, but is placed before them in the encoded sequence:

અ઼̣̈ <અ GUJARATI LETTER A, ◌̣̈ *three-dot nukta above*, ◌̣̇ *shadda*>

ઝ઼̣̇ <ઝ GUJARATI LETTER JHA, ◌̊̊ *two-circle nukta above*, ◌̣̇ *shadda*>

Glyph adjustments When ◌ૅ *three-dot nukta above* occurs with vowel signs that extend above and over the body of the consonant — ◌ે U+0AC7 GUJARATI VOWEL SIGN E, ◌ૈ U+0AC8 GUJARATI VOWEL SIGN AI, ◌ો U+0ACB GUJARATI VOWEL SIGN O, ◌ૌ U+0ACC GUJARATI VOWEL SIGN AU — it is reduced in size and rotated 15° clockwise in order to prevent clashing and to accommodate fit.

સૅ → સૅ <સ GUJARATI LETTER SA, ◌ૅ *three-dot nukta above*, ◌ે GUJARATI VOWEL SIGN E>

The same size reduction takes places when the THREE-DOT NUKTA occurs with ◌્ SUKUN and the width of the two signs exceeds the width of the base character:

હ્ૅ → હ્ૅ <હ GUJARATI LETTER HA, ◌ૅ *three-dot nukta above*, ◌્ *sukun*>

6 Characters Not Proposed

Section mark The sources show a mark resembling ❋, which is used as a sectioning mark. It corresponds to the ۝ U+06DD ARABIC END OF AYAH used in the parallel Arabic. The ❋ occurs in four different contexts: 1) independently; 2) followed by one or more digits, eg. ❋ ૩૭, for indicating the numbering of a text passage; 3) with a superscript consonant letter; and 4) with a superscript consonant letter and following digits. The source shows two superscript letters that occur with ❋: અ GUJARATI LETTER A and મ GUJARATI LETTER MA, written above the mark: ❋ and ❋. The ❋ section mark is not proposed for encoding as a separate character at present. It may be considered a glyphic variant of ⋆ U+066D ARABIC FIVE POINTED STAR or ⁕ U+2055 FLOWER PUNCTUATION MARK. These characters should be used, with any necessary changes to the glyphs, for representing the section mark.

Superscript letters As shown in the sources, the superscript forms of અ GUJARATI LETTER A and મ GUJARATI LETTER MA are used, respectively, for transliterating ◌ٔ U+0654 ARABIC HAMZA ABOVE and ◌ۢ U+06D8 ARABIC SMALL HIGH MEEM INITIAL FORM when the latter occur with ۝ ARABIC END OF AYAH. While these superscript characters are suitable candidates for encoding, they are not proposed for inclusion at present because the complete repertoire such characters has not yet been identified. Moreover, the usage of the superscript letters is not yet entirely understood. The sources indicate that not every Koranic annotation sign in Arabic is represented in Gujarati; for instance, the Arabic text in figure 2 shows usage of ◌ٕ U+0615 ARABIC SMALL HIGH TAH (line 6), but there is no corresponding Gujarati sign for it in figure 1.

Superscript characters in various Indic scripts are often used in footnotes and other contexts. A decision to encode superscript Gujarati letters, whether as independent letters or combining signs, should take into consideration the requirements for the broader usage of such letters within Gujarati orthography. Until more research is conducted on the usage of superscript letters in Gujarati, the Gujarati annotation marks for the Koran should be managed using text layout and formatting. ’

7 Character Data

Character properties Character properties given in the format of `UnicodeData.txt`:

```
0AFA;GUJARATI SIGN SUKUN;Mn;0;NSM;;;;;N;;;;;
0AFB;GUJARATI SIGN SHADDA;Mn;0;NSM;;;;;N;;;;;
0AFC;GUJARATI SIGN MADDAH;Mn;0;NSM;;;;;N;;;;;
0AFD;GUJARATI SIGN THREE-DOT NUKTA ABOVE;Mn;7;NSM;;;;;N;;;;;
0AFE;GUJARATI SIGN CIRCLE NUKTA ABOVE;Mn;7;NSM;;;;;N;;;;;
```

```
0AFF;GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE;Mn;7;NSM;;;;;N;;;;;
```

Linebreaking properties Linebreaking properties given in the format of `LineBreak.txt`:

```
0AFA;CM # GUJARATI SIGN SUKUN
0AFB;CM # GUJARATI SIGN SHADDA
0AFC;CM # GUJARATI SIGN MADDAH
0AFD;CM # GUJARATI SIGN THREE-DOT NUKTA ABOVE
0AFE;CM # GUJARATI SIGN CIRCLE NUKTA ABOVE
0AFF;CM # GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE
```

Syllabic categories Syllabic categories given in the format of `IndicSyllabicCategory.txt`:

```
# Indic_Syllabic_Category=Nukta
0AFD          ; Nukta              # Mn       GUJARATI SIGN THREE-DOT NUKTA ABOVE
0AFE          ; Nukta              # Mn       GUJARATI SIGN CIRCLE NUKTA ABOVE
0AFF          ; Nukta              # Mn       GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE

# Indic_Syllabic_Category=Gemination_Mark
0AFB          ; Gemination_Mark    # Mn       GUJARATI SIGN SHADDA

# Indic_Syllabic_Category=Cantillation_Mark
0AFA          ; Cantillation_Mark  # Mn       GUJARATI SIGN SUKUN
0AFC          ; Cantillation_Mark  # Mn       GUJARATI SIGN MADDAH
```

Positional categories Matra categories given in the format of `IndicPositionalCategory.txt`:

```
# Indic_Positional_Category=Top
0AFA..0AFF    ; Top               # Mn   [6] SIGN SUKUN .. SIGN TWO-CIRCLE NUKTA ABOVE
```

Names List Names list information in the format of `NamesList.txt`:

```
@         Signs for Arabic transliteration
0AFA      GUJARATI SIGN SUKUN
          * indicates a pause in recitation
          * also used for marking a vowel-less consonant
0AFB      GUJARATI SIGN SHADDA
          * used for marking a geminated consonant
0AFC      GUJARATI SIGN MADDAH
          * indicates vowel elongation during recitation
0AFD      GUJARATI SIGN THREE-DOT NUKTA ABOVE
0AFE      GUJARATI SIGN CIRCLE NUKTA ABOVE
0AFF      GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE
```

8 References

Ismāʾil, Gulāmalī. 1901–1903. *Anvārūl bayān phī taphsīril kuraʾān*. 3 vols. Bhāvanagar: Isnāʾaśarī ilēkṭrīk prīnṭīng prēs.

Pandey, Anshuman. 2014. “Proposal to Encode Gujarati Signs for the Transliteration of Arabic in ISO/IEC 10646”. N4574 L2/14-131. `http://www.unicode.org/L2/L2014/14131-gujarati-arabic.pdf`

Rajan, Vinodh. 2013a. “Proposal to Encode Gujarati Sign Triple Nukta”. L2/13-066. `http://www.unicode.org/L2/L2013/13066-gujarati-triple-nukta.pdf`

———. 2013b. "Proposal to Encode Gujarati Letter ZHA". L2/13-143. http://www.unicode.org/L2/L2013/13143-gujarati-zha.pdf

9 Acknowledgments

I am grateful to Iqbal Akhtar (Florida International University, Miami) for bringing these Gujarati signs to my attention and for providing specimens of their usage. I would also like to thank Roozbeh Pournader (Google) for offering insights into the Arabic orthography used in the *Qurʾān* and for providing feedback regarding the encoding of the characters proposed here.

This project was made possible in part through a Google Research Award, granted to Deborah Anderson for the Script Encoding Initiative, and a grant from the United States National Endowment for the Humanities (PR-50205-15), which funds the Universal Scripts Project (part of the Script Encoding Initiative at the University of California, Berkeley). Any views, findings, conclusions or recommendations expressed in this publication do not necessarily reflect those of Google or the National Endowment for the Humanities.

Copyright © 2015 Anshuman Pandey. All rights reserved.

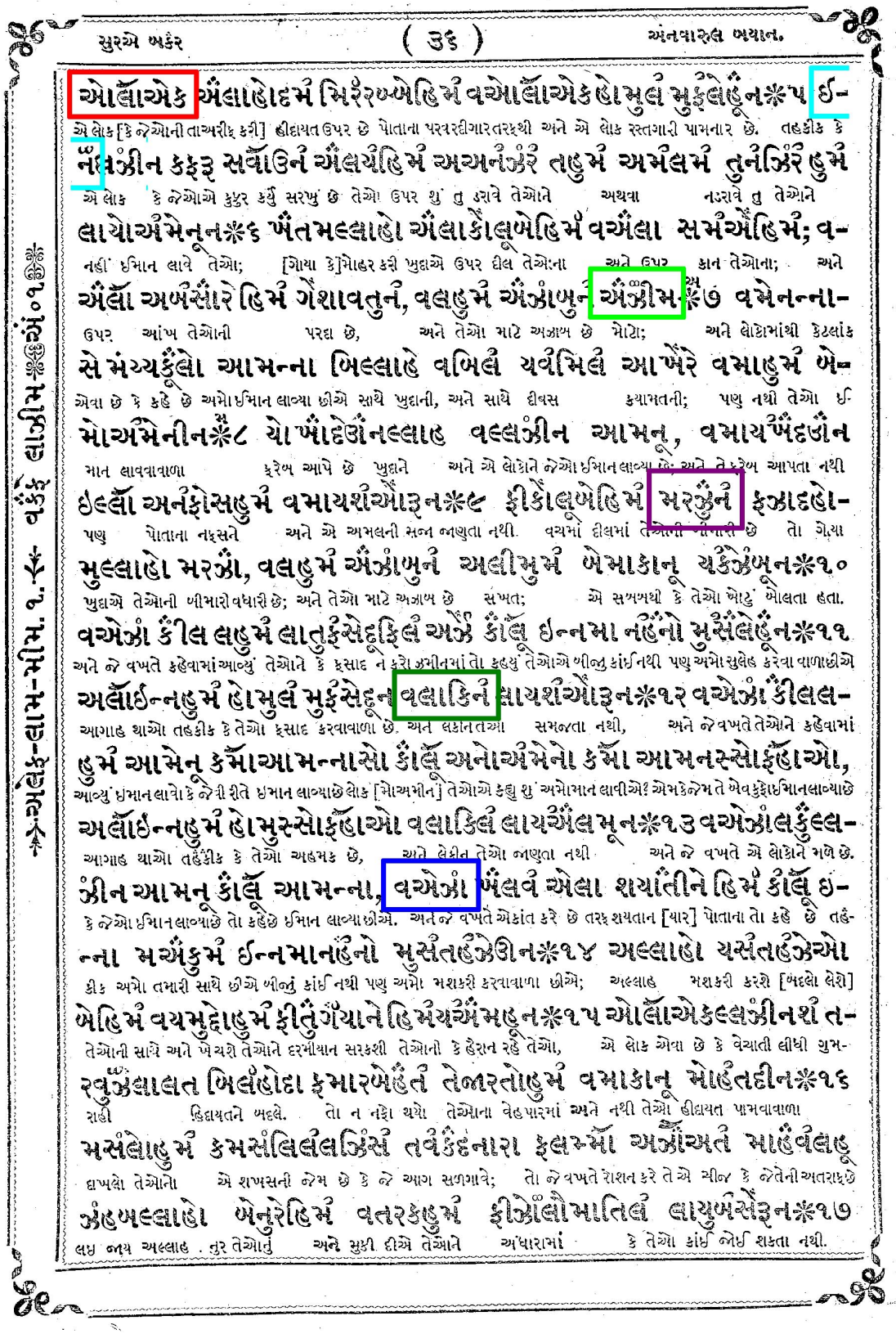

Figure 1: Page of the *Qur'ān* in Gujarati showing usage of the Arabic transliteration marks (from Gulāmalī Ismā'il 1901: 36). Arabic parallel shown in figure 2.

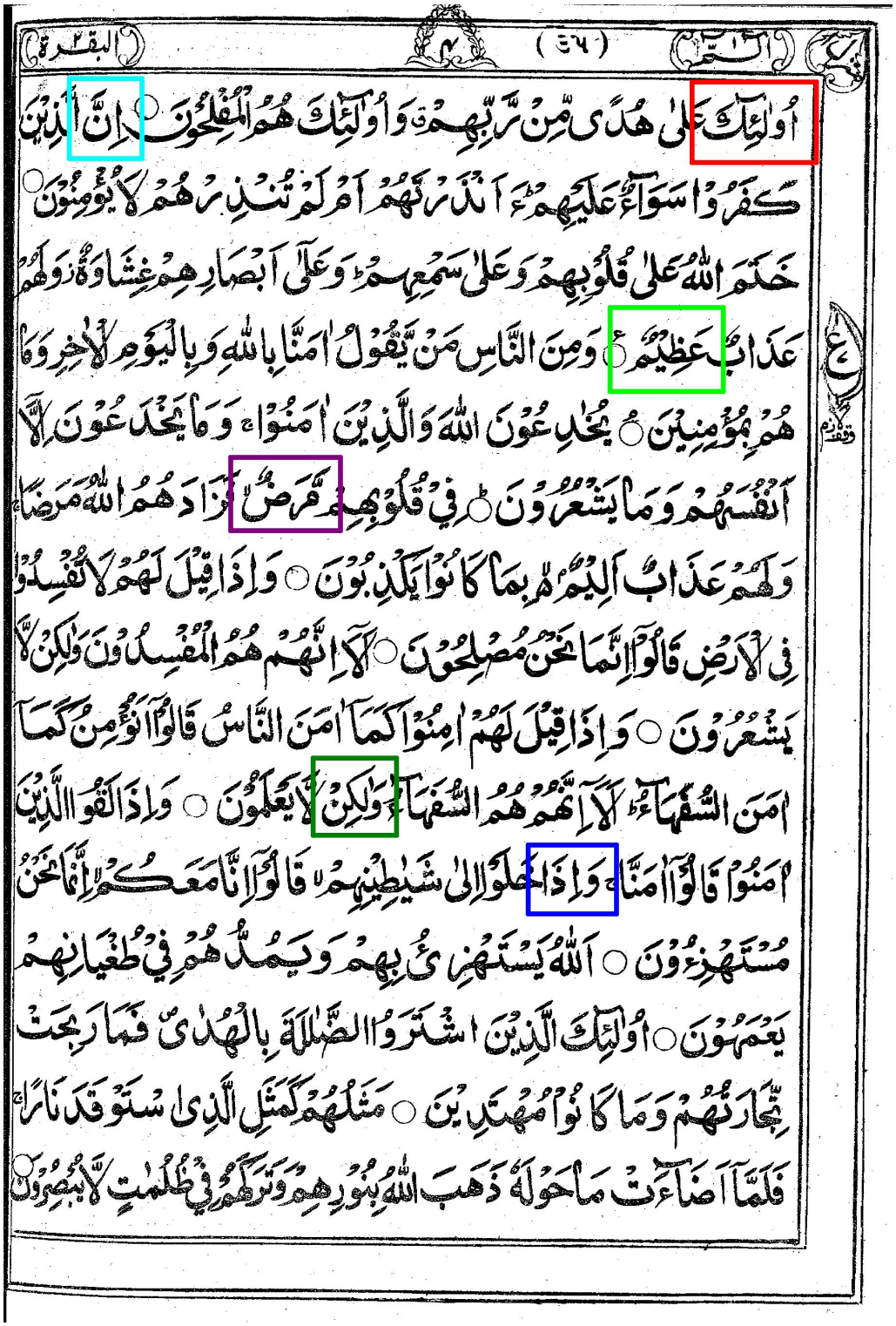

Figure 2: Page of the *Qurʾān* in Arabic (from Gulāmalī Ismāʾil 1901: 35). Gujarati parallel shown in figure 1.