Title: Proposal to Encode Gujarati Signs for the Transliteration of Arabic in ISO/IEC 10646
Source: Script Encoding Initiative (SEI)
Author: Anshuman Pandey (pandey@umich.edu)
Status: Liaison Contribution
Action: For consideration by UTC and WG2
Date: 2014-05-02

1 Introduction

This is a proposal to encode six additional characters in the 'Gujarati' block of the Universal Character Set (ISO/IEC 10646):

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 used for transliteration.

2 Description

These 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*.

The ◌ઃ GUJARATI SIGN SUKUN, ◌ૐ GUJARATI SIGN SHADDA, and ◌ૅ GUJARATI SIGN MADDAH occur with several base letters; however, the distribution of the three *nukta* signs is limited to particular base letters. In the available Ithnashari sources 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 ઝૼ, 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 ઝૼ as an atomic character, ie. *GUJARATI LETTER ARABIC ZAH. The situation 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. ઝૻ. 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.

3 Characters Proposed

3.1 GUJARATI SIGN SUKUN

The ◌ઃ GUJARATI SIGN SUKUN is used for indicating a pause during recitation. It can occur with vowel letters and consonants. When it occurs with a consonant, it can also represent a vowel-less consonant. 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:

અઃ <અ A, ◌ઃ *sukun*>

સઃ <સ SA, ◌ઃ *sukun*>

The *sukun* occurs, for example, in the phrase وَلٰكِنْ / વલાકિંઃ, highlighted in dark green in figures 1 and 2.

3.2 GUJARATI SIGN SHADDA

The ◌ૅ GUJARATI SIGN SHADDA represents consonant gemination. It is modelled upon ◌ّ U+0651 ARABIC SHADDA and corresponds to ◌𑍍 U+11237 KHOJKI SIGN SHADDA. It can also occur in conjuncts above half-letters.

સ્સ <સ SA, ◌ૅ *shadda*>

સ્સ્સ <સ SA, ◌ૅ *shadda*, ◌્ VIRAMA, સ SA, ◌ૅ *shadda*>

The *shadda* occurs, for example, in the phrase اِنَّ / ઇનّ, highlighted in cyan in figures 1 and 2.

3.3 GUJARATI SIGN MADDAH

The ◌૆ GUJARATI SIGN MADDAH represents the lengthening or sustain of a vowel during recitation. It is modelled upon the ◌ٓ U+0653 ARABIC MADDAH ABOVE. It is used in encoded text as follows:

ક૆ <ક KA, ◌૆ *maddah*>

કા૆ <ક KA, ◌ા VOWEL SIGN AA, ◌૆ *maddah*>

કુ૆ <ક KA, ◌ુ VOWEL SIGN U, ◌૆ *maddah*>

The *maddah* occurs, for example, in the word أُولَٰئِكَ / ઓલ૆ાએક, highlighted in red in figures 1 and 2.

3.4 GUJARATI SIGN THREE-DOT NUKTA ABOVE

The ◌ૈ GUJARATI SIGN THREE-DOT NUKTA ABOVE is used for representing the Arabic letters shown below. It can occur with vowel and consonant letters. The sign corresponds to ◌𑍌 U+11236 KHOJKI SIGN NUKTA.

Arabic	Gujarati	Encoded sequence
ع AIN	અૈ	<અ A, ◌ૈ *three-dot nukta above*>
ع AIN	ઇૈ	<ઇ I, ◌ૈ *three-dot nukta above*>
ق QAF	કૈ	<ક KA, ◌ૈ *three-dot nukta above*>
خ KHAH	ખૈ	<ખ KHA, ◌ૈ *three-dot nukta above*>
غ GHAIN	ગૈ	<ગ GA, ◌ૈ *three-dot nukta above*>
ط TAH	તૈ	<ત TA, ◌ૈ *three-dot nukta above*>
ص SAD	સૈ	<સ SA, ◌ૈ *three-dot nukta above*>
ض DAD	ઝૈ	<ઝ JHA, ◌ૈ *three-dot nukta above*>

The *three-dot nukta* occurs, for example, in the word مَرَضٌ / મરઝૈનું, highlighted in purple in figures 1 and 2.

When both *three-dot nukta above* and a dependent vowel sign occur with a base letter, the *three-dot nukta above* occurs before the vowel sign. 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.

કૃ̇ <ક KA, ◌̇ *three-dot nukta above*, ◌ુ VOWEL SIGN U>

3.5 GUJARATI SIGN CIRCLE NUKTA ABOVE

The ◌̊ GUJARATI SIGN CIRCLE NUKTA ABOVE is used for representing the following two Arabic letters:

Arabic	Gujarati	Encoded sequence
ث THEH	સ̊	<સ SA, ◌̊ *circle nukta above*>
ذ THAL	ઝ̊	<ઝ JHA, ◌̊ *circle nukta above*>

The *circle nukta above* occurs, for example, in the phrase وَإِذَا (وَ) / વએઝ̊ા, highlighted in blue in figures 1 and 2.

3.6 GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE

The ◌̎ GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE is used for representing the following Arabic letter:

Arabic	Gujarati	Encoded sequence
ظ ZAH	ઝ̎	<ઝ JHA, ◌̎ *two-circle nukta above*>

The *two-circle nukta above* occurs, for example, in the word عَظِيمٌ / અ̇ઝ̎ીમ, highlighted in light green in figures 1 and 2.

4 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.

Ordering When *sukun* occur together with one of the three above-base *nukta* signs or with *shadda*, the latter occurs first in the encoded representation, but it is positioned to the right of the *sukun* in the output:

અ્̇ <અ A, ◌̇ *three-dot nukta above*, ◌્ *sukun*>

ઝ્̎ <ઝ JHA, ◌̎ *two-circle nukta above*, ◌્ *sukun*>

સ્ <સ SA, ◌ *shadda*, ◌્ *sukun*>

The same principle applies to the co-occurrence of *shadda* and one of the three above-base *nukta* signs:

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

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

Glyph Adjustment 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 — then it is reduced in size and rotated in order to prevent clashing and to accommodate fit. 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:

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

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

5 Characters Not Proposed

The sources show a mark resembling ❋, which is used as a sectioning mark. This mark could be a glyphic variant of ⋆ U+066D ARABIC FIVE POINTED STAR. It corresponds to the ۝ U+06DD ARABIC END OF AYAH used in the parallel Arabic. The ❋ occurs in three different contexts: 1) independently, 2) followed by one or more digits (❋ ૩૭), and 3) with a superscript consonant letter. The source shows two superscript letters, eg. અ A and મ MA, written above the mark: ❋ and ❋.

The ❋ section mark is not proposed for encoding as a separate character at present. The ⋆ U+066D ARABIC FIVE POINTED STAR may be used, with appropriate changes to the glyph. The superscript letters may be represented as rich text. However, section mark may be proposed for encoding in the future if additional research shows it to be a distinct character used in Gujarati transliteration of Arabic.

6 Character Data

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

```
0AFx;GUJARATI SIGN SUKUN;Mn;0;NSM;;;;;N;;;;;
0AFx;GUJARATI SIGN SHADDA;Mn;0;NSM;;;;;N;;;;;
0AFx;GUJARATI SIGN MADDAH;Mn;0;NSM;;;;;N;;;;;
0AFx;GUJARATI SIGN THREE-DOT NUKTA ABOVE;Mn;7;NSM;;;;;N;;;;;
0AFx;GUJARATI SIGN CIRCLE NUKTA ABOVE;Mn;7;NSM;;;;;N;;;;;
0AFx;GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE;Mn;7;NSM;;;;;N;;;;;
```

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

```
0AFx;CM # GUJARATI SIGN SUKUN
0AFx;CM # GUJARATI SIGN SHADDA
0AFx;CM # GUJARATI SIGN MADDAH
0AFx;CM # GUJARATI SIGN THREE-DOT NUKTA ABOVE
0AFx;CM # GUJARATI SIGN CIRCLE NUKTA ABOVE
0AFx;CM # GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE
```

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

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

# Indic_Syllabic_Category=Gemination_Mark
0AFx          ; Gemination_Mark   # Mn       GUJARATI SIGN SHADDA
```

Matra Categories Matra categories given in the data format of `IndicMatraCategory.txt`:

```
# Indic_Matra_Category=Top
0AFx..0AFx    ; Top               # Mn   [6] SIGN SUKUN .. SIGN TWO-CIRCLE NUKTA ABOVE
```

7 References

Ismā'il, Gulāmalī. 1901–1903. *Anvārūl bayān phī taphsīril kura'ān*. 3 vols. Bhāvnagar: Isnā'aśarī ilēkṭrīk prīnṭīng prēs.

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

8 Acknowledgments

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

This project was made possible in part with support from the Script Encoding Initiative at the University of California, Berkeley.

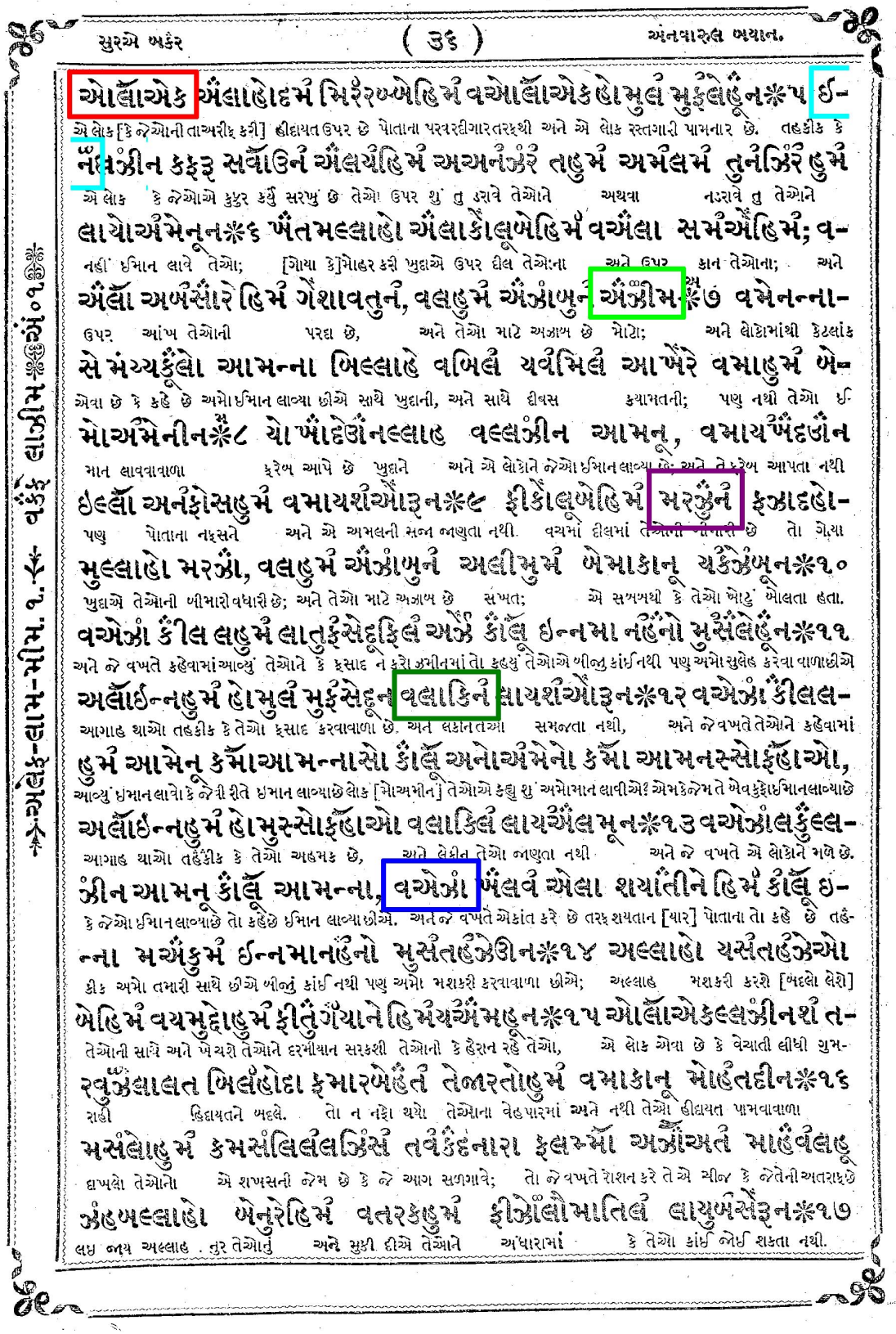

Figure 1: Page of the *Qur'ān* in Gujarati showing usage of the Arabic transliteration marks (from Gulāmalī Ismā'il 1901: 36). Arabic parallel shown in figure 2.

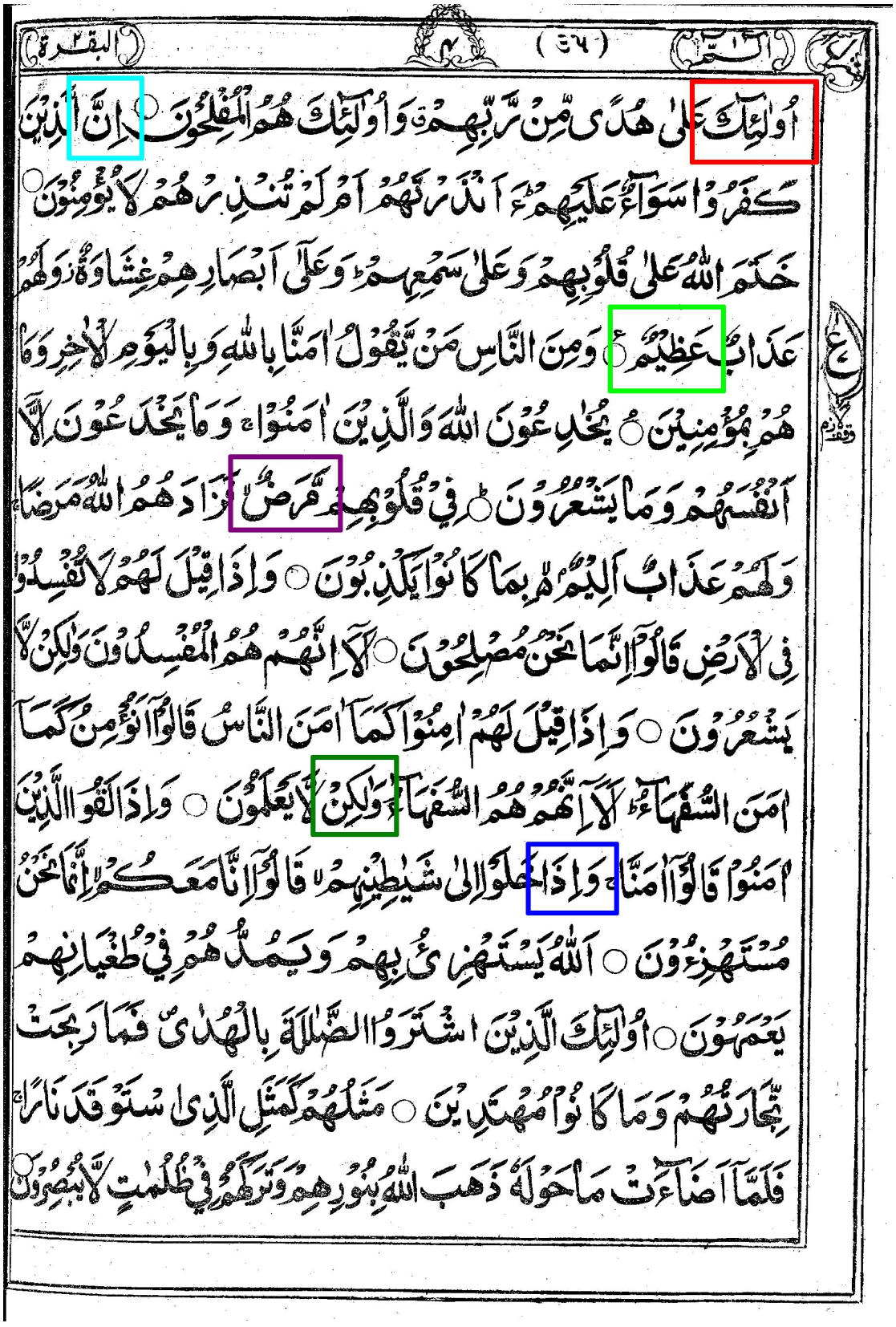

Figure 2: Page of the *Qurʾān* in Arabic (from Gulāmalī Ismāʾil 1901: 35). Gujarati parallel shown in figure 2.