

# Proposal for Thaana Script Reference Label Generation Rules for Root Zone

---

*Date:* 2025-05-23

*Document version:* 1.0

*Authors:* Thaana Script Generation Panel (Thaana GP)

## 1. Overview

This document specifies a set of Label Generation Rules (LGR) for the Thaana script, covering the Dhivehi language. It includes information on the features of the script, and the rationale behind the decisions for the proposed rules.

The formal specification of the Thaana script LGR can be found in the accompanying XML document:

proposal-thaana-lgr-23may25-en.xml

A non-normative HTML version of this file can be found at:

proposal-thaana-lgr-23may25-en.html

Labels for testing can be found in the accompanying text document:

thaana-test-labels-23may25-en.txt

## 2. Script for which the LGR is proposed

ISO 15924 Code: Thaa

ISO 15924 Key N°: 170

ISO 15924 English Name: Thaana

Latin transliteration of native script name: Thaana

Native name of the script: ތާނަ

Maximal Starting Repertoire [MSR] version: 5

The Unicode Standard, Version: 11.0

Unicode Range: 0780–07BF

### 3. Background

The Maldives is a South Asian group of 22 atolls across the Indian Ocean. As per the official census, the population of the Maldives is 515,122 (382,751 Maldivian and 132,371 expatriates), (National Bureau of Statistics, 2023).

Maldivian is also known as Dhivehi, Divehli, Mali, Malikh or Malki. Native names are **ދިވެހި** (Divehi), and **ދިވެހިބަސް** (Divehi-bas). [402]

Dhivehi, the official language of the Maldives has its roots in Sanskrit and belongs to the Indo-Aryan family. According to Naseema Mohamed, the earliest form of the script used to write Dhivehi was Eveylla Akuru, which had strong similarities to South Asian scripts such as Grantha, Elu and Vatteluttu. During the 12th & 13th century, Eveylla Akuru was used in copper-plate grants (Loamaafaanu). Over the centuries the script evolved into Dives Akuru and, Thaana script was introduced during the 18th century. Unlike the earliest script, Thaana is written from right-to-left. [401]

A dialect of Dhivehi (“Maliku Bas” or “Mahl”) is spoken in Minicoy, an island in Lakshadweep, India. The dialect is known to be written in the Thaana script, or a version of Devanagari script. Although there were early attempts at a modified Thaana script for the Mahl dialect called “Maluku Thaana”, this has been abandoned and reverted back to standard Thaana, utilizing the same character set and unicode ranges.

“Major dialects of Maldivian are Malé, Huvadhu, Mulaku, Addu, Haddhunmathee and Maliku. The Malé dialect of the Maldivian capital is considered the standard. In Minicoy the Maliku dialect is spoken and is known as Mahl or Maliku bas.” [402]

This Root zone LGR aims to cover the usage across all dialects. Thaana use across these dialects is consistent, and no one dialect has a different set of characters or dialect-specific rules for writing in Thaana.

The example of Thaana script usages can be found in the newspaper, media and others. [403] [404] [405] [406] [407]

#### 3.1 Thaana Character Categorization

Thaana has 25 consonants and 11 vowels. The first nine letters of Thaana are derived from the Arabic/Persian numerals (see below). The vowels are written above or below the letters.

### 3.1.1 Basic consonants

The following table illustrates the basic consonants and their phonology. There are 25 everyday consonants for native Dhivehi. [401]

Table 1: Thaana script basic consonants

Code point	Glyph	Name	Transliteration	Approximate Pronunciation
U+0780	ހ	THAANA LETTER HAA	h	<i>hat</i>
U+0781	ށ	THAANA LETTER SHAVIYANI	ś	<i>shy</i> (palletized)
U+0782	ނ	THAANA LETTER NOONU	n	<i>net</i>
U+0783	ރ	THAANA LETTER RAA	r	<i>ram</i>
U+0784	ބ	THAANA LETTER BAA	b	<i>baby</i>
U+0785	ޅ	THAANA LETTER LHAVIYANI	ɭ	A retroflexed l close to settle
U+0786	ކ	THAANA LETTER KAAFU	k	<i>kin</i>
U+0787	އ	THAANA LETTER ALIFU	NULL	NULL
U+0788	ވ	THAANA LETTER VAAVU	v	<i>very</i>
U+0789	މ	THAANA LETTER MEEMU	m	<i>men</i>
U+078A	ފ	THAANA LETTER FAAFU	f	<i>farm</i>
U+078B	ދ	THAANA LETTER DHAALU	d	<i>then</i>
U+078C	ތ	THAANA LETTER THAA	t	<i>three</i>
U+078D	ލ	THAANA LETTER LAAMU	l	<i>let</i>
U+078E	ގ	THAANA LETTER GAAFU	g	<i>goal</i>
U+078F	ސ	THAANA LETTER GNAVIYANI	ñ	<i>canyon</i>
U+0790	ސ	THAANA LETTER SEENU	s	<i>see</i>

Code point	Glyph	Name	Transliteration	Approximate Pronunciation
U+0791	ⵎ	THAANA LETTER DAVIYANI	d	<i>do</i>
U+0792	ⵏ	THAANA LETTER ZAVIYANI	z	<i>zero</i>
U+0793	ⵑ	THAANA LETTER TAVIYANI	t	<i>tap</i>
U+0794	ⵒ	THAANA LETTER YAA	y	<i>yet</i>
U+0795	ⵓ	THAANA LETTER PAVIYANI	p	<i>pit</i>
U+0796	ⵔ	THAANA LETTER JAVIYANI	j	<i>jam</i>
U+0797	ⵕ	THAANA LETTER CHAVIYANI	c	<i>chin</i>
U+07B1	ⵙ	THAANA LETTER NAA	dn	<i>Gaṇeśa</i>

The consonant ⵒ (THAANA LETTER ALIFU, U+0787) is typically a null consonant which has no direct pronunciation. It serves one of three functions:

- As a vowel carrier for diphthongs (along with a preceding consonant-vowel pair).
- As a glottal-stop when carrying the vowel SUKUN U+07B0 at the end of a sentence.
- As a gemination of the following consonant when used in the middle of a word, when combined with SUKUN U+07B0.

The consonant ⵙ (THAANA LETTER NAA, U+07B1) is a dialect specific consonant. Sometimes called “baru noonu” (heavy n) or “Dnaviyani”, represents the retroflex nasal and was dropped from standard Thaana in favor of ⵕ (THAANA LETTER NOONU, U+0782). However, the consonant has seen a resurgence in usage in online writing and social media, and a decision was made to include it in the repertoire at this time, despite it not being officially recognized as part of the base consonants. This was done given that it is culturally relevant, particularly to the southern dialects.

### **THAANA LETTER NOONU (U+0782) and THAANA LETTER RAA (U+0783)**

These characters are unique in that they can be used without vowels in specific circumstances. THAANA LETTER NOONU (U+0782) when used by itself without a vowel, can be used to pre-nasalize the next consonant. THAANA LETTER RAA (U+0783) when used by itself without a vowel, is used to the “r” sound in borrowed words such as those that appear on words ending with -er, -or and so on (ex: Doctor). It is valid, and often practised however to add a vowel to Raa such that the “-er” becomes “-ru”, or to completely replace it with “-aa”. Each of them can be at the end of a label, or can be followed by a

consonant. They can also follow themselves, but a vowel is expected after the second occurrence. They cannot start a label under any circumstances.

### 3.1.2 Vowels

The following table illustrates 11 Thaana vowels and their usage. [401]

Table 2 - Thaana script vowels

Code point	Glyph	Name	Transliteration	Approximate Pronunciation
U+07A6	◌	THAANA ABAFILI	a	sun
U+07A7	◌	THAANA AABAAFILI	ā	father
U+07A8	◌	THAANA IBIFILI	i	mill
U+07A9	◌	THAANA EEBEEFILI	ī	meet
U+07AA	◌	THAANA UBUFILI	u	full
U+07AB	◌	THAANA OOBOOFILI	ū	rule
U+07AC	◌	THAANA EBEFILI	e	bet
U+07AD	◌	THAANA EYBEYFILI	ē	bale
U+07AE	◌	THAANA OBOFILI	o	going
U+07AF	◌	THAANA OABOAFILI	ō	phone
U+07B0	◌	THAANA SUKUN	NULL	NULL

Sukun in itself has no pronunciation. It can indicate a null vowel sound, glottal stop, or gemination of nasal consonants, depending on where it occurs.

### 3.1.3 Extended consonants for Arabic and other loan words

These code points are excluded from the Thaana script RZ-LGR (see 5.2).

Since Dhivehi has a lot of Arabic loan words, an additional set of letters were created in 1957, to transliterate the loan words accurately, by adding diacritics (dots) to the existing letters. An exception to this is 𑌛 U+079C, which was later added to represent the sound /ʒ/ in English words such as treasury and vision. [401]

Table 3 - Thaana script extended consonants for Arabic

Code point	Glyph	Name	IPA	Arabic Transliteration
U+0798	ﺘ	THAANA LETTER TTAA	θ	ث
U+0799	ﻫ	THAANA LETTER HHAA	ħ	ح
U+079A	ﺦ	THAANA LETTER KHAA	x	خ
U+079B	ﺬ	THAANA LETTER THAALU	ð	ذ
U+079C	ﺰ	THAANA LETTER ZAA	ʒ	N/A
U+079D	ﺶ	THAANA LETTER SHEENU	ʃ	ش
U+079E	ﺺ	THAANA LETTER SAADHU	s <sup>ʕ</sup>	ص
U+079F	ﺪ	THAANA LETTER DAADHU	d <sup>ʕ</sup>	ط
U+07A0	ﺖ	THAANA LETTER TO	t <sup>ʕ</sup>	ض
U+07A1	ﺰ	THAANA LETTER ZO	ð <sup>ʕ</sup>	ظ
U+07A2	ﺀ	THAANA LETTER AINU	ʕ	ع
U+07A3	ﻏ	THAANA LETTER GHAINU	ɣ	غ
U+07A4	ﻕ	THAANA LETTER QAAFU	q	ق
U+07A5	ﻭ	THAANA LETTER WAAVU	w	و

## 4. Overall Development Process and Methodology

Thaana LGR was initiated at an ICANN outreach seminar held in Male' in 2018. However, the first introductory meeting was held online in January 2024 to develop the Thaana script Reference LGR for the second level (Thaana script Ref. LGR). The Thaana script Ref. LGR, finalized after the public comment, was [published](#) in November 2024.

In September 2024, while the Thaana script Ref. LGR was being finalized, the Thaana Script Root Zone LGR Generation Panel (Thaana GP) was [formed](#). The Thaana GP consists of five

members. They are Unicode standard and software engineer, Linguist, Academia, Font Type designer, and representative from government and public work.

The GP reviews the work done in the Thaana script Ref.LGR and uses it as an input to develop the Thaana script RZ-LGR.

The GP’s working methodology follows the Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels ([RZ-LGR Procedure](#)). The first GP meeting was held in September 2024, fortnightly online meetings are held to review updates and the feedback is incorporated after every meeting.

For various statistical analysis, the Thaana GP compiled their own corpus of raw text data obtained from online news websites, dictionaries, and other publications. The resulting corpus contains 2,109,844 non-unique Thaana strings, which also captures both correct and incorrect usage of the script in the general public domain. Here, incorrect refers to typing mistakes, grammatical errors and such, that a general script user may produce in everyday usage as well. The GP considered this characteristic to be useful to validate potential implications of rules developed for the Root Zone LGR.

The GP conducted two public consultation sessions: One online session on 21st January, 2025 and one physical session on 30th January 2025 in Maldives. Both sessions were attended by representatives in the tech and design community, Government, linguists, media persons and internet service providers.

## 5. Repertoire

Based on RZ-LGR Procedure, code point analysis of the Thaana script was conducted from the code points included by the [MSR] to be validated and used in the RZ-LGR. This section includes a list of code points recommended for inclusion and exclusion from the repertoire.

### 5.1 Included Code Points

The repertoire includes 36 code points of the basic alphabet and diacritics (vowels) used in Thaana script in everyday general use. Note that the hyphen and digits are not allowed in the root zone.

Table 4 – Thaana script RZ-LGR included code points

Code Point	Glyph	Name	Category	Ref
U+0780	ހ	THAANA LETTER HAA	Thaana Consonant	[401] [402]
U+0781	ށ	THAANA LETTER SHAVIYANI	Thaana Consonant	[401] [402]
U+0782	ނ	THAANA LETTER NOONU	Thaana Consonant	[401] [402]

Code Point	Glyph	Name	Category	Ref
U+0783	ر	THAANA LETTER RAA	Thaana Consonant	[401] [402]
U+0784	ب	THAANA LETTER BAA	Thaana Consonant	[401] [402]
U+0785	ل	THAANA LETTER LHAVIYANI	Thaana Consonant	[401] [402]
U+0786	ك	THAANA LETTER KAAFU	Thaana Consonant	[401] [402]
U+0787	ا	THAANA LETTER ALIFU	Thaana Consonant	[401] [402]
U+0788	و	THAANA LETTER VAAVU	Thaana Consonant	[401] [402]
U+0789	م	THAANA LETTER MEEMU	Thaana Consonant	[401] [402]
U+078A	ف	THAANA LETTER FAAFU	Thaana Consonant	[401] [402]
U+078B	د	THAANA LETTER DHAALU	Thaana Consonant	[401] [402]
U+078C	ت	THAANA LETTER THAA	Thaana Consonant	[401] [402]
U+078D	ن	THAANA LETTER LAAMU	Thaana Consonant	[401] [402]
U+078E	ي	THAANA LETTER GAAFU	Thaana Consonant	[401] [402]
U+078F	ح	THAANA LETTER GNAVIYANI	Thaana Consonant	[401] [402]
U+0790	س	THAANA LETTER SEENU	Thaana Consonant	[401] [402]
U+0791	ع	THAANA LETTER DAVIYANI	Thaana Consonant	[401] [402]
U+0792	ج	THAANA LETTER ZAVIYANI	Thaana Consonant	[401] [402]
U+0793	ح	THAANA LETTER TAVIYANI	Thaana Consonant	[401] [402]
U+0794	ر	THAANA LETTER YAA	Thaana Consonant	[401] [402]
U+0795	ق	THAANA LETTER PAVIYANI	Thaana Consonant	[401] [402]
U+0796	ع	THAANA LETTER JAVIYANI	Thaana Consonant	[401] [402]
U+0797	چ	THAANA LETTER CHAVIYANI	Thaana Consonant	[401] [402]
U+07A6	ا	THAANA ABAFILI	Thaana Vowel	[401] [402]
U+07A7	ا	THAANA AABAAFILI	Thaana Vowel	[401] [402]
U+07A8	ا	THAANA IBIFILI	Thaana Vowel	[401] [402]
U+07A9	ا	THAANA EEBEEFILI	Thaana Vowel	[401] [402]
U+07AA	ا	THAANA UBUFILI	Thaana Vowel	[401] [402]
U+07AB	ا	THAANA OOBOOFILI	Thaana Vowel	[401] [402]
U+07AC	ا	THAANA EBEFILI	Thaana Vowel	[401] [402]
U+07AD	ا	THAANA EYBEYFILI	Thaana Vowel	[401] [402]
U+07AE	ا	THAANA OBOFILI	Thaana Vowel	[401] [402]



Given the relatively low frequency of usage of Arabic consonants, the ability to swap them out with base consonants as well as a tendency to be used incorrectly, it was decided that the conservative approach would be to exclude the extended consonants from the Root Zone Repertoire.

Table 5 – Thaana script RZ-LGR excluded code points

#	Code Point	Glyph	Name
1	U+0798	ﺗﺘﺎ	THAANA LETTER TTAA
2	U+0799	ﻫﻫﺎ	THAANA LETTER HHAA
3	U+079A	ﻛﻫﺎ	THAANA LETTER KHAA
4	U+079B	ﺗﻫﺎ	THAANA LETTER THAALU
5	U+079C	ﺯﺎ	THAANA LETTER ZAA
6	U+079D	ﺷﻪﻧﯘ	THAANA LETTER SHEENU
7	U+079E	ﺳﺎﺩﻫﯘ	THAANA LETTER SAADHU
8	U+079F	ﺩﺎﺩﻫﯘ	THAANA LETTER DAADHU
9	U+07A0	ﺗﻮ	THAANA LETTER TO
10	U+07A1	ﺯﻭ	THAANA LETTER ZO
11	U+07A2	ﺍﻳﻨﯘ	THAANA LETTER AINU
12	U+07A3	ﻏﻬﺎﻳﻨﯘ	THAANA LETTER GHAINU
13	U+07A4	ﻗﺎﺍﻓﯘ	THAANA LETTER QAAFU
14	U+07A5	ﯞﺎﺍﯞﯗ	THAANA LETTER WAAVU

## 6. Variants

### 6.1 In-Script Variants

If the extended consonants for Arabic and other loan words were included in the repertoire, the consideration of variants between the extended consonants and the basic consonants would need to be included. However, the extended consonants for Arabic are excluded from the Thaana script RZ-LGR. Therefore, the variants for core consonants are not required.

There is only one in-script variant set with two members as shown in the table below. The disposition is blocked.

Table 6 – Thaana script in-script variant

Set	Code Point	Glyph	Name	Code Point	Glyph	Name
1	U+0782	س	THAANA LETTER NOONU	U+07B1	ز	THAANA LETTER NAA

These consonants are considered variants because they are phonetically similar (u+0782 NOONU voiced alveolar nasal, and u+07B1 NAA voiced retroflex nasal) and are used interchangeably.

Example: Dhondheena (name of a bird) can be interchangeably written in these two forms:

ދަހަންދީނާ : U+078B U+07AE U+0782 U+07B0 U+078B U+07A9 **U+07B1** U+07A7

ދަހަންދީނާ : U+078B U+07AE U+0782 U+07B0 U+078B U+07A9 **U+0782** U+07A7

## 6.2 Cross-Script Variants

### 6.2.1 Analysis based on the background of the Thaana script

As previously mentioned, the first nine glyphs of the script are derived from the glyph forms for Arabic numerals. The following table illustrates the derivation.

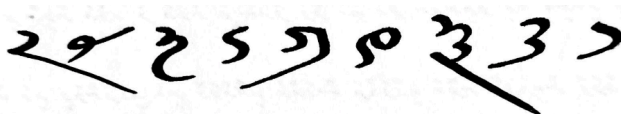

Modern Thaana	Old Thaana	Persian	Arabic
١	۱	۱	۱
٢	۲	۲	۲
٣	۳	۳	۳
٤	۴	۴	۴
٥	۵	۵	۵
٦	۶	۶	۶
٧	۷	۷	۷
٨	۸	۸	۸
٩	۹	۹	۹

Fig. 1: How the first nine letters were derived [408]

For completeness, given the similarity in glyph forms, adding variants for the first nine consonants was considered for variant candidates. However, given that modern Thaana is written with a slant and is not easily confusable with the Persian/Arabic numerals it was decided not to include them as cross-script variants at this time.

Furthermore, the next nine consonants in Thaana were derived from the old Dhivehi alphabet forms script, called Dives Akuru (Unicode Dives Akuru U+11900-U+1195F).

The graphic below shows the Dives akuru digits 1-9, and Thaana consonants derived from them [408].

ASCII Digits	9 8 7 6 5 4 3 2 1
Dives Akuru script	
Thaana script	

However, given that Dives akuru is not present in the reference LGR, and the fact that it is rarely used in either writing or in fonts, these letters were not considered part of the repertoire. Therefore, not in the scope of variant consideration.

### 6.2.2 Thaana script and Arabic script analysis

The following pairs in Table 7 were analyzed between Thaana and Arabic.

Table 7 – Thaana - Arabic cross-script analysis

Thaana Code Point	Glyph	Name	Arabic Code Point	Glyph	Name
U+0780	↵	THAANA LETTER HAA	U+0631	ر	ARABIC LETTER REH
U+0784	⦿	THAANA LETTER BAA	U+06BE	ه	ARABIC LETTER HEH DOACHASHMEE
U+0788	↶	THAANA LETTER VAAVU	U+0648	و	ARABIC LETTER WOW
U+0789	↷	THAANA LETTER MEEMU	U+062F	د	ARABIC LETTER DAL

The Thaana RZ-LGR only allows a consonant to be followed by a vowel, and the Arabic RZ-LGR does not include vowels, therefore, it is considered as similarity cased but not raised to the variant level.

Invalid Thaana Label: ڤوڤو

Valid Thaana Label: ڤوڤو

Valid Arabic Label: هودر

Therefore, there is no cross-script variant between Thaana-Arabic defined.

### 6.2.3 Thaana script and other scripts

Beside Arabic script, the GP conducts variant analysis with all other 25 other scripts integrated in the [RZ-LGR version 5](#). As the Thaana RZ-LGR only allows a consonant to be followed by a vowel which appears above or below the consonant, therefore it is unlikely to create a whole label confusion between a Thaana label and the single-level scripts labels. These single-level scripts include Armenian, Chinese (Han), Ethiopic, Georgian, Greek, Hebrew, Japanese (Hiragana, Katakana), Korean (Han, Hangul), and Latin scripts. Though some of these scripts present diacritics above or below a letter, there is no base letter that looks confusingly similar with a Thaana letter.

The scripts with the multi-level feature have been analyzed which include Bangla, Devanagari, Gujarati, Gurmukhi, Kannada, Khmer, Lao, Malayalam, Myanmar, Oriya, Sinhala, Tamil, Telugu, and Thai scripts.

For base consonants, some cross script similarities do exist such as U+07B1 ( ڤ ,Thaana Letter Naa) with U+057B ( զ ,Armenian Small Letter Jheh), U+0786 ( ڤ ,Thaana Letter Kaafu) with U+0076 ( v , Latin Small Letter V), or U+078E ( ڤ ,Thaana Letter Gaafu) potentially being similar to “s”. However, after analysing these in the context of the whole label rules, which disallow Thaana base consonants without vowels, which negates the need to encode the same as cross script variants, the GP concluded that at this time, no out of script variants are required as far as base consonants are concerned.

Some of the multi-level featured script includes vowels which look similar to Thaana vowels. For example, U+07A8 ( ڤ , THAANA IBIFILI) and U+0956 ( ॐ ,DEVANAGARI VOWEL SIGN UE), or U+07B0 ( ڤ ,THAANA SUKUN ) and U+0B02 ( ॰ ,Oriya ORIYA SIGN ANUSVARA). However, there are no base characters which can create the whole label confusion between Thaana script and these scripts.

In summary, although there are some similar glyphs, within the scope of the root zone, the GP is of the opinion that none of them have cause for concern within the context of the rules defined for Thaana, and therefore no cross-script variant is defined for Thaana script RZ-LGR.

## 7. Whole Label Evaluation Rules and Contextual Rules

Thaana is an abugida script, and is usually written in CV form, i.e. consonant-vowel pairs.

The general exception to this is the consonants Noonu (U+0782) and Raa (U+0783). These two consonants can either appear with or without a vowel, depending on the context. These two consonants were defined as a more flexible class N in Thaana Ref. LGR for the second-level to allow them to appear without a vowel.

However, as seen in the frequency distribution chart in Appendix 1 Fig. 3 for word-length, even number character sequences are more common than odd number sequences, indicating that the usage of Noonu and Raa without vowels is relatively rare. It should be noted that the graph is from a large corpus, including many misspellings and errors i.e. accidentally dropped vowels, repeated vowels etc which also contributes to odd-number lengths.

An additional graphs, Fig. 4 and Fig.5, illustrate that consecutive vowels (an error) are more common (117,840) than occurrences of U+782 without a vowel (96,033) or U+783 without a vowel (28,487). This is compared against 13.45 million occurrences of CV (consonant with vowel).

As noted previously in 3.1, usages of Raa (U+0783) without a vowel can and often are replaced with a variation carrying a vowel (often ر U+0783 U+07AA) such that the “-er” sound becomes “-ru”.

For examples:

Transliteration of ‘doctor’ can be written in two forms interchangeably.

ڊڪٽر U+0791 U+07AE U+0786 U+07B0 U+0793 U+07A6 U+0783

ڊڪٽر U+0791 U+07AE U+0786 U+07B0 U+0793 U+07A6 U+0783 U+07AA

Transliteration ‘waiter’ can be written in two forms interchangeably.

ڦوٽر U+0788 U+07AC U+0787 U+07A8 U+0793 U+07A6 U+0783

ڦوٽر U+0788 U+07AC U+0787 U+07A8 U+0793 U+07A6 U+0783 U+07AA

For Noonu (U+0782), usage without a vowel indicates a pre-nasalization of the next consonant. Apart from a few exceptions where dropping it changes the meaning of the word itself, it is generally acceptable to do so.

For example:

Transliteration ‘mango’ can be written in two forms interchangeably.

مانڱو U+0787 U+07A6 U+0782 U+0784 U+07AA

مانڱو U+0787 U+07A6 U+0784 U+07AA

With these considerations, the GP decided to take on the stricter and more conservative approach for the root zone in restricting the Noonu (U+0782) and Raa (U+0783) to behave strictly the same way as other consonants, i.e. always written with a vowel. Additional constraints related to vowel usage to consider where:

- a vowel will never begin a label, and
- a vowel can never follow another vowel.

In Thaana, these two instances will never occur. It also be noted that repeated vowels, i.e. a vowel followed by a vowel, is a common typing error which results in invisible characters as the vowels will usually overlap one another.

This proposal defines the following named character classes:

- Consonants (C): U+0780 - U+0797, U+07B1
- Vowels(V): U+07A6 - U+07B0

Thaana Script LGR includes the following contextual rules:

1. V only follows a C
2. C must be followed by a V

These two rules, in combination, ensure that all labels follow the required CV structure, while also ensuring a label never starts with a V, always ends with a V, and the case of VV never occurs.

The formal specification of the Thaana script LGR can be found in the accompanying XML document:

proposal-thaana-lgr-23may25-en.xml

#### ***Addition Note:***

During the public consultation, it was raised there may be a need of a sequence of ދިރި (Dh-R, U+078B U+0783) because this could be a candidate of an IDN ccTLD. ދިރި stands as an abbreviation for “ދިވެހިރާއްޖެ” (U+078B U+07A8 U+0788 U+07AC U+0780 U+07A8 U+0783 U+07A7 U+0787 U+07B0 U+0796 U+07AC), which is the common local name for “Maldives”.

The GP has consulted with the Integration Panel and the LGR procedure and found that encoding one specific label in the RZ-LGR is against the predictability principle. In addition, the sequence has not been formally confirmed as a selected IDN ccTLD by the community. The GP decided not to allow the sequence in this version. Note that the process for [IDN ccTLD Fast Track](#) was taken into consideration.

## 8. Contributors

- Naail Abdul Rahman ([kudanai@gmail.com](mailto:kudanai@gmail.com))
- Aminath Saeed ([aminath.s@mnu.edu.mv](mailto:aminath.s@mnu.edu.mv))
- Ali Fawaz Shareef ([a.f.shareef@mnu.edu.mv](mailto:a.f.shareef@mnu.edu.mv))
- Abdulla Shafeeu(Abo) ([abo@encrea.com](mailto:abo@encrea.com))
- Hassan Irfan ([irfantgo@gmail.com](mailto:irfantgo@gmail.com))
- Women In Tech Maldives ([info@womenintechmv.org](mailto:info@womenintechmv.org))

## 9. References

- [MSR] Integration Panel, "Maximal Starting Repertoire - MSR-5 Overview and Rationale", <https://www.icann.org/en/system/files/files/msr-5-overview-24jun21-en.pdf> (Accessed on 10 September 2024)
- [401] Dhivehi Writing Systems by Naseema Mohamed, page 7-8, NCLHR, 1999. ISBN 99915-71-91-4
- [402] Maldivian (ދިވެހި), Omniglot, <https://www.omniglot.com/writing/thaana.htm> (Accessed on 5 February 2025)
- [403] That Maldivian Blog, Raajje\_blog, <https://thatmaldivesblog.wordpress.com/learn-dhivehi/> (Accessed on 1 March 2025)
- [404] Dhivehi Language, Hassan Hameed, <https://www.hassanhameed.com/dhivehi-language/> (Accessed on 1 March 2025)
- [405] Dhivehi newspaper, Mihaaru, <https://mihaaru.com/> (Accessed on 1 March 2025)
- [406] Dhivehi media company, SUN MEDIA GROUP, <http://sun.mv/> (Accessed on 1 March 2025)
- [407] Media, News company, Raajje, <http://raajje.mv/> (Accessed on 1 March 2025)
- [408] Thaana: History, Development and Change, Hassan Hameed, <https://www.hassanhameed.com/dhivehi-language/thaana-history-development-and-change/> (Accessed on 1 March 2025)



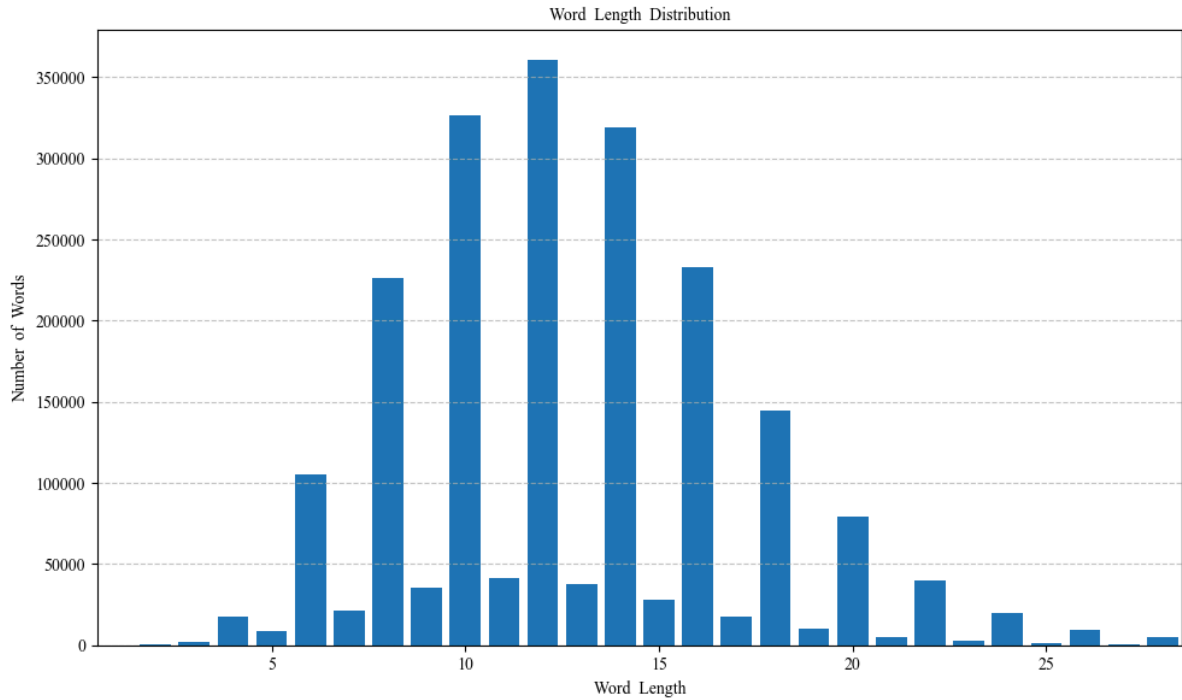


Fig. 4: Distribution of word length across the corpus. As Thaana is mostly written in CV pairs, it is expected to peak at even-length words. Odd-length words are either words with occurrences of either U+0782, U+0783 without vowels, or from repeated vowels (errors). Generated by Thaana GP.

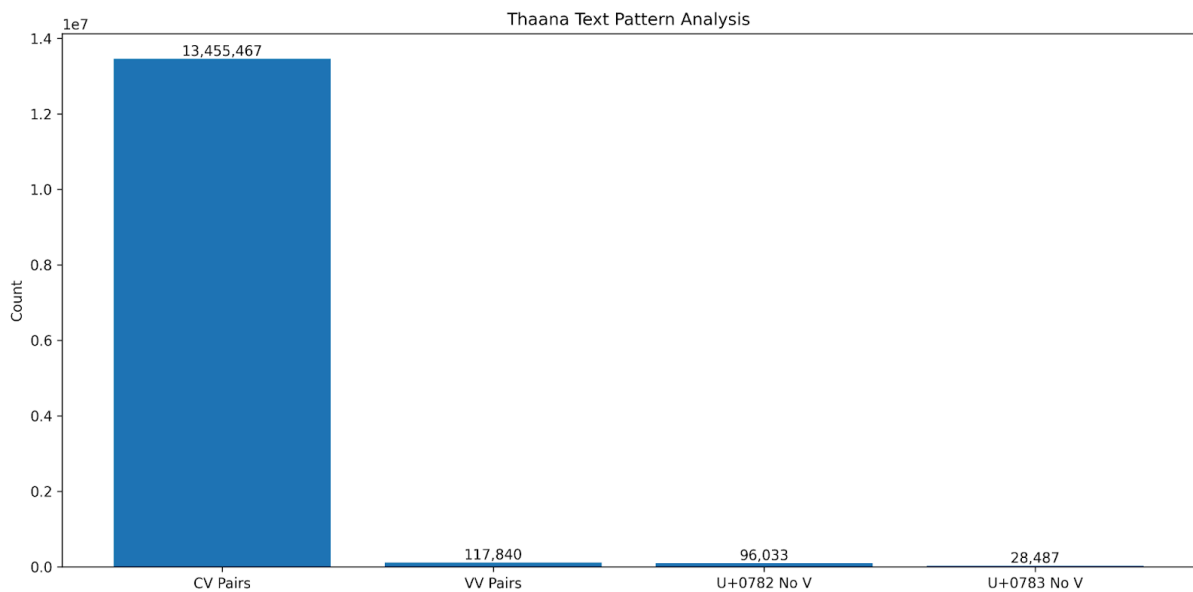


Fig. 5: Frequency of CV pairs, VV pairs (errors) and occurrences of U+0782, U+0783 without vowels to illustrate their relatively infrequent use. Generated by Thaana GP.