

# The Story of the Official Languages of Sri Lanka, Sinhala, and Tamil on the Multilingual Internet and their current status

Harsha Wijayawardhana B.S. in Biochemistry (Miami), CITP (UK), FBCS (UK)  
COO/CTO Theekshana and Board Member, LK Domain Registry and Road Development Authority of Sri Lanka  
Chair, Local Language Working Group and ICANN's Universal Acceptance Steering Group  
[wijayawardhana@gmail.com](mailto:wijayawardhana@gmail.com), විජයවර්ධන@ඉතැපැල්.ලංකා

# Introduction

- Today, the Internet has become ubiquitous and is twelve thousand days or thirty-five years old. It is used by more than sixty percent of the world's population.
- The current Internet is more versatile than the first generation of the Internet, which came into existence in the latter part of the nineteen sixties.
- It also supports almost all worldwide scripts

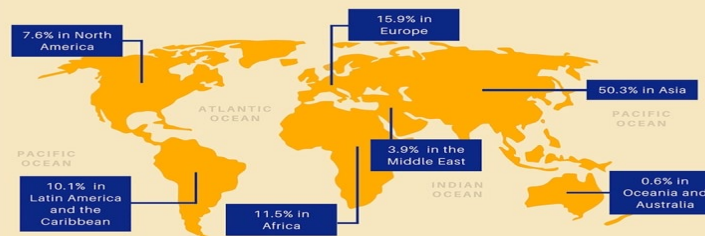
# Introduction cont...

During the last thirty-five years, the Internet has passed through three generations.

Reference:  
<https://websitesetup.org/news/internet-facts-stats/>

## Asia Has the Largest Percentage of Internet

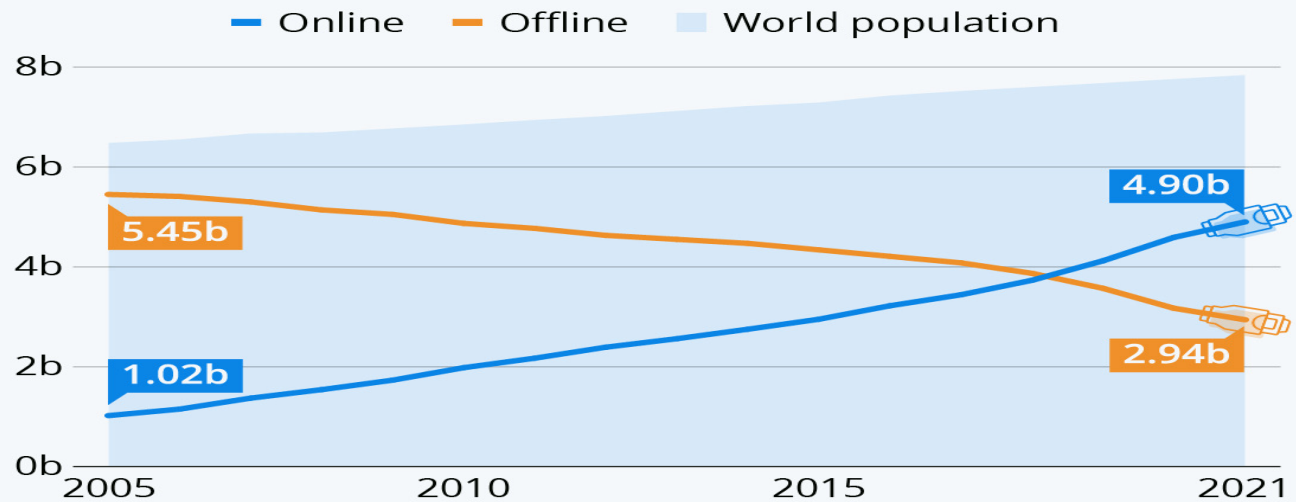
Asia has the largest percentage of Internet users by continent/region.



# Internet User's growth

## Disconnected: 2.9 Billion People Still Offline

Estimated number of individuals worldwide using/not using the internet



Source: ITU




# Multilingual Internet cont

WorldAtlas CONTINENTS COUNTRIES GEOGRAPHY EDUCATION SOCIAL SCIENCE

## The World's Most Popular Writing Scripts

Rank	Name of script	Type	Population actively using (in millions)
1	Latin Latin	Alphabet	over 4900
2	Chinese 汉字 漢字	Logographic	1340
3	Arabic العربية	Abjad	660+
4	Devanagari देवनागरी	Abugida	608+
5	Bengali-Assamese বাংলা-অসমীয়া	Abugida	300
6	Cyrillic Кириллица	Alphabet	250
7	Kana かな カナ	Syllabary	120
8	Javanese	Abugida	80
9	Hangul 한글	featural	70.7
10	Telugu తెలుగు	Abugida	74
11	Tamil தமிழ்	Abugida	70
12	Gujarati ગુજરાતી	Abugida	48
13	Kannada ಕನ್ನಡ	Abugida	45
14	Burmese မြန်မာ	Abugida	39
15	Malayalam മലയാളം	Abugida	38
16	Thai ไทย	Abugida	38
17	Sundanese	Abugida	38
18	Gurmukhi ਗੁਰਮੁਖੀ	Abugida	22
19	Lao ວາອ	Abugida	22
20	Cherokee ᏍᎦᏏᎠ	Abugida	21
21	Ge'ez ግዕዝ	Abugida	18
22	Sinhala සිංහල	Abugida	14.4
23	Hebrew עברית	Abjad	14
24	Armenian Հայերեն	Alphabet	12
25	Khmer ខ្មែរ	Abugida	11.4
26	Greek Ελληνικό	Alphabet	11
27	Lontara	Abugida	7.6
28	Tibetan ལྷནས་	Abugida	5
29	Georgian ქართული	Alphabet	4.5
30	Modern Yi ꯀꯃ	Syllabary	4
31	Mongolian ᠮᠣᠩᠭᠣᠯ	Alphabet	2
32	Tifinagh	Abjad	2

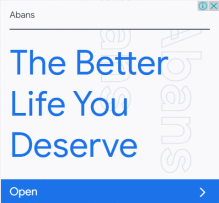


WorldAtlas CONTINENTS COUNTRIES GEOGRAPHY EDUCATION SOCIAL SCIENCE

Windows taskbar: The World..., Sinhala Scri..., Presentatio..., Multilingua..., \*Untitled - ... 11:45 AM 10/31/2022

WorldAtlas CONTINENTS COUNTRIES GEOGRAPHY EDUCATION SOCIAL SCIENCE

21	Ge'ez ግዕዝ	Abugida	18
22	Sinhala සිංහල	Abugida	14.4
23	Hebrew עברית	Abjad	14
24	Armenian Հայերեն	Alphabet	12
25	Khmer ខ្មែរ	Abugida	11.4
26	Greek Ελληνικό	Alphabet	11
27	Lontara	Abugida	7.6
28	Tibetan ལྷནས་	Abugida	5
29	Georgian ქართული	Alphabet	4.5
30	Modern Yi ꯀꯃ	Syllabary	4
31	Mongolian ᠮᠣᠩᠭᠣᠯ	Alphabet	2
32	Tifinagh	Abjad	2



Windows taskbar: The World..., Sinhala Scri..., Presentatio..., Multilingua..., \*Untitled - ... 11:46 AM 10/31/2022

# The World Before the Unicode

- Before the advent of the Unicode, numerous encoding standards that were unique to each language or group of languages had been introduced worldwide.
- The American Standard Code for Information Interchange (ASCII) became the most popular and was also known as ISO 646: it became one of the many standards that came as a group of standards belonging to a 7-bit character encoding.
- American Standard Association, ASA (now known as American National Standard Institute or ANSI) developed ASCII based upon the Telegraph Code in 1964.

# The World Before the Unicode cont.

- ISCII:

Indian Script Code for Information Interchange (ISCII) is a coding scheme for representing various writing systems of India. It encodes the main Indic scripts and a Roman transliteration. The supported scripts are: Bengali-Assamese, Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil, and Telugu.

- Sri Lanka too came up with its own character code by the beginning of nineties called Sri Lanka Standard Sinhala Character Code for Information Interchange.

(SLSCII). This was later known as SLS 1134 version I.

# The emergence of Unicode---Universal Character Encoding

- With the introduction of Universal Character Set, All most all character sets are handled. This came to know worldwide as Unicode or Unicode Consortium
- The current version (15) of Unicode Code contains 149,186 characters and covering 161 modern and historic scripts as symbols, emojis, colors, formatting codes as well.
- Sinhala began its journey being encoded in Unicode version 3 in 1999.
- Later, Sinhala numerals were added as early as 2010. Lith Illaklam into Basic Multilingual Plane and Sinhala illakkam into SMP.



# Emergence cont.

- Success of the Unicode led all operating systems in the world moving to Unicode. Unicode has been used for Internationalization and Localization.
- Unicode standard is implemented in recent Technologies, Modern Software, Operating Systems, XML, etc.
- The Unicode character repertoire is synchronized with ISO/IEC 10646, each being code-for-code identical with the other.

# Unicode cont...

0D80 Sinhala 0DF7

	0D8	0D9	0DA	0DB	0DC	0DD	0DE	0DF
0	ඌ	ඍ	ඎ	ඏ	ඐ	එ		
1	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
2	ඐ	එ	ඒ	උ	ඌ	ඍ	ඎ	ඏ
3	ඐ	එ	ඒ	උ	ඌ	ඍ	ඎ	ඏ
4	ඍ	ඎ	ඏ	ඐ	එ	ඒ	උ	ඌ
5	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
6	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
7	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
8	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
9	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
A	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
B	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
C	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
D	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
E	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
F	ඒ	උ	ඌ	ඍ	ඎ	ඏ		

The latest Version of Sinhala Unicode 15.5

0D80 Sinhala 0DF7

	0D8	0D9	0DA	0DB	0DC	0DD	0DE	0DF
0	ඌ	ඍ	ඎ	ඏ	ඐ	එ		
1	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
2	ඐ	එ	ඒ	උ	ඌ	ඍ	ඎ	ඏ
3	ඐ	එ	ඒ	උ	ඌ	ඍ	ඎ	ඏ
4	ඍ	ඎ	ඏ	ඐ	එ	ඒ	උ	ඌ
5	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
6	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
7	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
8	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
9	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
A	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
B	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
C	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
D	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
E	ඒ	උ	ඌ	ඍ	ඎ	ඏ		
F	ඒ	උ	ඌ	ඍ	ඎ	ඏ		

Sinhala Unicode version 7 with Numerals

111E0 Sinhala Archaic Numbers 111FF

	111E	111F
0	අ	ආ
1	ඇ	ඈ
2	ඈ	ඉ
3	ඉ	ඊ
4	ඊ	උ
5	උ	ඌ
6	ඌ	ඍ
7	ඍ	ඎ
8	ඎ	ඏ
9	ඏ	ඐ
A	ඐ	එ
B	එ	ඒ
C	ඒ	උ
D	උ	ඌ
E	ඌ	ඍ
F	ඍ	ඎ

This number system is also known as Sinhala Illakkam. This number system does not have a zero place holder concept, unlike the Sinhala astrological numbers, Sinhala Lith Illakkam, encoded in the range 0DE6-0DEF.

**Historical digits**

- 111E1 ආ SINHALA ARCHAIC DIGIT ONE
- 111E2 ඇ SINHALA ARCHAIC DIGIT TWO
- 111E3 ඈ SINHALA ARCHAIC DIGIT THREE
- 111E4 ඉ SINHALA ARCHAIC DIGIT FOUR
- 111E5 ඊ SINHALA ARCHAIC DIGIT FIVE
- 111E6 උ SINHALA ARCHAIC DIGIT SIX
- 111E7 ඌ SINHALA ARCHAIC DIGIT SEVEN
- 111E8 ඍ SINHALA ARCHAIC DIGIT EIGHT
- 111E9 ඎ SINHALA ARCHAIC DIGIT NINE

**Historical numbers**

- 111EA ඍ SINHALA ARCHAIC NUMBER TEN
- 111EB ඎ SINHALA ARCHAIC NUMBER TWENTY
- 111EC ඏ SINHALA ARCHAIC NUMBER THIRTY
- 111ED ඐ SINHALA ARCHAIC NUMBER FORTY
- 111EE එ SINHALA ARCHAIC NUMBER FIFTY
- 111EF ඒ SINHALA ARCHAIC NUMBER SIXTY
- 111F0 උ SINHALA ARCHAIC NUMBER SEVENTY
- 111F1 ඌ SINHALA ARCHAIC NUMBER EIGHTY
- 111F2 ඍ SINHALA ARCHAIC NUMBER NINETY
- 111F3 ඎ SINHALA ARCHAIC NUMBER ONE HUNDRED
- 111F4 ඏ SINHALA ARCHAIC NUMBER ONE THOUSAND

Sinhala Illakkam in version 7

# Unicode cont ...

Sinhala code table

	130	131	132	133	134	135	136	137
0		ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
1			ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
2	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
3	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
4		ඵ	ඵ	ඵ	ඵ			ඵ
5	ඵ	ඵ	ඵ	ඵ				ඵ
6	ඵ	ඵ	ඵ	ඵ	ඵ			ඵ
7	ඵ	ඵ	ඵ	ඵ	ඵ		ඵ	ඵ
8	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
9	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
A	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
B	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
C	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
D		ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
E	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ
F	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ	ඵ

Michael Everson's Proposal for Sinhala in 1997

# Unicode Standard cont...

- Unicode Consortium Website covers not only character sets, but other detailed areas such as normalization of characters, collation and Bidirectional Text.
- The Unicode standard defines three and several other encodings exist, all in practice variable-width encodings. The most common encodings are the ASCII-compatible UTF-8, the ASCII-incompatible UTF-16 (compatible with the obsolete UCS-2), and the Chinese Unicode encoding standard GB18030 which is not an official Unicode standard but is used in China and implements Unicode fully.

# Rakaaraansaya and Yansaya issue

- 𐌆𐌿𐌸 : 0db1/0dca/0daf
  - Letter NA+Hal Kirima+Letter Da
- 𐌆𐌿𐌸𐌵 : 0db1/0dca/200c/0daf
  - Letter Na+Hal Kirima+ ZWJ+Letter Da

# Rakaaraansaya and Yansaya cont ...

- ශ්‍රී (Correct Form) -> ශ්‍රී+0dca (hal Kirima)+200c (ZWJ)+ඊ
- when 200c is removed: ශ්‍රී
- ශ්‍රී ලංකා (correct form): ශ්‍රී ලංකා (Country Name)
- සත්‍ය (correct Form) : සත්‍ය (truth in Sinhala)

# Repaya

- In addition, ZWJ is used for rendering Repaya or Reph form and is one of the three consonant conjuncts Sinhala has.
- Repaya form: 0dca 0dbb 200D ( ේ + ර + ZWJ)
- For instance, හරෂ will become හෂී.

# Internationalized Domain Names (IDNs)

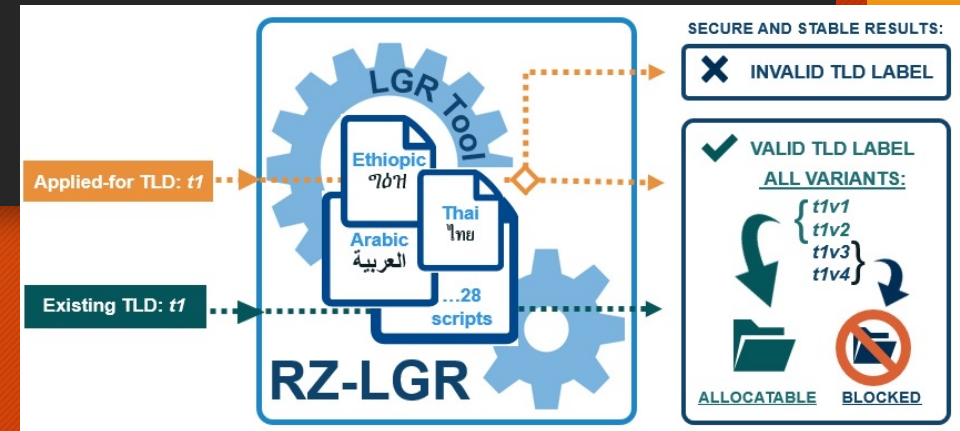
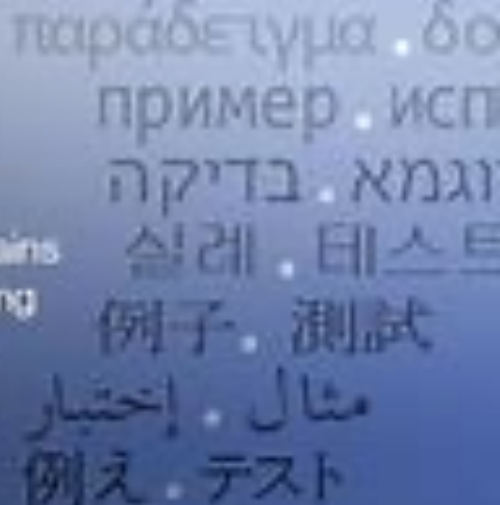
- An Internationalized Domain Name is a Internet Label which contains at least one label displayed in a software application in a whole or part in Non Latin Script or Alphabet such as Arabic, Hebrew, Sinhala, Tamil, etc.
- These Non Latin Scripts are encoded in Computer Systems or devices using Multi Bytes Unicode. However, Internationalized Domain Names are stored in Domain Name Systems (DNS) using special ASCII strings which is known as Punycode.
- විශ්වසම්මුති.ලංකා equivalent Punycode: xn--n0cva4aafp8cd8eiw.xn--fzc2c9e2c



# IDNs...

## — Internationalized Domain Names

New generic Top-Level Domains can be any language, including non-Latin scripts like Arabic, Chinese and Cyrillic.



# IDNs cont ...

.city



.legal



## IDNs cont...

- IDNs originally proposed by Martin Durst and implemented in 1990 by Tan Juay Kwang and Leong Kok Yong under the guidance of Tan Tin Wee
- Although the Domain Name System supports non-ASCII characters, applications such as e-mail and web browsers restrict the characters which can be used as domain names for purposes such as a hostname. Strictly speaking, it is the network protocols these applications use that have restrictions on the characters which can be used in domain names, not the applications that have these limitations or the DNS itself

## IDNs cont...

- To retain backward compatibility with the installed base, the IETF IDNA Working Group decided that internationalized domain names should be converted to a suitable ASCII-based form that could be handled by web browsers and other user applications IDNA specifies how this conversion between names written in non-ASCII characters and their ASCII-based representation is performed.
- ICANN issued guideline for the use of IDNA in 2003.
- The conversions between ASCII and non-ASCII forms of a domain name are accomplished by a pair of algorithms called ToASCII and ToUnicode.

# Sinhala GP - Languages Using Sinhala Script

- ⦿ Sinhala Script is primarily used in Sri Lanka to write Sinhala language which belongs to Indo-European Language family and Indo-Arya Sub family
- ⦿ The Script is Abugida Script which sprang from family of Southern Brahmi Script to which Telugu, Malayalam and Tamil belong to
- ⦿ Languages covered by the script:
  - Sinhala
  - Pali
  - Sanskrit

# Code Point Repertoire

- Starting from MSR-3, the repertoire includes:
  - 72 code points
  - 4 sequences
- The repertoire excludes:

#	Unicode Code Point	Glyph	Character Name	Reason for exclusion
1	0D8E	සෘ	SINHALA LETTER IRUUYANNA	Usage unknown
2	0D8F	ඌ	SINHALA LETTER ILUYANNA	Usage unknown
3	0D90	ඌඹ	SINHALA LETTER ILUUYANNA	Usage unknown
4	0D9E	ඳ	SINHALA LETTER KANTAJA	Not in modern usage
5	0DA6	ඳ	SINHALA LETTER SANYAKA	Only used in the word ‘ඉඳුරු’ (this word is used to call dogs)
6	0DDF	ඹ	SINHALA VOWEL SIGN GAYANUKITTA	Usage unknown
7	0DF3	ඹ	SINHALA VOWEL SIGN DIGA GAYANUKITTA	Usage unknown

# Cross-Script Variant Analysis

- Sinhala GP concluded there is no cross-script variant rules
- Following are confusable cases
  - U+0D82 (SINHALA SIGN ANUSVARAYA, ീ)

Sinhala	Telugu	Kannada	Malayalam
ീ (U+0D82)	ీ (U+0C02)	ೀ (U+0C82)	ഀ (U+0D02)

- U+0D83 (SINHALA SIGN VISARGAYA, ു)

Sinhala	Devanagari	Gujarati	Telugu	Kannada	Malayalam
ു (U+0D83)	ः (U+0903)	ઃ (U+0A83)	ృ (U+0C03)	ೃ (U+0C83)	ഃ (U+0D03)

# Universal Acceptance

- Universal Acceptance is coined by Ram Mohan of Aflias in 2001.
- He put forward as Universal Acceptance to represent every Top Level Domain regardless of Script, Number of Characters and how new it is.
- For the principle of Universal Acceptance to be realized, all valid domain names and email addresses must be accepted, validated, stored, processed and displayed correctly and consistently by all Internet-enabled applications, devices





# Universal Acceptance cont...



ACCEPT



VALIDATE



STORE



PROCESS

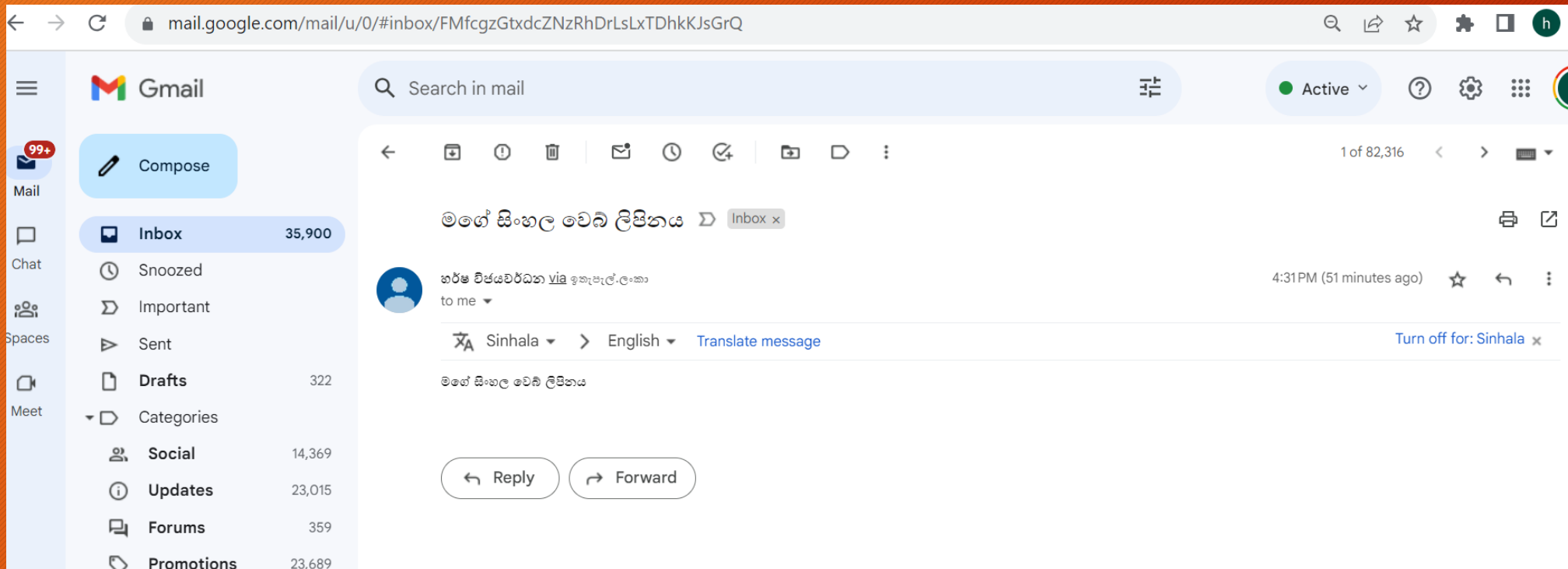


DISPLAY

# SLS Tamil Input Standard

- Sri Lanka standardized the Tamil Input in 2008 by releasing SLS 1326. Sri Lanka Tamil users prefer using the Renganathan keyboard, where the Tamil vowel modifier Kombuva is typed before a consonant.
- To formalize, Tamil input in Sri Lanka, the ICTA, under its local initiative, standardized the Renganathan keyboard as SLS 1326. LLWG participates actively with many groups of Tamil users and scholars worldwide, providing input into maintaining the Tamil Unicode standard.

# Universal Acceptance



# Future

- Encode Rakaaraansaya and yansaya with forward and backward compatibility to the existing content on the Internet using Unicode Normalization forms: Canonical (NFC, NFD), and Compatibility(NFKC, NFKD) Normalization Forms.
- Theekhana configured ඉන්ටර්නේට්.ලංකා to have email accounts in English, Sinhala, and Tamil in three languages and scripts receiving mail into a single mail folder. A user, if the person asks for email addresses in three scripts, can send any mail using any or all of the three scripts.

# Future cont ...

- The present Internet will eventually evolve into the Metaverse. Theekshana is exploring how Domain labels will work in the above scenario.