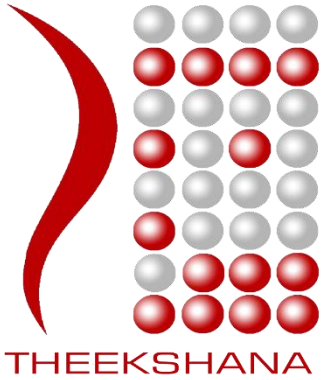


The recent work on Internationalized Domain Names (IDNs) in South and East Asian Scripts and especially in Sinhala

Harsha Wijayawardhana B.Sc.,FBCS,CITP

COO/CTO Theekshana R & D,

Director, BoC and LK Domain Registry

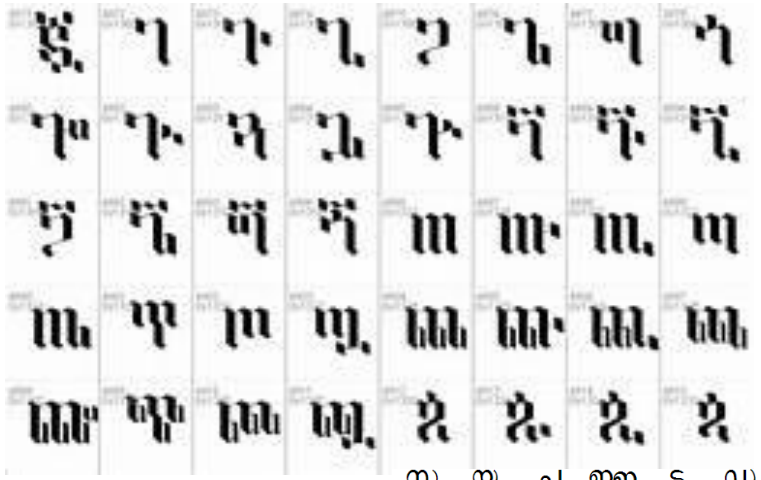
About Sinhala

- Sinhala belongs to Indo-European Language family and sub family of Indo- Arya.
- Closest Relatives of Sinhala Language are from North Indian family such as Hindi, Urdu, Bengali, Orya, etc.
- Europe has several of Indo-European Sub-families such as German, Romanic, and East Slavic.
- English, German and Swedish belong to German Sub Family whereas French, Portuguese, Spanish and Italian belong to Romanic Languages.

සිංහල

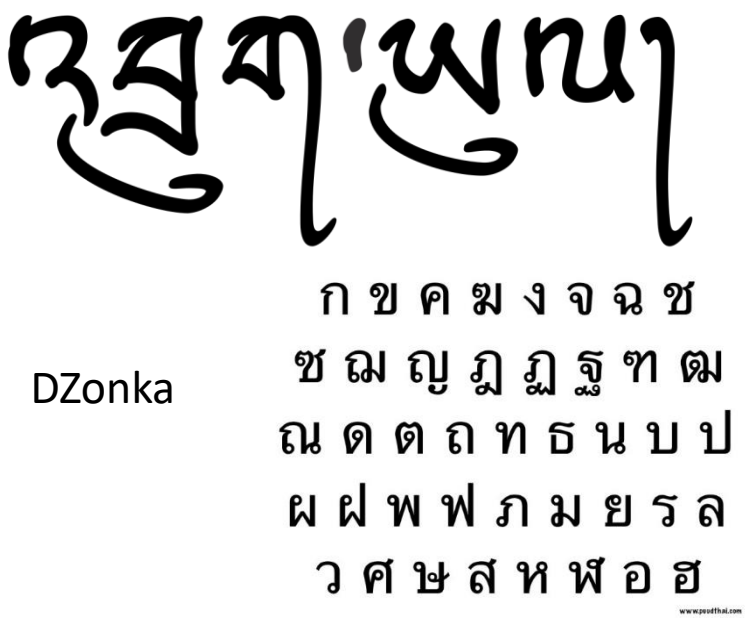
Written Script

- Modern Indian and East Asian scripts derive from Brahmi Script and they belong to Abugida Scripts



Ethiopic Script

Malayalam



DZonka

ก ข ค ฆ ง จ ฉ ช
ซ ฌ ญ ฏ ฏ ฐ ฑ ฒ
ณ ด ต ถ ท ธ น บ ป
ผ ฝ พ ฟ ภ ม ย ร ล
ว ศ ษ ส ห ฬ อ ฮ

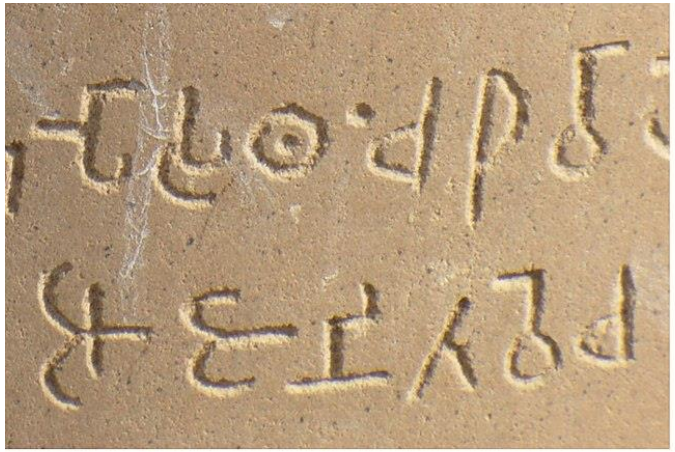
Thai

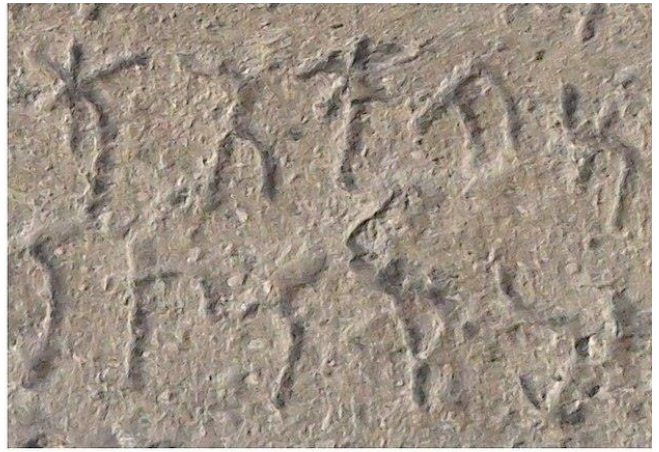
Brahmi

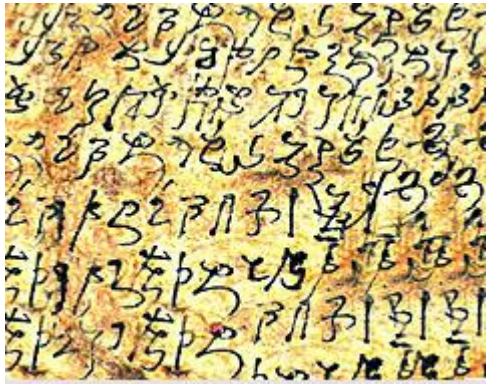
- All these Indian, East Asian Scripts derived from Brahmi.
- Although it was once believed that Emperor Asoka used Brahmi as his regal script, it has discovered that oldest Brahmi is found in Sri Lanka and Southern India.
- Brahmi is usually written from right to left. In Sri Lanka, eighteen (18) Rock inscriptions have discovered with Brahmi written from right to left dating to 3rd century BC.
- In North Eastern part of Asoka's Empire, his edicts were written in Prakrit using Kharoshthi Script. Kharoshthi was written from right to left.

Brahmi cont.

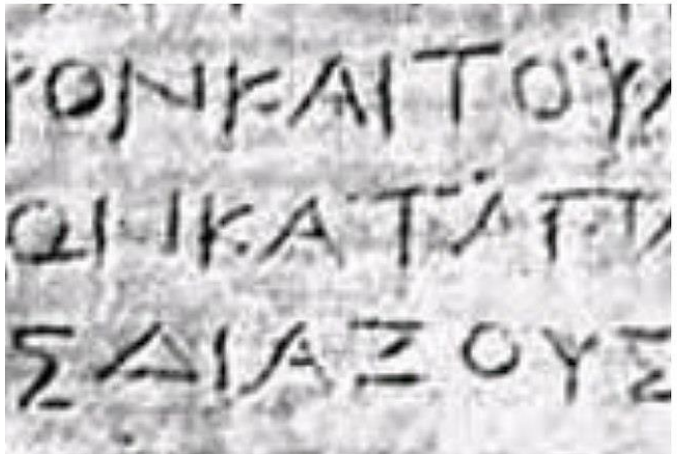
Brahmi and Karoshthi

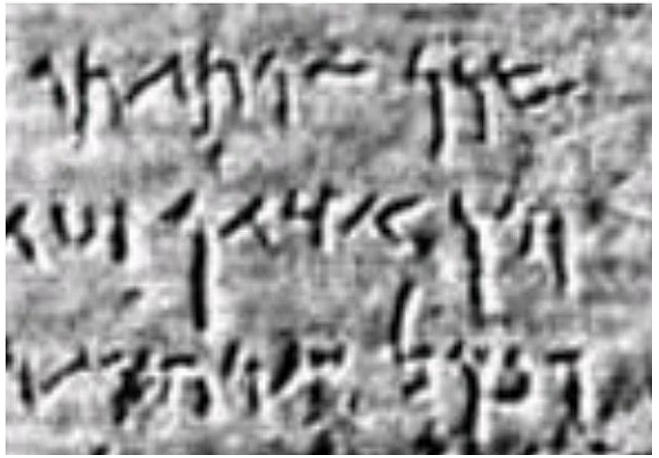


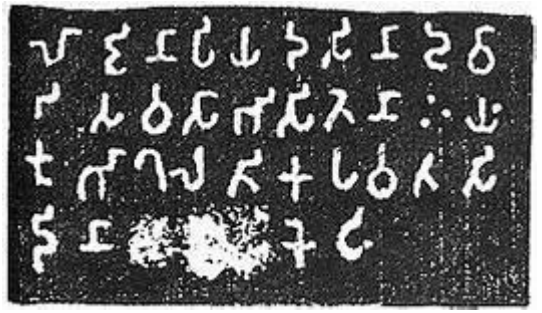




Kharosthi Script

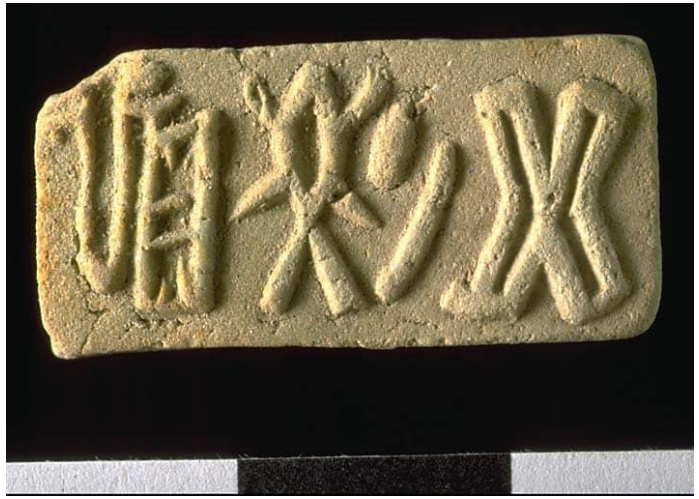






Brahmi Script cont.

- It was earlier postulated that Brahmi derived from Aramaic.
- Another hypothesis is Brahmi derived from Indus Valley script.



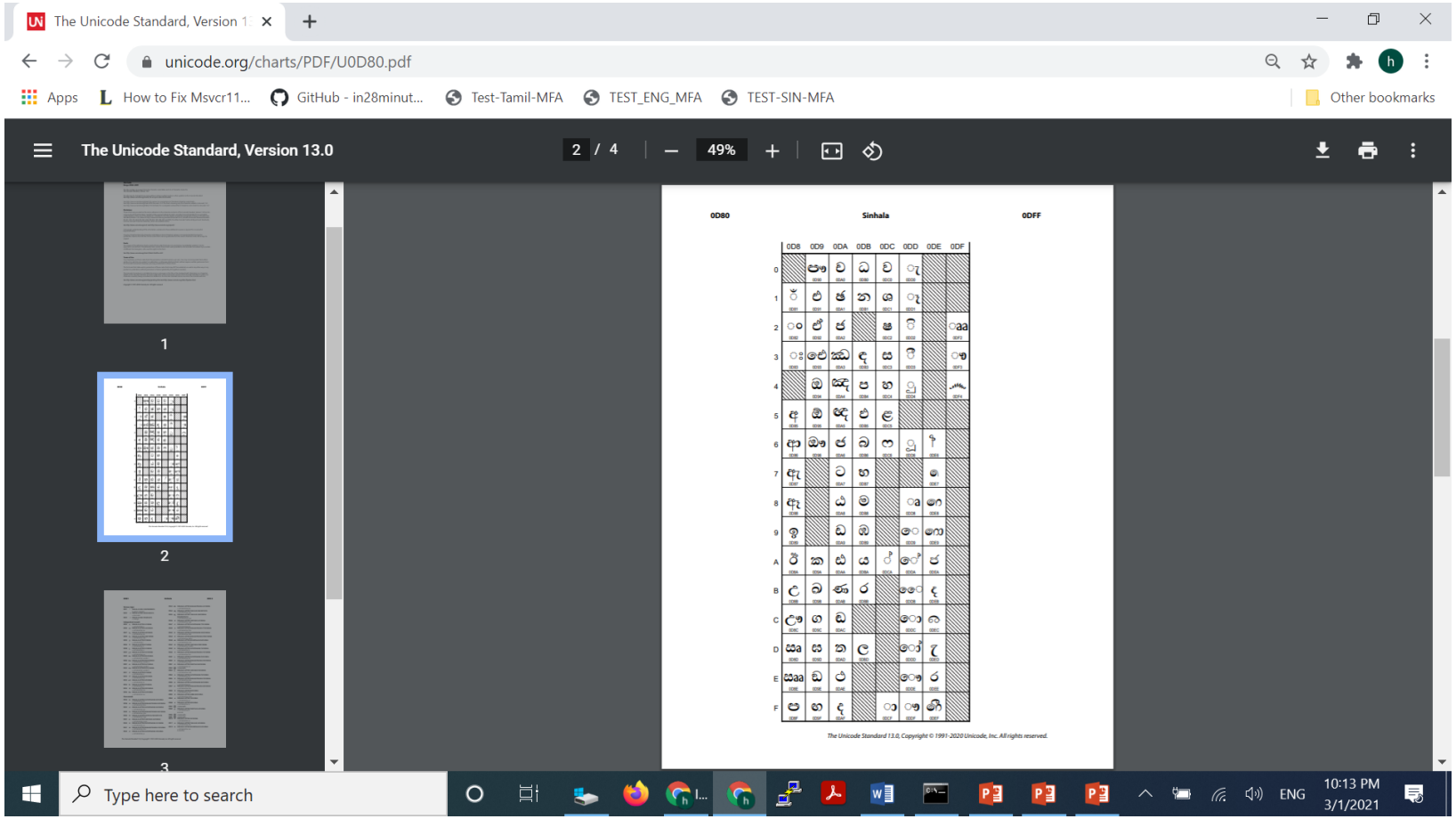
Sinhala and other Indic Scripts

- Sinhala Script derived from Southern Brahmi Script.
- Subsequently, Sinhala Script was influenced by Pallava Grantha.
- Eastern Scripts such as Khmer, Myanmar, etc. influenced by Pallava Script.
- Sinhala has two Vowels ෟ and ෲ.
- These two vowels are found in East Asian Scripts.

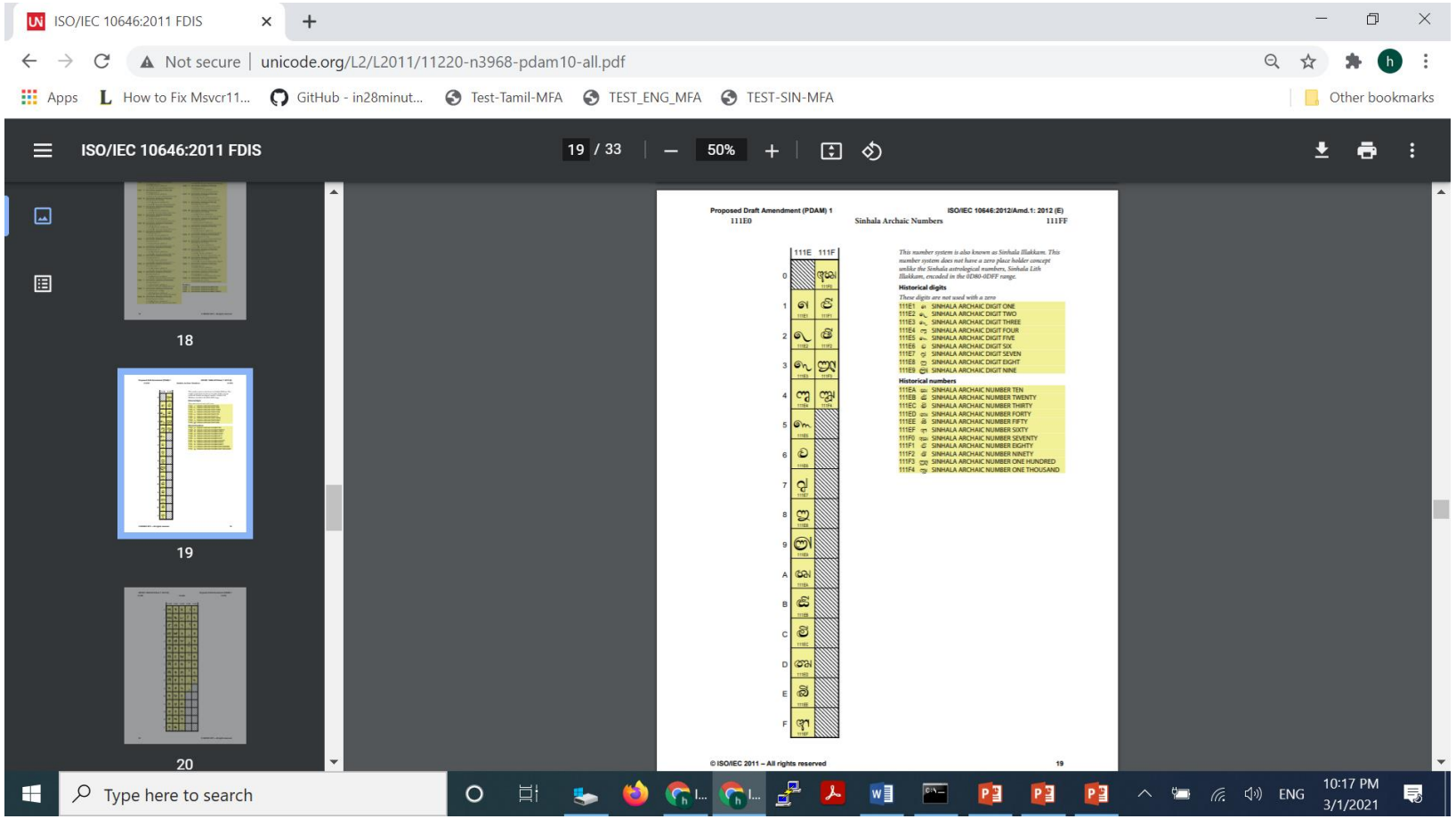
Sinhala Computerization

- Various attempts on Sinhala Computerization begun in early eighties.
- Sinhala was not encoded until 1996.
- Sinhala Unicode Implementation began in early two thousand.,
- Sri Lanka came up with input standard SLS 1134 and in 2010, Sinhala Unicode was updated for the last time with the third revision by including Sinhala Numerals.
- Sri Lanka also has its own standard for Tamil input since Sri Lanka Tamil community prefers to use input method deviating from Indian Implementation: SLS 1326 of 2008.

Unicode Version 13.1

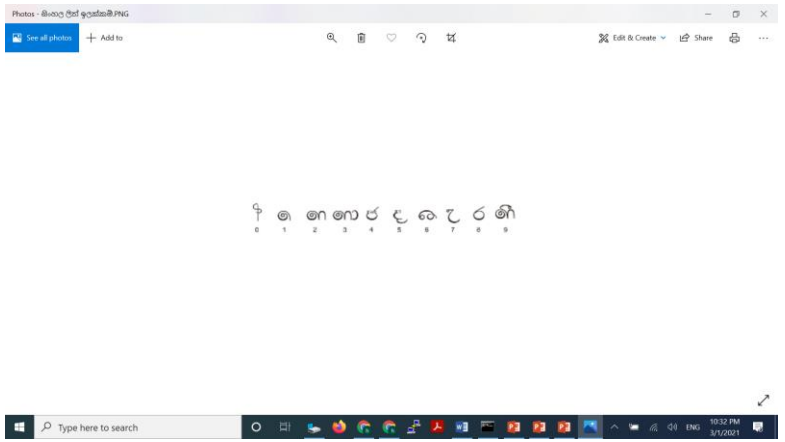


Sinhala Archaic Numbers in Supplementary Multilingual Plane (SMP)



Sinhala Numerals

- Although five ways of writing numbers found, only two have been encoded.
- The Sinhala Illakkam and Lith Illakkam have evolved from Brahimi numerals.
- Lith Illakkam has a zero but not the Sinhala Illakkam.



IDNs

- Sri Lanka began working in Internationalized Domain Names early Two Thousand becoming very first country to come up with two Non Latin Script ccTLDs: .ලංකා and .இலங்கை
- Sri Lanka further worked on Second Level for ccTLDs.
- Language Technology Research Lab (LTRL) of University of Colombo School of Computing (UCSC) under Dr. Ruvan Weerasinghe and University Moratuwa under the guidance of Prof. Gihan Dias carried out research into IDNs.
- One of the earliest research papers on IDNs by LTRL:
- https://www.researchgate.net/profile/Chamila-Liyanage-2/publication/228715570_Implementation_of_Internet_Domain_Names_in_Sinhala/links/0deec52148a10198fc000000/Implementation-of-Internet-Domain-Names-in-Sinhala.pdf

Sinhala Generation Panel of ICANN

- ICANN formed the Sinhala Generation Panel in December 2017. Indian Generation Panels met in Sri Lanka while Sri Lanka formed its panel and began in earnest.
- By mid 2018, Sri Lanka was ready to release first- draft on rules and by August, we managed to finish most of the work and released for the public comment. At present, we have released rules for second level.

Sinhala GP – Languages Using Sinhala Script

- Languages covered by the script:
  - Sinhala
  - Pali
  - Sanskrit
- Starting from (Maximal Starting Repertoire)MSR-3, the repertoire includes:
  - 72 code points
  - 4 sequences

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	ඍaa	SINHALA LETTER IRUUYANNA	Usage unknown
2	0D8F	ඏ	SINHALA LETTER ILUYANNA	Usage unknown
3	0D90	ඐ	SINHALA LETTER ILUUYANNA	Usage unknown
4	0D9E	ඞ	SINHALA LETTER KANTAJA	Not in modern usage
5	0DA6	ඦ	SINHALA LETTER SANYAKA	Only used in the word ‘ඉඳුඃ’ (this word is used to call dogs)
6	0DDF	ෟ	SINHALA VOWEL SIGN GAYANUKITTA	Usage unknown
7	0DF3	ෳ	SINHALA VOWEL SIGN DIGA GAYANUKITTA	Usage unknown

In-Script Variant Analysis

- Sinhala GP decided the following are in-script variant code points due to similar shapes and characters which could be used interchangeably
  - ඝ (U+0DC3) and ඝ (U+0D9D)
  - ඛ (U+0DB6) and ඛ (U+0D9B)
  - ඟ (U+0DC4) and ඟ (U+0DB7)
  - ච (U+0DA0) and ච (U+0DC0)
  - ඹ (U+0D94) and ඹ (U+0DB9)
  - ඵ (U+0D91) and ඵ (U+0DB5)
  - සa (U+0D8D) and සa (U+0DC3 U+0DD8)
  - ඓ (U+0D93) and ඓ (U+0DB5 U+0DD9)
  - ඒ (U+0D92) and ඒ (U+0DB5 U+0DCA)
  - ඕ (U+0D95) and ඕ (U+0DB9 U+0DCA)

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)

Cross-Script Variant Analysis

- Following are confusable cases (cont.)
  - Sinhala and Malayalam

Sinhala	Malayalam
ශ (U+0D9C)	ശ (U+0D36)
ශ්‍ර (U+0DC1)	ശ്ര (U+0D36)
◌ා (U+0DCF)	◌ാ (U+0D3E)

  - Sinhala and Myanmar

Sinhala	Myanmar
ශ (U+0D9C)	သ (U+1010)
ශ්‍ර (U+0DC1)	သ္ရ (U+107B)

Whole Label Evaluation Rules

- Code point category
  - C → Consonant
  - V → Vowel
  - M → Matras / Vowel Signs
  - B → Anusvara (Bindu)
  - X → Visarga
  - H → Halanta / Virama
  - J → Sannjakas

- Whole Label Evaluation Rules
  - H: must be preceded by C
  - M: must be preceded by C or J
  - X: must be preceded by either V, C, or M
  - B: must be preceded by either V, C, J or M

Some of the Issues -

- IDNA 2003 did not support hidden Characters such ZWJ, ZWNJ, etc.
- Without those characters, in Sinhala and some other Indic scripts had issue of rendering some of commonly used characters such as ශ්‍රී without country name cannot be written. ZWJ is needed for Rakar, and Yansa forms to render.
- IDNA 2008 now support the above and puny code allows these characters. And they allowed now in the second level.

Thank you

Questions ?