
ICANN
ANNUAL GENERAL
60
ABU DHABI
28 October–3 November 2017

Neo-Brahmi Generation Panel Meeting

Overall Progress

Ajay Data, Mahesh Kulkarni, Udaya Narayana Singh, NBGP Co-Chairs

Presented by Akshat Joshi

ICANN 60
31 October 2017

Why the name Neo-Brāhmī ?

Background:

- ICANN began “IDN Variant TLDs” initiative with six generation panels, based on specific scripts from major writing systems.
  - Latin, Greek, Cyrillic, CJK, Arabic and **Devanagari**
- Devanagari was representative of the **Brāhmī** family scripts
- The Brāhmī script
  - progenitor of all scripts used to write Modern Indo-Aryan languages
  - Dravidian
  - to a lesser extent scripts of the Tibeto-Burman and Munda families.
  - Also adopted by a large number of cultures in Southeast Asia to transcribe their languages: Burmese, Thai, Lao, Khmer (in South-East Asia), and others in Central Asia
- The **Neo-Brāhmī** group is so named to cover all such scripts used today and which are based on Brāhmī

Features of Brāhmī based scripts

- Brāhmī is written from **left to right**
- It has an angular shape. As it evolved this angular feature was gradually replaced by rounded shapes in cultures where palm leaves where used as a medium of written communication.
- The main feature of Brāhmī is the written syllable or **akshara** a concept admitting a **full Consonant or Vowel as a node**
- Vowels admit Vowel Modifiers such as nasals or vowel lengtheners
- **Consonants** are at times **modified by** a combining mark functioning as "vowel-killer" (termed **Halanta**), truncating the following vowel and thereby constituting a conjunct
- In turn these can be modified by vowel signs: Matras and further by Nasals or Vowel lengtheners
- The **adjuncts** to the Vowel or Consonant nodes are **appended in a strict rule-order**
- **This feature has been remarkably stable** over the evolution of Brāhmī and has been followed by all the later Indic and Southeast scripts derived from the script.

Principal Neo-Brāhmī Languages South Asian Scripts

- **Devanāgarī**: Devanāgarī is currently used for 11 out of 22 official languages of India (Boro/Bodo, Dogri, Hindi, Kashmiri, Konkani, Maithili, Marathi, Nepali, Sanskrit, Santhali and Sindhi) and around 45 other languages especially the related Indo-Aryan languages: Bagheli, Bhili, Bhojpuri, Himachali dialects, Magahi, Newari and Rajasthani and its dialects: Marwari, Mewati, Shekhawati, Bagri, Dhundhari, Harauti and Wagri. Nepali is the official language of Nepal as well as one of the official languages of the state of Sikkim in India. Spoken by over 30 million people. The script is also used in Fiji to represent Fiji Hindi. Hindi is also used in Mauritius, Malaysia, England, Canada, South Africa, Indonesia as well as emigrant communities around the world.
- **Gujarati**: Used for writing Gujarati and Kacchi, Gujarati is extensively spoken in large parts of Africa, Madagascar, UK and the USA as well as by emigrant communities around the world.
- **Gurumukhi** which evolved separately in the Northern family is used to write the Punjabi language in the Indian state of Punjab and elsewhere in India. Gurmukhi stabilized around the 16th century when it was used to transcribe the holy Granth Sahib.
- **Bengali**: Often termed as Bangla by linguists and grammarians is historically related and similar in design to the Devanāgarī script and with one or two exceptions has the same consonant and vowel set. Bengali is used to transcribe quite a few languages of which the most prominent are Assamese and Manipuri. The former differs from Bengali in a few consonant characters. The same is the case with Manipuri which today is also written in Meetei Mayek.

Principal Neo-Brāhmī Languages South Asian Scripts

- **Oriya [Odia]** can be traced back to the Ashokan inscriptions: 3rd century B.C.E. Because of the prevalence of a large number of tribal languages belonging to the Munda and Dravidian families in the state of Odisha (Orissa), the Oriya script is used in writing these languages.
- Sinhala used for writing Sinhala language and at times also Pali, is derived from Brāhmī as early as the third-second century B.C.E. Although it belongs historically to the Northern family, it has been considerably influenced by the early Grantha script of South India.
- **Tamil** (also spelt as "Tamizh") More than any other script derived from Brāhmī, it is highly alphabetical in nature and admits no ligatures with the exception of two consonant conjuncts. Apart from being the official language of Tamil Nadu, Tamil is also an official and national language of Sri Lanka and one of the official languages of Singapore. Tamil is also spoken by significant minorities in Malaysia, England, Mauritius, Canada, South Africa, Fiji, Indonesia, as well as emigrant communities around the world.
- **Kannada and Telugu** are closely related scripts used to write two Dravidian languages: Kannada in the state of Karnataka, and Telugu in Telangana and Andhra. Over the centuries, Brāhmī evolved with marked characteristics in the south. Around the tenth century, these crystallised into the Old Kannada script, used where both Kannada and Telugu are now spoken. By around 1500, this script divided into Kannada and Telugu. As a result, there are very few differences between these two scripts.
- **Malayalam**: Subject to reforms, modern Malayalam has introduced alphabetic writing into the script, although the main structure of Malayalam still adheres to the akṣara.

Current status of Neo-Brahmi GP Work

- Devanagari LGR is almost ready and has been sent to Integration Panel for their review.
- In the course of this presentation, we plan to present the same in detail.
- Once the Devanagari LGR is finalized, other LGRs will soon follow the suite. Respective script/language community members have already begun the ground work.

Before starting with the

Devanagari LGR

let us take a look at

Akshar Formalism that binds brahmi based scripts

Character classification

Components of the Syllable

- **Consonants(C) :**
  - क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द ध न ऩ प फ ब भ म य य़ र ऱ ल ळ ऴ व श ष स ह ग़ ज़ ड़ ब़
- **Vowels (V) :**
  - ऍ अ आ इ ई उ ऊ ऋ ऎ ऐ ए ऐ ऑ ओ ओ औ ॲ अं आं औं अ॒ आ॒
- **Vowel Signs / Matras (M) :**
  - ॅ ा ि ी ु ू ृ े ै ॉ ॊ ो ौ ॅ ं ः ॅ ॅ ॅ
- **Vowel modifiers (D) :** ँ ं ः
- **Halant (H) :** ्
- **Nukta (N) :** ़

Formalism at a glance ...

Formalism Illustrated...

- **Variables :**

Dash	→	Hyphen -
Digit	→	Indo-Arabic digits [0-9]
C	→	Consonant
V	→	Vowel
M	→	Matra
D	→	Anusvara/Bindi/Tippi/Sunna
B	→	Chandrabindu/Anunasika/Arasunna
X	→	Visarga/Aytham
H	→	Halant/Chandrakala/Virama
A	→	Addak
N	→	Nukta
Y	→	Avagraha/Praslesham
L	→	Chillu
Z	→	Khanda Ta
k	→	Number of possible Consonant Halanta Sequences

Formalism Illustrated…

- Formalism Operators :

|	→	Alternative
[ ]	→	Optional
*	→	Variable Repetition
( )	→	Sequence Group

Formalism Illustrated…

Consonant-Syllable →

*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]
| [CH]Z
| L[HC[D|H|M[D]]]
| AC[D|X|M[D|X]]

Vowel-Syllable → V[D|B|X|BD|BX]

Syllable → Consonant-Syllable [Y] | Vowel-Syllable[Y]

IDN-Label → (Syllable | digit)*([dash](Syllable | digit))

Formalism Illustrated..

Consonant-Syllable :

*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]
| [CH] Z
| L[HC [D | H | M[D]]]
| AC[D | X | M[D|X]]

ABNF Illustrated..

Consonant-Syllable :

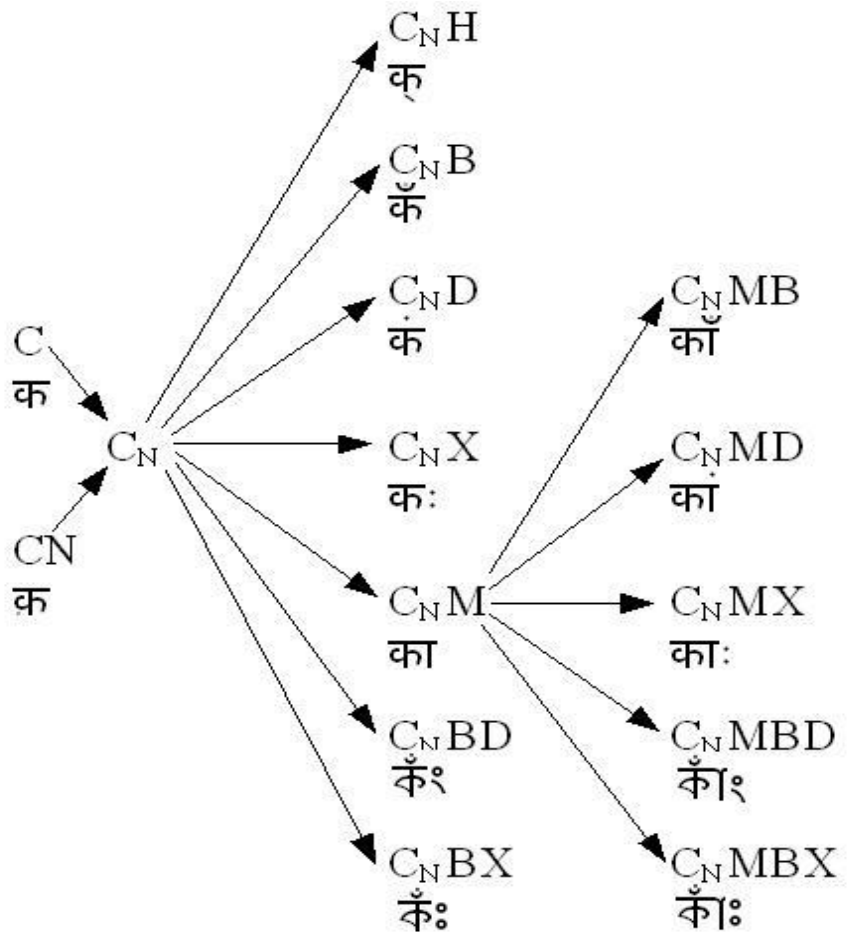
*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]

| [CH] Z

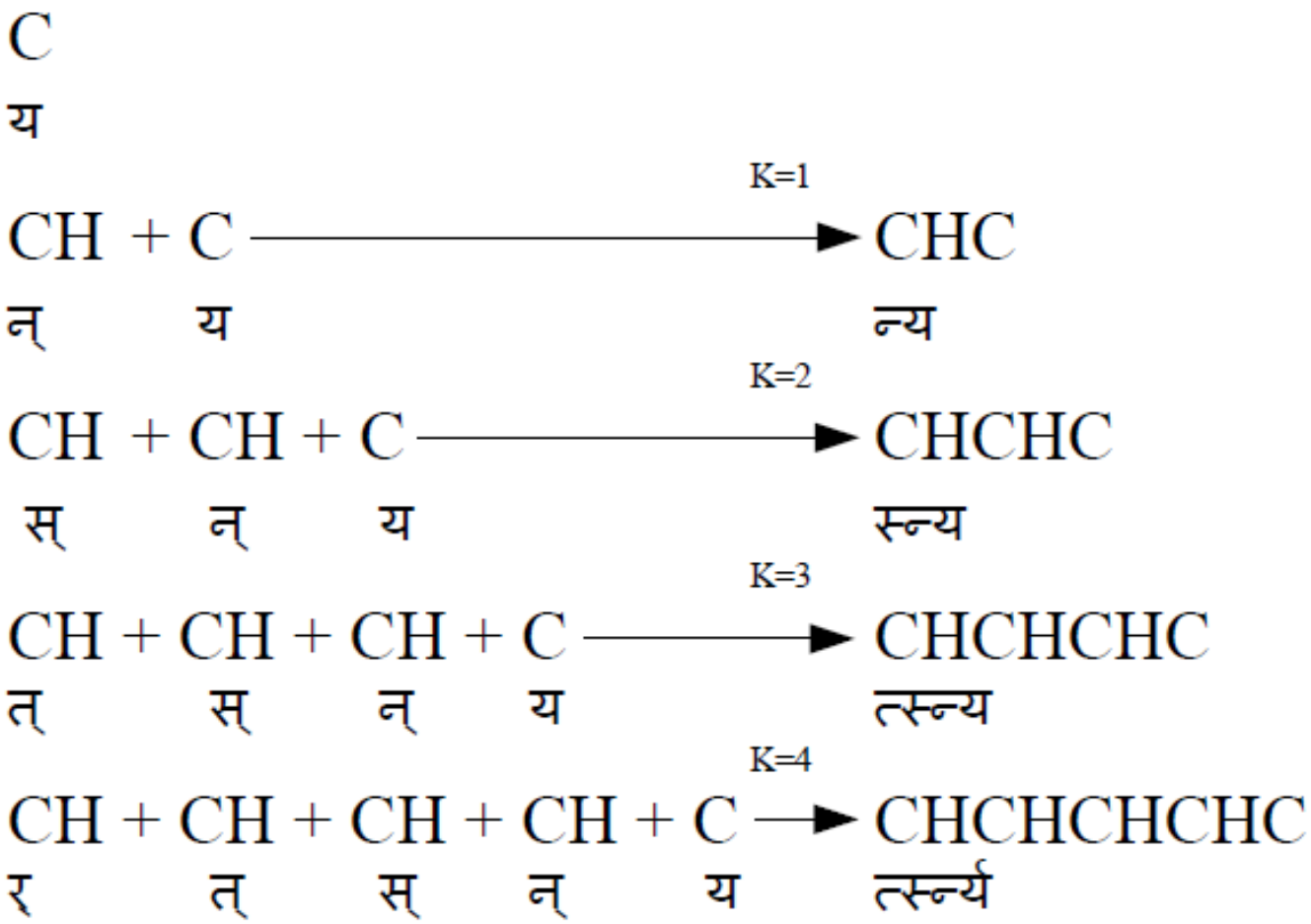
| L[HC [D | H | M[D]]]

| AC[D | X | M[D|X]]

*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]



*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]



ABNF Illustrated...

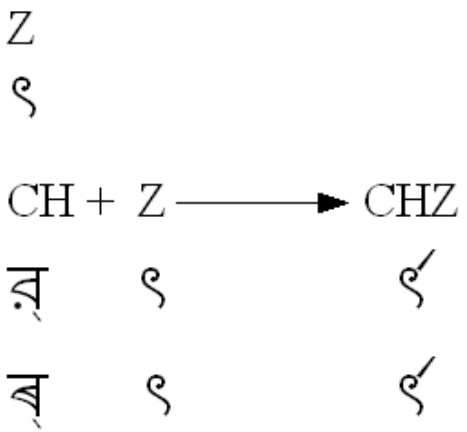
Consonant-Syllable :

*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]
| [CH] Z
| L[HC [D | H | M[D]]]
| AC[D | X | M[D|X]]

Consonant Syllable continues...

Syllable with Khanda Ta only exists in Bangla and Assamese language.

[CH] Z



ABNF Illustrated...

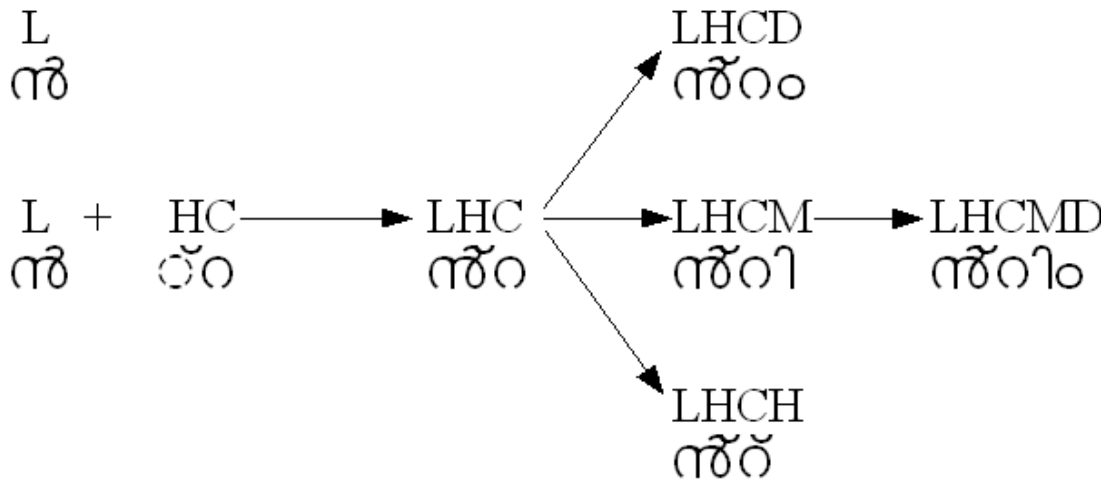
Consonant-Syllable :

*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]
| [CH] Z
| L[HC [D | H | M[D]]]
| AC[D | X | M[D|X]]

Consonant Syllable continues...

Syllable with Chillu characters only exists in Malayalam language.

L[HC [D | H | M[D]]]



ABNF Illustrated...

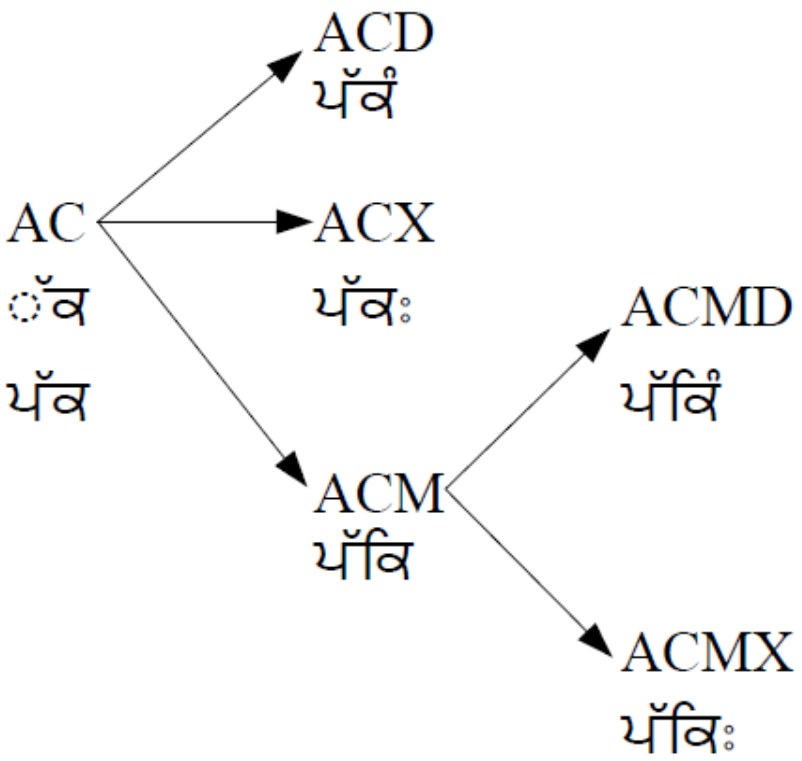
Consonant-Syllable :

*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]
| [CH] Z
| L[HC [D | H | M[D]]]
| AC[D | X | M[D|X]]

Consonant Syllable continues...

Syllable with Addak only exists in Punjabi language.

AC[D | X | M[D|X]]



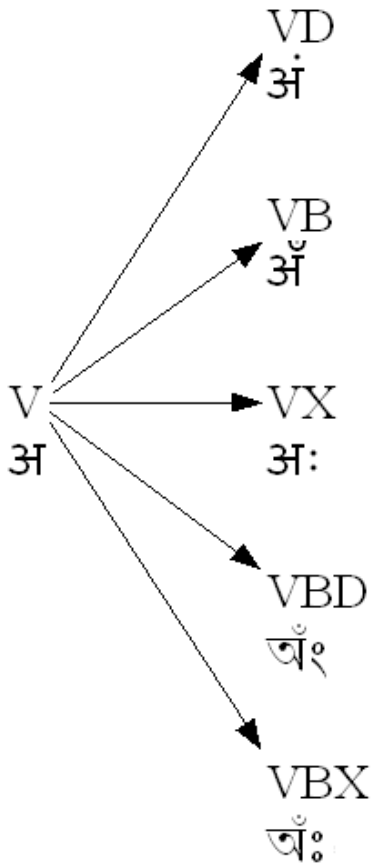
ABNF Illustrated..

Vowel-Syllable :

V [D | B | X | BD | BX]

Vowel Syllable continues...

V [D | B | X | BD | BX]



ABNF Illustrated...

Syllable :

Consonant-Syllable | Vowel-Syllable

-where

Consonant-Syllable →

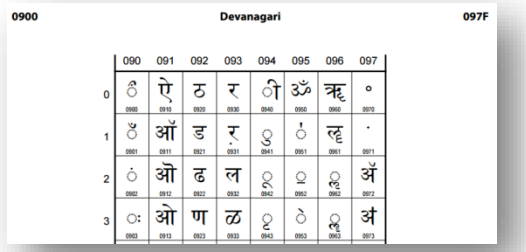
*k(C[N]H) C[N] [H|D|B|X|BD|BX|M[D|B|X|BD|BX]]
| [CH]Z
| L[HC[D|H|M[D]]]
| AC[D|X|M[D|X]]

Vowel-Syllable → V[D|B|X|BD|BX]

Devanagari LGR

Devanagari LGR

- **Fundamental Blocks:**
  - Code point repertoire
  - Whole Label Evaluation rules
  - Variant Rules



अद्रक
अद्रक
अद्रक

Root LGR procedure

- Binding principles:
  - LONGEVITY PRINCIPLE
  - LEAST ASTONISHMENT PRINCIPLE
  - INCLUSION PRINCIPLE
  - SIMPLICITY PRINCIPLE
  - PREDICTABILITY PRINCIPLE
  - STABILITY PRINCIPLE
  - LETTER PRINCIPLE

Root LGR procedure

- **LONGEVITY PRINCIPLE**

  The panels are supposed to begin using the latest version of Unicode, but also to take into consideration the stability of Unicode character properties.

  - Devanagari LGR proposes almost all those characters which have been fairly stable over past many versions of the Unicode.
  - Most of the characters exist in the Unicode right from Unicode Version 1.
  - Some characters were encoded in successive versions of Unicode which were very language specific. Hence their usage too was unambiguous.
  - Latest character(s) included by NBGP was introduced in Unicode 6.0 in Oct. 2010.

Root LGR procedure

- **LEAST ASTONISHMENT PRINCIPLE**

  The Least Astonishment Principle aims at ensuring that the allocated code points included in the zone repertoire are useful as elements in unique identifiers. To the extent that a code point is confusing to the user population or can be used in surprising ways –whether to members of the original linguistic target community or, in the case of the root, to members of other linguistic communities – use of the code point fails to adhere to the Least Astonishment Principle in that context.

  - NBGP has documented some cases in it's analysis of Variants which can be confusing to the majority of the Devanagari community itself.
  - However, for some languages, those cases are an absolute must.
  - NBGP has taken a pragmatic stand of proposing equivalent (blocking) variants wherever possible.

Root LGR procedure

- **INCLUSION PRINCIPLE**

  The procedure is an example of the Inclusion Principle in action, since every rule or code point is excluded until reviewed and then explicitly included.

  - This methodology was adopted and diligently followed during the character selection process.

Root LGR procedure

- **SIMPLICITY PRINCIPLE**

  Part of the point of having the integration panel is that it performs a check of the Simplicity Principle. The integration panel cannot possibly include experts in every language and script, but the members must have general knowledge of Unicode, IDNA, DNS, or all of the above. If any member of the integration panel cannot understand the rationale for inclusion of some rule, then that member will not support the rule, and it will not proceed. This is the purpose of the unanimity requirement for the integration panel.

  - NBGP has to the extent possible, tried to frame all the WLE rules in simple yet understandable way.
  - Also, NBGP does have community members on-board with general knowledge of Unicode, IDNA and DNS.
  - During the drafting phase, none of the rules proposed had contrasting views on it.

Root LGR procedure

- **PREDICTABILITY PRINCIPLE**

  The proposal follows the Predictability Principle in much the same way it follows the Simplicity Principle: if the integration panel does not immediately agree with the recommendations of the generation panel, or if members of the integration panel disagree with each other, that is a good reason to suppose that the rule in question is not really predictable.

  - At this stage, NBGP hopes that this gets complied with during IP review. ☺

Root LGR procedure

- **STABILITY PRINCIPLE**

  Especially in the case of the root zone, the Stability Principle is less a matter of guidance and more a statement of fact. The proposed procedure attempts to minimize the possibility that any label generation rule will be permitted for the root zone without that rule having been considered as carefully as possible for any negative consequences. If there is a failure such that the integration panel determines that a previously active rule needs to be removed, this procedure requires that the procedure itself be subject to review.

  - As discussed in Longevity principle.

Root LGR procedure

- **CONSERVATISM PRINCIPLE**

  The proposal is consistent with the Conservatism Principle in two ways. First and most important, because the integration panel is supposed to reject anything it does not positively think is safe, the Conservatism Principle is built in to the integration panel's criteria. Second, in the event of disagreement between the generation and integration panels, the proposed rule that is the subject of the disagreement is automatically excluded from the root label generation rules.

  - NBGP has tried to follow the principle while designing the LGR.

Devanagari Code Block - MSR

Convention:

White: Blocked by IDNA Protocol

Pink: Blocked by MSR

Yellow: Permitted by MSR for final decision by the GPs.

Proposed Devanagari LGR Code point repertoire

- ऱ
- ऩ
- ळ
- ऴ
- ॺ

Proposed Devanagari LGR – Additional rule for ऱ (U+0931)

Sr. No.	Unicode Code Points	Sequence	Character Names	Unicode General Category (gc)	Reference
1.	0931 094D 092F	ऱ्य	DEVANAGARI LETTER RRA DEVANAGARI SIGN VIRAMA DEVANAGARI LETTER YA	Lo Mn Lo	[INSCRIPT]
2.	0931 094D 0939	ऱ्ह	DEVANAGARI LETTER RRA DEVANAGARI SIGN VIRAMA DEVANAGARI LETTER HA	Lo Mn Lo	[INSCRIPT]

Proposed Devanagari LGR
- WLE Rules -

C	Consonant
M	Matra
V	Vowel
B	Anusvara
D	Chandrabindu
X	Visarga
H	Halant / Virama
N	Nukta
S	Eyelash Reph (C1HC2) where C1 is 0931 (ऱ - DEVANAGARI LETTER RRA) H is 094D (◌् - DEVANAGARI SIGN VIRAMA) C2 is either - 092F (य - DEVANAGARI LETTER YA) or 0939 (ह - DEVANAGARI LETTER HA)

- N: must be preceded only by either of specific set of Cs, Vs and Ms

The specific Cs are:

- क (U+0915)
- ख (U+0916)
- ग (U+0917)
- ज (U+091C)
- ड (U+0921)
- ढ (U+0922)
- फ (U+092B)

The specific Vs are:

- आ (U+0906) (Required in Santhali language)
- ओ (U+0913) (Required in Santhali language)

The specific Ms are:

- ◌ा (U+093E) (Required in Santhali language)
- ◌ो (U+094B) (Required in Santhali language)

- H: must be preceded by C or CN
- M : must be preceded by C or CN
- X: must be preceded by either of V, C, N or M
- B: must be preceded by either of V, C, N or M
- D: must be preceded by either of V, C, N or M
- V: Can NOT be preceded by H
- Case of Eyelash-Reph

Proposed Devanagari LGR

- Variants -

- Root LGR procedure ideally does not want homographic variants to be part of this variant set
- Most of the **prominent** variants in Brahmi Based scripts/languages **are Homographic**
- However, there are two categories of these Homographic Variants
  - Pure homographs (visual confusability due to shape)
  - Perceptive homographs (visual confusability due to deviation from known norms to majority of script community)

Proposed Devanagari LGR

- Variants -

- Confusingly similar
- NOT BEING PROPOSED AS VARIANTS

घ U+0918	ध U+0927
भ U+092D	म U+092E
ष्ट U+0937 U+094D U+091F	ष्ठ U+0937 U+094D U+0920

Proposed Devanagari LGR

- Variants -

- Variants due to requirement of Santhali of having Nukta follow specific Vowels and Matras
- PROPOSED AS VARIANTS

Variant 1	Variant 2
आ U+0906	आ़ U+0906 U+093C
ओ U+0913	ओ़ U+0913 U+093C
◌ा U+093E	◌ा़ U+093E U+093C
◌ो U+094B	◌ो़ U+094B U+093C

Proposed Devanagari LGR

- Variants -

- Variants due to Unique characters required for Kashmiri
- PROPOSED AS VARIANTS

Variant 1	Variant 2
ऀ U+0973	अं U+0905 U+0902
ऺ U+093A	◌ं U+0902
ऻ U+0974	आं U+0906 U+0902
ऻ U+093B	◌ा◌ं U+093E U+0902
ऐ U+090E	ऐ U+0910
◌ै U+0946	◌े U+0947
ऻ U+0975	औ U+0914
ऻ U+094F	◌ौ U+094C

Possible Variant Cases - Cross script

- Confusingly similar

DEVANAGARI SCRIPT	COGNATE SCRIPT	CODEPOINT IN COGNATE SCRIPT
VOWELS		
उ 0909	Bangla	ঙ 0993
उ 0909	Gurmukhi	ਤ 0A24
ऋ 090B	Gujarati	ઋ 0AE0
CONSONANTS		
क 0915	Bangla	ক 0995
ग 0917	Gujarati	ગ 0A97
ग 0917	Gurmukhi	ਗ 0A17
घ 0918	Gurmukhi	ਘ 0A2C
घ 0918	Gujarati	ઘ 0A98
ङ 0919	Gujarati	ઙ 0A99
छ 091B	Gujarati	છ 0A9B
ञ 091E	Gujarati	ઞ 0A9E

DEVANAGARI SCRIPT	COGNATE SCRIPT	CODEPOINT IN COGNATE SCRIPT
ट 091F	Gurmukhi	ਟ 0A17
ठ 0920	Gujarati	ઠ 0AA0
ठ 0920	Gurmukhi	ਠ 0A20
ड 0921	Gujarati	ડ 0AA1
ढ 0922	Gurmukhi	ਢ 0A2B
त 0924	Gujarati	ત 0AA4
ध 0927	Gujarati	ધ 0AA7
न 0928	Gujarati	ન 0AA8
न 0928	Bangla	ন 09A8
न 0928	Bangla	ণ 09A3
प 092A	Gujarati	પ 0AAA
प 092A	Gurmukhi	ਧ 0A17
प 092A	Gurmukhi	ਪ 0A2A

DEVANAGARI SCRIPT	COGNATE SCRIPT	CODEPOINT IN COGNATE SCRIPT
प 092A	Gurmukhi	ਪ 0A6B
म 092E	Gurmukhi	ਸ 0A38
म 092E	Gujarati	મ 0AAE
य 092F	Gujarati	ય 0A9A
र 0930	Gujarati	ર 0AAE
र 0930	Gurmukhi	ਰ 0A15
ल 0932	Bangla	ল 09B2
व 0935	Gujarati	વ 0AB5
श 0936	Gujarati	શ 0AB6
श् 0936+094D	Bangla	ঽ 09BD
ष 0937	Gujarati	ષ 0AB7
स 0938	Gujarati	સ 0AB8
ह 0939	Gujarati	હ 0AB9

DEVANAGARI SCRIPT	COGNATE SCRIPT	CODEPOINT IN COGNATE SCRIPT
Nukta characters		
ग़ 095A Or 0917+094D	Gurmukhi	ਗ਼ 0A5A
ढ़ 095D Or 0922+094D	Gurmukhi	ੜ 0A5E

F2F meetings

- Second F2F meeting in Nepal
  - Held between 24th to 26th May 2017 in Kathmandu, Nepal
  - Active participation from community in Nepal which provided great insights to the GP on Nepali, Newar and Tamang

- (Scheduled) Third F2F meeting in Colombo, Srilanka (Mid-Dec 2017)
  - Outreach to the Tamil community in Srilanka
  - Co-ordination with the Sinhala Generation Panel

Community Outreach – Language coverage

- Since the last update to community by NBGP, the Devanagari LGR has considered quite a number of languages falling within the EGIDS scale 1-4
- Previously covered:
  - Bodo, Dogri, Hindi, Kashmiri, Konkani, Marathi, Maithili, Nepali, Sanskrit, Santhali, Sindhi
- Recently covered:
  - Avadhi, Bhatri, Bhojpuri, Chhattisgarhi, Halbi, Kinnauri, Kukna, Limbu, Magahi, Newar, Panchpargania, Sadri, Tamang, Eastern, Wagdi
- Unable to cover: Saraiki

Feedback received from Integration Panel on draft on 8th Sep. 2017 and action taken

- Changes in wording
  - Wherever required, the changes were made.
  - Wherever change was not possible, more elaboration was added
- More explanation of certain aspects
  - Rationale behind exclusion set characters – yet to be added
- Major modifications suggested to XML submitted
  - The GP had not used the online tool at this point for validation hence there were validation errors, were rectified in next submission
- Harmonization of similar terms with a unique name
  - Done (Eyelash Ra/Reph, Matra/Vowel sign)
- Request for a more varied set of References
  - Inclusion of more reference online e.g. omniglot. Added.
- Request for reconsideration of Halanta ending variant as this is a null mapping variant which has no precedant in LGRs so far (However left upto GP to take a final call)
  - This was to safeguard possible confusion, however during the call with the community, this variant proposition was taken back

Feedback received from Integration Panel on draft on 11th Oct. 2017 and action taken

- More clarification sought on Sanskrit vis-à-vis Modern and Archaic usage
  - This will be added
- Changes in wording
  - Wherever required, the changes will be made
  - Wherever change not possible, more elaboration will be added
- Comments on 0944
  - This is a mistake and will be corrected. Method of correction yet to be finalized.
- Some un-addressed points during last review
  - The rationale behind certain decisions could not be provided to the IP due to major focus on getting the document done, this will be done

Feedback received from Integration Panel on draft on 11th Oct. 2017 and action taken

- Case of Halanta ending variant and discussion on the same to be added
  - Will be added
- Small changes to XML on technical details
  - Will be done
- More nukta characters e.g. For Konkani as given on Omniglot
  - This needs more discussion as the language community differs from Omniglot
- Changes to the XML
  - Descriptive sections to be populated – Will be added
  - Reference to adherence to RFC 7940 – Will be added
- Summarization of WLE, Variants in XML – Will be done

Future plan of action

- As decided by NBGP earlier, NBGP plans to get Devanagari LGR through and thereafter consider it as a basis for other script LGRs.
  - The Devanagari LGR can provide all the necessary placeholder sections of the document and the nature of the text required.
  - NBGP however does not assume that all the script LGRs can be easily modeled after Devanagari going by the rules. Some script specific rules are definitely expected for other scripts and they will need to be formulated carefully.
- However, NBGP members have already begun working on some of the scripts viz. Bengali, Gujarati, Gurmukhi, Kannada and Telugu.

धन्यवाद !