Briefing on Variance in Devanāgarī by Case-Study team to ICANN representatives

Singapore, 20 June 2011: 5:00-6:30 pm.

Report by Nicholas Ostler [NDMO] <nicholas@ostler.net>

In India, there are 22 official languages, 19 of them using scripts derived from the ancient Brahmi (via Siddha script), The most populous of these modern scripts is Devanagari, (the principal others being Gujarati, Gurmukhi, Eastern Nagari, Oriya, Telugu, Malayalam, Tamil, Kannada), the other 3 languages using Perso-Arabic script. Brahmi-derived scripts are left-right, Perso-Arabic right-left. If variance is fully analysed for Devanagari, it should be a simple matter to apply these principles to the other Brahmi-derived scripts (and hence the languages which use them).

Dr Kulkarni gave the following exposé, drawing on many years' experience in language technology at CDAC.

He distinguished three potential (general) sources of variance, with instances in Devanagari:

Homographs (where characters look similar - especially when the typeface is small):

e.g

द + र = द्र (da + ra = dra)

द + ग = द्ग (da + ga = dga)

द + न = द्न (da + na = dna)

Among this type there are also 'single-character variants', which are just look-alike characters, notably

घ gha

ध dha)

cf. English 1 vs l vs I, 0 vs O

Homophones (where character-strings sound similar)

e.g.

हिंदी (analysable as hĩ-dī = 'hindī')

हिन्दी (analysable as hi-ndī = 'hindī')

cf. English 'color' vs 'colour'

Case-variants (which are not present in Devanagari (– nor Chinese or Arabic))

e.g.

ASDFG vs asdfg in Latin

[(and similarly)

АСДФГ vs асдфг in Cyrillic

ΑΣΔΦΓ vs αδφγ in Greek

I note in passing that variants due to homographs between scripts, evident as between pairs of the three European scripts just mentioned, are unlikely to afflict Devanagari, Chinese or Arabic with their systematically very different shapes: though there might be the odd curiosity such as T vs Chinese 丁, and l vs Arabic ا. - NDMO]

The Indian practice is to concern itself **only** with homographs, as above. Homophones are therefore no concern (though both spellings above for 'Hindi' are in practice in totally free variation. [It may also be noted that the homographs above, also permit further homophones:

e.g.

द्र for dra

द्ग for dga

द्न for dna

[These are not correctly drawn, since there should be no space between the two glyphs.]

Since these result from sequences of Unicode codes distinct from those that produce their homophones above [inserting ZWNJ between the two], they are not considered variants in Indian practice.

Multiple homophones are of no more concern, since they do not look alike, and are characterizable by distinct strings of Unicode codes: e.g.

क + ् + ष = क्ष (ka + *halant* + ṣa -> kṣa)

क + ् + ZWNJ + ष = क्‌ष (ka + *halant* + ṣa -> kṣa)

क + ् + ZWJ + ष = कष (ka + *halant* + ṣa -> kṣa)

[This last not properly drawn: the क should lack its hook, as in क्स.)

ZWJ means 'zero-width joiner', and ZWNJ 'zero-width non-joiner'.

When registering strings of Devanagari, the "first come, first served" principle is observed. Therefore, if a word incorporating a homophone is registered, all words which differ only through substituting one of the recognized homophones for another are blocked. Once you have registered अद्रपि (adrapi), no-one could have अद्नपि (adnapi) or अद्गपि (adgapi).

[I gathered that 'single-character variants' (as instanced above) do not have this blocking effect. - NDMO]

The Indians organize their Devanagari characters into **variant tables**, which are language-independent, and refer to the script as a whole. [They are comparable in form to the variant tables which Andrew Sullivan envisaged, I believe - NDMO]

They also organize the whole array of possible characters in the Devanagari script into **language tables**, which simply show the whole array of characters [organized phonetically - as per the traditional Indian order - NDMO] but with those used by the language distinguished by a yellow background.

In the variant-assessment tool which Dr Kulkarni exhibited, possible strings are considered for variants within the subset of characters used for a particular language. However, he agreed that for registration purposes this is of little account: a Hindi-speaker (or anyone) could perfectly well register a character-string that was valid only in Marathi. E.g. the character ळ ḷ is used in Marathi but not in Hindi, therefore कमळ (kamaḷa) would be recognized (and probably registrable) as a Marathi label, not as a Hindi one.

The Indian practice was to exclude registrations of strings which were not possible words in some language. Therefore the subscript dot (*nuqtā*) is only permitted in conjunction with the restricted set of characters that allow it: क़ ख़ ग़ ज़ ड़ ढ़ फ़ य़. And there is no question of allowing, e.g., the string of Unicode codes which [through a transformation I didn't check - NDMO] would correspond to "Google" in the Devanagari would not be registrable: the first item would be a character which can only occur non-initially.

Strings of consonants without intervening vowels (so-called conjunct consonants) can get quite long (up to 5 in a row - [e.g. कात्स्न्र्य *kā-rtsnya* 'wholeness' - NDMO]) so all of these combinations have to be assessed for possible homograph variance.

*

Dr Govind finished with some speculations about problems affecting future (cross-lingual?) variants, when it might be necessary to find Indian equivalents for gTLD's such as .com. Deriving it from 'com(merce)' he said the best Hindi version might be व्यापार *vyāpār*, rather than वणिज् *vaṇij*. But in Tamil, the choice might better fall on வணிகம் *vaṇigam*. He suggested, therefore, that there might be scope for conflict over variants in the future (i.e. conflict based on languages within India rather than just scripts) but did not expatiate.