Computerised writing for small languages

Pat Hall, Language Technology Kendra, Patan, Nepal

ABSTRACT

If knowledge is to be shared on the Internet between members of a linguistic community, then their language needs to be written and the writing encoded for the computer. The problem of achieving this for small languages is illustrated with the case study of Nepal. Nepal has over 120 languages, with only the national language Nepali having any modern computer support. Nepali is relatively easy, since it is written in Devanagari which is also used for Hindi and other Indian languages, though with some local differences. I focus on one particular language, that of the Newar people, which has a mature written tradition spanning more than one thousand years, with several different styles of writing, and which, as yet, has no encoding of its writing within Unicode. I also look at the many unwritten languages of Nepal and the frustrated aspirations of their speakers. I explore why this happens, looking for answers in the standardisation processes and in the different and competing interests and incentives of the people involved. Finally I suggest what small linguistic communities around the world can do to access information and knowledge using their own language.

KEYWORDS

Small languages, Unicode, standardisation, computers, writing, script, Nepal, Newar.

1. Introduction

If we are going to make knowledge available to everybody in their own languages, then those languages must be written and the writing encoded for the computer. There are 7106 living languages worldwide according to Ethnologue (Lewis, Simons and Fennig 2013). Developed at SIL (Lewis and Simons 2010), it is the first edition of Ethnologue to assess the vitality of languages using a system of 13 levels (EGIDS): 0 to 10, with two levels subdivided in two. At levels 0 to 4, languages are actively written and deemed not in danger. These levels currently comprise only 215 languages, at 82.7% of the world's population. Levels 6b to 10 are considered endangered, with the language not being passed on to the next generation: this amounts to 2,434 languages, 34.3% of the total, but only 1.1% of the world's population. Levels 5 and 6a contain more than half of the world's languages and lie at the boundary between the two:

> 5: The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.
> 22% of languages, 9.8% of the world's population.
> 6a: The language is used for face-to-face communication by all generations and the situation is sustainable.
> 36% of languages, 6.4% of the world's population.

If assessed at vitality levels 1 to 5, the language is written. While Ethnologue's original criterion for being written was the Bible, this does not actually indicate that the language is, in any real sense, written. In recent years the view of what constitutes active written use now includes

use of the written language in newspapers and magazines, the availability of dictionaries and grammars, and the written language taught in schools. Ethnologue's EGIDS levels can provide us with useful guidance.

It may seem that all level 5 languages and above are well served; to be written they must have an orthography, and today that must mean some form of computer support. But are they? It may be that a language is written in a script that is not supported by computers, or in an imposed script which has not fully respected the phonology of the spoken language and is disliked by the linguistic community. We will see examples of all of these later in this paper.

Level 5 languages and above may need some investment in their writing, but this is even more true for level 6a languages. In deciding in which languages to invest to help them become fully written, an aid agency or similar may need some other criterion, and may look to speaker population size. How many people would be helped? The median size for 6a languages is 10,000, which gives us an alternative simple criterion based on mother-tongue speaker population.

The discussion above suggests that if knowledge is to be made widely available across the Internet, we need computer support for languages of at least 10,000 speakers and/or with status 6a or better. I will apply this to Nepal in section 2, showing how few Nepalese languages are actually actively written, and then in section 3 look at one particular language, Nepal Bhasa, with a written tradition spanning more than a thousand years but which does not yet have its method of writing encoded for the computer, though it is moving slowly in this direction. In Section 4 I look at the encoding of other unwritten languages in Nepal and what can be done for them.

The objective of this paper is to consider what can be done for all marginalised languages, learning from the experience of Nepal. This is picked up in Section 5, which looks both at how to use the existing standardisation mechanisms of Unicode and ISO, and how the processes of Unicode and ISO might be reformed.

2. Nepal case study background

Nepal lies on the border between the high plateaus of central Asia to the north and the low-lying plains of India to the south. Linguistic communities have migrated into Nepal from both directions; those from the north brought with them languages of the Tibeto-Burman family, while communities migrating from the south brought with them languages of the Indo-Aryan family.

2.1 Nepalese languages

Once in Nepal, communities remained completely isolated by steep valleys, high mountains and thick forests, leading to the evolution of many distinct languages, given as 70 (Toba 1992) when I first stayed in Nepal in 1997, then 92 in the 2001 census, but now put by Ethnologue at 120, though this increase in number seems mostly related to distinguishing dialects within larger groups. The 2011 census of Nepal which allowed citizens to name their own language reported 123 languages, and included several dialects of Nepali but no distinctions between different forms of other major languages. The research community's knowledge of the linguistic landscape has evolved over the years and is continuing to evolve through the ongoing Linguistic Survey of Nepal.

Ethnologue's total of 120 languages includes eight languages that have no recorded speakers in Nepal, either dormant or extinct, (EGIDS 9 or 10). Of the remaining 112 languages, 17 languages have much larger speaker populations in neighbouring countries – the Indo-Aryan languages Maithili, Bhojpuri, Avadhi, Urdu, Hindi, Bengali, Marwari, Angika and Kisan, the Tibeto-Burman languages Tibetan, Mechi (Bodo), Lepcha and Byansi, the Austro-Asiatic languages Santhali, Mundari, Kharia, and the Dravidian language Kurux. If we remove these languages as potentially well served by the neighbouring countries, this leaves 95 languages that are predominantly Nepalese. Using the statistics given in the descriptions Nepal's languages in the 17th edition of Ethnologue (Lewis, Simons and Fennig 2013), I have compiled the table shown in Table 1, arranging the languages in descending order of size.

Colleagues in Nepal and I are keen to enable all these languages to be used on computers and the Internet, but if we only had limited resources to invest, which of these 95 languages should we choose? There are 40 languages with more than 10,000 speakers and 43 languages with EGIDS status of 6a or better, though these only partially overlap. Surprisingly some languages with large populations such as Magar (770,000) and Gurung (352,000) are marked as endangered at 6b, while some languages of very few speakers such as Helambu Sherpa (3990) and Koi (2640) are marked as developing 5.

Language	speakers	ST	CL	rank
Nepali	11,100,000	1	IA	1
Tamang, Eastern	1,180,000	4	TB	2
Newar	825,000	3	TB	3
Tharu, Dangaura	500,000	5	IA	4
Magar, Eastern	462,000	6b	TB	5
Bantawa	371,000	6b	TB	6
Tharu, Rana	336,000	5	IA	7
Limbu	334,000	5	TB	8
Tamang, Western	323,000	5	TB	9
Magar, Western	308,000	6b	TB	10
Tharu, Chitwania	285,000	5	IA	11
Tharu, Kochila	258,000	5	IA	12
Dotyali	250,000	4	IA	13
Gurung, Eastern	227,000	6b	TB	14
Rajbanshi	130,000	5	IA	15
Gurung, Western	125,000	6b	TB	16
Sherpa	122,000	5	TB	17
Tamang, Southwestern	109,000	6a	TB	18
Tharu, Kathariya	106,000	6a	IA	19
Tamang, Northwestern	55,000	6a	TB	20
Jumli	40,000	6a	IA	21
Chepang	36,800	6b	TB	22
Danuwar	31,800	7	IA	23
Thulung	30,000	5	TB	24
Sunwar	26,611	6b	TB	25
Kham, Western Parbate	24,500	5	TB	26
Thangmi	24,200	6b	TB	27
Kayort	22,000	6a	IA	28
Majhi	21,800	6b	IA	29
Ghale, Southern	21,500	6a	TB	30
Kham, Sheshi	20,000	6b	TB	31
Kulung	18,700	6b	TB	32
Khaling	18,000	5	TB	33
Dhimal	17,300	6b	TB	34
Sonha	14,700	7	IA	35
Yakkha	14,600	6b	TB	36
Kham, Gamale	13,100	6a	TB	37
Bahing	12,600	6a	TB	38
Chamling	12,100	6b	TB	39
Darai	10,200	6b	IA	40
Dolpo	8,000	5	TB	41
Kham, Eastern Parbate	7,500	6b	TB	42
Loke	7,500	6a	TB	43
Jirel	7,070	6b	TB	44
Kumhali	6,530	7	IA	45
Mugom	6,500	6a	TB	46
Thakali	6,440	7	TB	47
Sampang	6,000	6b	TB	48
Lhomi	5,660	5	TB	49
Puma	5,000	6b	TB	50
Lohorung	4,970	6b	TB	51
Kyirong	4,790	6a	TB	52
Tsum	4,790	6a	TB	53
Wambule	4,470	5	TB	54
Ghale, Northern	4,440	6b	TB	55
Humla	4,000	6a	TB	56
Helambu Sherpa	3,990	5	TB	57
Tamang, Eastern Gorkha	3,980	6a	TB	58
Manangba	3,740	6b	TB	59
Nachering	3,550	7	TB	60
Belhariya	3,500	6b	TB	61
Chhintang	3,500	6b	TB	62
Bote	2,820	6b	IA	63
Koi	2,640	5	TB	64
Dumi	2,500	7	TB	65
Yamphu, Southern	2,500	6b	TB	66
Tichurong	2,420	6a	TB	67
Raji	2,410	6b	TB	68
Athpariya	2,000	6b	TB	69
Chantyal	2,000	6b	TB	70
Jerung	2,000	6b	TB	71
Kaike	2,000	6a	TB	72
Nubri	2,000	6a	TB	73
Thudam	1,800	6a	TB	74
Wayu	1,740	7	TB	75
Yamphu	1,720	6b	TB	76
Kagate	1,500	6a	TB	77
Mugali	1,500	7	TB	78
Walungge	1,500	6b	TB	79
Chhulung	1,310	7	TB	80
Kuke	1,300	6a	TB	81
Lumba-Yakkha	1,200	6b	TB	82
Raute	830	6b	TB	83
Naaba	770	6a	TB	84
Seke	700	6a	TB	85
Nar Phu	600	6a	TB	86
Tilung	310	8a	TB	87
Dungmali	220	6b	TB	88
Chukwa	100	8a	TB	89
Lingkhim	97	?	TB	90
Baram	50	8b	TB	91
Musasa	50	6b	IA	92
Saam	23	8b	TB	93
Kusunda	7	8b	LI	94
Bhujel	3	7	TB	95

Table 1. The 95 Nepalese languages

ST – status EGIDS level; CL – language family TB=Tibeto-Burman, IA=Indo-Aryan.

Do we invest in the endangered languages with large populations, or the more vigorous small languages? It depends, of course, on what we wish to achieve. If we want to enable wider communication of government policy and develop knowledge translated from Nepali or English it could be the larger languages, but if it is to enable long-term survival of the languages then maybe we should choose the more vigorous languages however small. But why not choose both?

2.2 Nepalese languages with written traditions

While most of the 17 cross-border languages have strong written traditions, only three of the 95 primarily Nepalese languages have any tradition of being written:

- Nepali, historically known as Khas, Parbatiya and Gorkhali, with 11,826,953 mother-tongue speakers in 2011, has been written for around 300 years. It is now the official language of the nation, and spoken by most of the population.
- Newar, known as Nepal Bhasa within the linguistic community, and as Newari[1] in much writing in English about the language, has 846,557 mother-tongue speakers, has been written for over a thousand years in

a number of scripts, and is still actively used by a community rightly proud of its history.
- Limbu with 343,303 mother-tongue speakers, has a few historical sacred texts dating back a few hundred years, but no real written tradition, though it is now used by language activists.

Nepali is written in Devanagari, with this enshrined in law. In 1999 Allen Tuladhar submitted a proposal[2] to the ISO SC2/WG2 committee for a distinct encoding for the writing of Nepali, to include three common consonant compounds (conjuncts) — [tra], [ksha], and [gya] — that collate (sort) separately and are taught as part of the basic alphabet in Nepal. This proposal was rejected (Hugh McG Ross 1999) mostly because the conjuncts were encoded separately in other South Asian scripts, treated as conjuncts with the collation differences handled through collation algorithms. The rejection seemed justified, and software was then developed for the Nepali language working with the Unicode Devanagari code block (e.g. Bal, Gurung and Hall 2006).

While Ethnologue records Nepali at EGIDS level 1, it only reports Newar at level 3 and Limbu at level 5. This is not surprising because successive Nepalese governments from the late 18th century until 1990 suppressed all languages other than Nepali. While the writing of Limbu was probably only ever used for special cultural and religious texts, Newar writing was used for a wide range of purposes until the overthrow of the regime by the Gorkhas in the mid-18th century. Cross-border languages, particularly Maithili and Bhojpuri, also have their own mature literature with their own distinctive scripts — Mithilaksha or Tirhuta, and Kaithi respectively. Ethnologue claims that most languages of Nepal are written, mostly in Devanagari, but while a field linguist may have adapted Devanagari to write the language for the purpose of study and possibly a follow-on translation of the Bible, use of the writing is unlikely to have been much more than this until recently.

Indic writing in the Devanagari and Bengali scripts has been printed in movable type since the early 1800s, with the type evolving and being simplified over the centuries (see, for example, Fiona Ross 1999). When computers became used for writing and publishing, the encoding of Devanagari and other Indic scripts was undertaken in India, leading to the *Indian Script Code for Information Interchange – ISCII*. (BIS 1991). Devanagari was planned for inclusion in ISO 8859 as part 12 (Czyborra 1998), with the expectation that ISCII's codes would be adopted. However, ISO 8859 was superseded by Unicode, which included code blocks for Devanagari and other major Indic scripts from the start, adapted from the 1988 version of ISCII. While in ISCII all the scripts of India had been unified within a single table with the different scripts selected by appropriate font, in Unicode these scripts were disunified into separate code blocks.

The encoding of Limbu was added to the Unicode Standard in April 2003. Limbu has a traditional script, Sirijanga, which is claimed to have been invented in the 9th century, revived in the 17th century by Te-ongsi Sirijonga, and then again in 1925 when the script was formally named "Sirijanga." Limbu standardisation began in 1999 by McGowan and Everson, followed by a proposal by Everson and Michailovsky in 2002. Michailovsky had done considerable field research among the Limbu, learning about their writing in context. In 2011 Pandey proposed two additional composite characters, [tra] and [gya]; however, these could be viewed as conjuncts with the existing [sa-i] interpreted as a virama[3].

3. Nepal Basha and Nepal Lipi

The Newar people had been the rulers of the Kathmandu valley for many centuries before they were conquered by the Gorkhas from a neighbouring Himalayan kingdom to the west in the 1750s. They call the Kathmandu valley '*Nepal*,' their language '*Nepal Bhasa*' and their writing '*Nepal Lipi*' or '*Nepaalalipi*,' respectively the language and writing of the Kathmandu valley.

3.1 The styles of writing Nepal Bhasa

You can see Newar writing carved into stone or wood, or embossed in brass or other metals, in temples around the valley as in Figure 1. There are two distinct *styles*[4] of writing, an ornate style with many long downward strokes called '*Ranjana*,' and a more rounded style '*Prachalit*.' Examples from a modern newspaper are shown in Figure 2.

Figure 1. Newar writing from the Golden Temple in Patan, Kathmandu valley, Nepal (© the author).

Figure 2. Extract from newspaper with Ranjana headline, and Prachalit text.

Rabison Shakya (2002) identifies a third style Bhujimmola, while Hemraj Shakyavansha writing in 1985 had identified nine styles. Different styles appear to have been used for different purposes – Ranjana for sacred and religious texts, Prachalit for everyday secular writings and Bhujimmola for administrative purposes.

As noted by Ethnologue, much current writing of Nepal Bhasa is done in Devanagari, though this is deprecated by the Newar community. Hack fonts[5] are available for both Ranjana and Prachalit (e.g. Shakya 2002), but there is hope for an encoding in Unicode of Nepal Lipi for which high quality open type fonts can then be produced.

3.2 Early attempts to encode Nepaalalipi

In 2000, Everson proposed a code block '*Newari'* illustrated with Ranjana characters, and a code block '*Nepali*' illustrated with Prachalit characters (Everson 2000a, b). These drafts included the same three conjuncts, [tra], [ksha], and [gya], as in the earlier failed proposal for Nepali. While Unicode experts were confused about Nepal's languages and about the way they are written, this confusion was mostly overcome in later proposals formally submitted to the standardisation bodies (e.g. Everson 2009a).

In 2002, Rabison Shakya proposed tables of the basic characters of the writing for Ranjana, Prachalit, and Bhujimmola, but the Prachalit tables had significantly more characters, which intrigued me. There are "extra" vowels of pre-composed vowels with diacritics. To analyze what was going on here, I rearranged the consonants of Shakya's Prachalit table to show the seven 'extra' consonants in the groups or vargs used in South Asia to

organise a script, but also used in arranging phoneme tables. My rearranged table is shown in Figure 3. All of these extra consonants look like combinations of two or more basic consonants, but could they represent something special? Hale and Shrestha (2005) showed that they are phonologically distinct aspirated or breathy consonants, different from a consonant followed by a [ha], and written as if the [ha] came first.

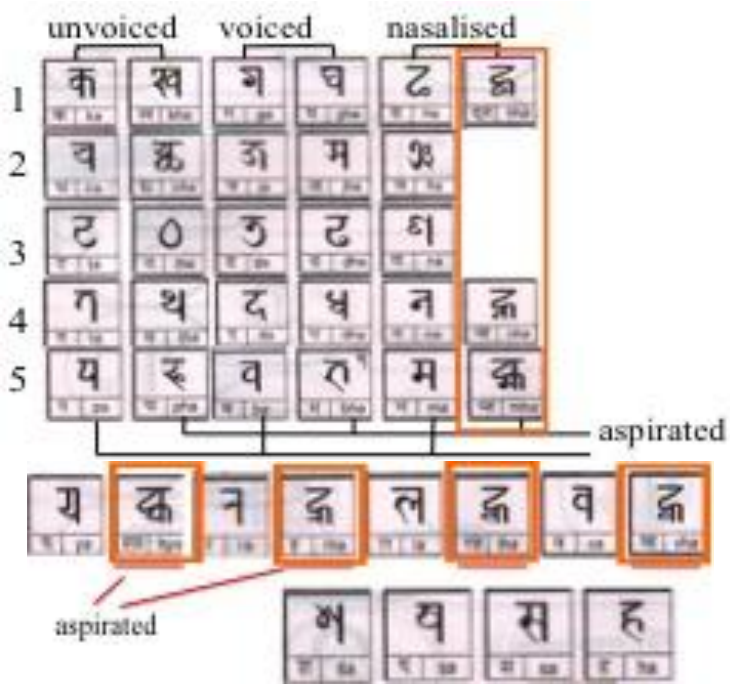

Figure 3. Prachalit in Vargs, highlighting the 'extra' aspirated (breathy) consonants (adapted by the author from Shakya 2002: 7)

Tuladhar (1999) mentions Newar, and in the mid-2000s Newar activists began to take an interest in Unicode standards for Newar writing, with ideas expressed as several proposed code blocks. A meeting of the Nepal Lipi Guthi in July 2008 explored the idea that each style of writing should be separately encoded and that the actual shapes of the characters should themselves be standardised.

In 2009 the Script Encoding Initiative funded Michael Everson to work on Ranjana. Everson (2009a: 1) advocated:

> Since Rañjana is visually and structurally similar to the Lañtsa and Wartu scripts used for Buddhist Sanskrit documents in Tibet (China), Bhutan, Mongolia, Nepal, Sikkim and Ladakh (India) it has been considered [it] would be practical to merge these two scripts (Lañtsa and Wartu) with Rañjana for encoding purposes.

In the email discussion that followed, a very strong lobby of Newars proposed that Prachalit be encoded first, as the style of writing used in the daily life of the Newars. I personally argued that Ranjana and Prachalit were equivalent, just different visual expressions of the same underlying

writing to support the language Nepal Bhasa. After much debate, Everson posted a second document (2009b: 1) which stated:

> **2 *PRACALIT***
> There are many scripts in this group, mostly distinguished by their headlines. The major difference between the PRACALIT scripts and the RAÑJANA scripts is the way in which *-e* and *-ai* are made by changing the top bar (see Figure 20). Two major varieties are distinguished, and there is not yet enough evidence available to determine whether or not it is appropriate to encode them separately from one another

Everson includes a table, his table 20 of characters copied from elsewhere, which contrasts a number of characters and the way they are composed with diacritic vowels in the Ranjana, Prachalit and Bhujimmol styles. Viewed as styles, the design differences between these are very small indeed when compared to the design differences between the many Roman fonts available on laptops. In discussions about scripts within the coding community these differences are characterised as "shaping behavior." However, this is not a term that appears in the Unicode Glossary and it is difficult to see what is meant by it and why it is significant. Everson concluded:

> **Encoding considerations**. It should first be said that some members of the user community have criticized the idea of unifying these 'scripts.' It may be that this is a misunderstanding of the UCS; the analogy of the Latin script with its Gaelic and Fraktur variants, however, is probably applicable, which is why the recommendations here have been made (2009b: 2).

This seemed to wrap it up: all the styles of writing the Newar language would be unified to a single encoding, a view discussed at a meeting of the Nepal Lipi Guthi in March 2010. However I was very wrong.

3.3 2011/13 encoding proposals and debates

In 2011 Anshuman Pandey (2011h) circulated a proposal for Prachalit, and updated this with a proposal for '*Newar*' (Pandey 2012a) which included a very similar code block with a few extra characters. The document included considerable argument supporting the name '*Newar*' for the code block, though this was then contested by many members of the Newar community who preferred '*Nepaalalipi*,' their traditional collective name for their scripts. Deborah Anderson pointed out in correspondence that *Nepaalalipi* could meet with difficulties and be delayed on the grounds that '*Lipi*' translates as 'script' and names of code blocks should not include the word 'script.' In response, the Newar community pointed to the two Japanese syllabaries Hiragana and Katakana with code blocks whose names contain '*gana*' and '*kana*' which also mean 'script.' This offence was further compounded by a well-researched 35 page document from Pandey (2012b) arguing further for

'*Newar*' as the correct choice. In 2014 this name issue was resolved, as discussed later.

Pandey's Newar proposal also discussed a number of "additional consonantal forms," shown in Figure 4, exactly those extra aspirated consonants included in Shakya's Prachalit table (Figure 4). Pandey stated that these should not be separately encoded, but should be viewed as conjuncts which have been written wrongly — in all of these the [ha] is written first at the top of the composite character, but if when pronounced the breathy [ha] follows, then the [ha] should correctly be written at the bottom.

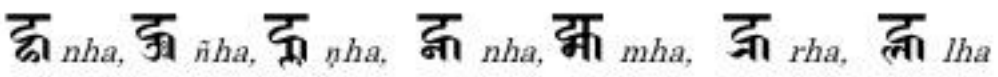

Figure 4. Newar characters — Are these consonants or conjuncts? (adapted by the author from Pandey 2014a)

In 2012 Devdass Manandhar and colleagues submitted an alternative proposal to the Unicode Technical Committee (UTC) describing Newar writing from a Newar perspective including the breathy consonants of Figure 4. At their May 2012 meeting the UTC considered both proposals, (UTC 2012a). After comparing the proposals to focus our discussions in Nepal (Hall 2012a), I found that there was strong agreement on many aspects of the proposals, apart from the breathy consonants and two concerns which are not matters for encoding.

While it is clear that the breathy consonants are present in the spoken language, why are they written in this 'wrong way?' There are contrasting pairs of words with different meanings between the breathy consonant and the consonant followed by [ha]. To avoid ambiguity in the writing, the breathy consonants need to be written differently, and writing them in the reverse order is one way to achieve this. Noonan (2003) reports a similar problem in the writing of Chantyal using Devanagari. The Nepal Lipi Guthi discussed the need for new glyphs to be designed for these characters, and was assured that if these glyphs were used in a number of documents, the case for them would be much stronger, however this is no longer necessary. A very detailed discussion of the issues involved has been given by Whistler (2014a) in the context of the later discussions described below.

Discussions continued throughout 2012 and 2013. I started drafting a composite proposal from the Pandey and Manandhar proposals (Hall 2012b) and distributed it expecting a positive response, but obtained very little feedback and the idea died; it was too early for a consolidated proposal. However Manandhar modified his proposal and issued 2 redrafts (2012b, 2013). The UTC discussed these proposals respectively at the November 2012 and May 2014 meetings.

In late 2013 Sinclair posted a letter in which he declared his purpose was to:

> support […] the sound proposal for Newar by Anshuman Pandey, ISO/IEC JTC1/SC2/WG2 N4184 L2/12-003, and in order to point out, as briefly as possible, fundamental flaws in the related proposals N4322 (Dev Dass Manandhar *et al*, 'Nepālalipi') (2013: 1).

With, sadly, no sense of looking for their relative merits in the interest of a rapprochement.

From early 2012 through to 2014, Pandey announced many trips to Nepal in order to meet with and consult the Newar community; each time a sequence of meetings in Kathmandu was prepared for the visits, but Pandey has never been able to make that trip for a combination of personal and bureaucratic reasons.

3.4 2014 – Enter the Unicode Technical Committee (UTC)

In early 2014 the Script Encoding Initiative at Berkeley obtained funds for a trip to Nepal to consult the linguistic community there. A meeting was fixed for October 4th to 7th 2014, which fell in the middle of the Dashain holiday period, an important time for families to meet and celebrate. The trip by the UTC's Deborah Anderson and Peter Constable was nevertheless welcomed by people in Nepal. The meetings were a great success, attended by many leading Newars, which attested the importance that the community places on the encoding of their writing. Anderson (2014) wrote a comprehensive report of that meeting, documenting the concerns and preferences of the active users of Newar writing.

Anderson's report led to a debate by members of the UTC, particularly about the inclusion of the breathy consonants as atomic characters. Pandey (2014a) defended his position, offering just one example from a website constructed in 2009. Whistler (2014a) gave a well-argued and balanced evaluation of the proposals for and against atomic encodings. On balance it favoured atomic encodings for the breathy consonants of Nepal Bhasa, taking into account the relationship between the use of the script for writing Nepal Bhasa and for writing Sanskrit.

Whistler (2014b), with support from the Newar group in Kathmandu, then wrote a proposal to the UTC for a code block called "Newa" which included the contentious breathy consonants. The proposal was in turn approved by the UTC on 30th October 2014 for inclusion in Unicode 9.0. Slots for both Newar and Ranjana appeared in the forward planning Roadmap of Unicode (Everson *et al* 2012).

Still not fully resolved is whether Ranjana and Prachalit are sufficiently different to warrant separated encoding, and what the status of

Bhujimmola is. Is it more important that Ranjana and Prachalit are both ways of writing the same language Nepal Bhasa, or that Ranjana is visually similar to the Lañtsa and Wartu writing in Bhutan and Tibet? The position from Unicode is clear: usage by particular languages should not be taken into account, but visual issues need to be.

The position of the designer of usable systems by people from within a single linguistic community, however, is quite different, as we will see. Discussions are already under way to prepare a font for Newa (Prachalit/Nepaalalipi) in anticipation of the standard. Once the font is produced, should they then move on to produce fonts for Ranjana and Bhujimmola styles of writing that work to the same set of codes, to the benefit of writers of Nepal Bhasa?

4. Other Nepalese languages and their encoding

When the Indian constitution first scheduled its official languages, Maithili was viewed as a dialect of Hindi. This decision was vigorously contested and eventually led to the inclusion in 2004 of Maithili as a distinct scheduled language, written in Devanagari. The traditional style of writing, known as Mithilaksha or Tirhuta, was treated as an exotic script for use in wedding invitations and similar events, though there were discussions on whether or not it could be unified with Bangla or Devanagari.

In 2008 a font for Tirhuta called Janaki was produced in Nepal, mapped to the Devanagari code block, with the advantage that existing documents encoded in Devanagari could be rendered in Tirhuta by a simple change of font. Then in 2011 Pandey proposed a separate encoding of Tirhuta, arguing briefly that it could not be unified with Bengali, but not discussing the situation with respect to Devanagari. Tirhuta is now part of the Unicode standard. Similarly, the Kaithi script used historically across north India and notably for Bhojpuri was proposed in 2007 by Pandey and accepted into Unicode in 2009; while visual similarities with Bengali writing were noted, unification with Bengali was not considered.

Field linguists in Nepal aiming to document the languages they study have always improvised a means of writing the languages, usually based on Devanagari (Noonan 2003). Regmi (2008) has proposed a writing system based on Devanagari for all the languages of Nepal, with this attracting some interest in government circles (Regmi *et al* 2012). But for most linguistic communities Devanagari is seen as the writing of historical oppression, and they prefer some other distinctive writing system.

Activists for some of these languages have created their own distinctive writing, with proposals that have reached discussion on standardisation (Anderson 2012) – these proposals have been packaged for Unicode by Pandey - (2011a,b,i) for Sunawar, (2012c,g) for Bantawa, (2012d) for

Gurung, (2012f) for Magar, and (2012j) for Dhimal. Part of the drive for writing these languages comes from Sikkim in India where they are also spoken, with the official newspaper The Sikkim Herald published in 11 languages with distinctive scripts and typography, as seen in Figure 5.

Figure 5. The Sikkim Herald in 11 languages (with permission from Mark Turin).

The pursuit of visually distinctive writing is one way of marking the writers' identity (Sebba 2009), but this can be achieved by distinctive fonts, and does not require separate encodings. This observation led me to propose unifying these proposals with Nepal Bhasa, encouraged by members of the Newar community (Hall 2012c), to cover all the Tibeto-Burman languages of Nepal, based on phonological arguments. The proposal was discussed at UTC meeting 133 (2012b) and rejected, with a person tasked to inform me with reasons - although I never received anything. Two negative comments were received, one by personal email from Michael Everson and the other more publicly by Sinclair (2013) who declared it "nightmarish."

Unwritten languages are clearly of no interest to Unicode. The UNESCO guide to writing unwritten languages (Robinson and Gadelii, 2003 — see also Cahill and Karan, 2008) recommends basing the writing on some dominant alphabetical writing of the region, which is exactly what Noonan has documented and Regmi proposed in their focus on Devanagari. The acceptance of the Newa proposal with features appropriate for Himalayish languages opens up the possibility of basing their writing on Newa.

5. How can a small language join the information society?

A language represented in the computer is able to share information in that language through modern communication technologies, possibly with translations to and from other languages, particularly the international languages English, Chinese and Spanish.

5.1 Problems with Unicode processes

From the case study of Nepal it is apparent that working with Unicode may not always be easy, but if the writing of a language is to fully join the information society, it must use encodings that are part of the Unicode standard. So what are the strengths and weaknesses of Unicode?

Unicode encodes writing and not languages
Unicode does not accept phonological arguments except in tightly constrained contexts about particular characters; there is absolutely no interest in unwritten languages. Examples of writing are looked at and deemed related if they look similar, unrelated if they look different, disregarding evidence concerning what language is being written. This was evident in the proposal to unify Ranjana with Lantsu and Wartu based on visual similarity rather than with Prachalit based on language. As a consequence of this logic, it appears (to me at least) that historical writing of dead texts takes precedence over living languages.

Coding starts with activists who may have limited expertise
Coding starts with activists who write the initial proposal and collect the evidence. Inevitably these activists have a narrow focus, knowing a lot about their own culture but little about the other areas important for coding. In Nepal we initially were championed by a font developer who clearly had a limited understanding of writing systems though was very adept politically; more recently, it has been a very able historian who appears to have a limited understanding of language and technology. Regrettably these activists do not know what they do not know, but have an undue influence.

Bias is towards separation not unification of codes
The practice in standardisation was originally to unify, as is seen in the Roman writing captured initially in ASCII (ISO646) and now in the Latin code blocks of Unicode. Different European languages use significantly different character sets and rules for diacritics and ligatures, but early codes like ASCII had limited flexibility, unifying all European writing within the single set of codes of ISO 646, while all Indic writing was unified within the single set of ISCII codes. Technology like Unicode enables multiple codes to co-exist, so when Indic writing was taken into Unicode the ways of writing different languages were separated, though this was not done for Roman writing. Why not?

It clearly helps in the proposal writing process to have people involved who are experienced and know what needs to be done, and the writers of a large number of proposals are rewarded by acclaim within the coding community, with one even getting a special write-up in the New York Times. While their contributions are much appreciated, unfortunately all this has the perverse consequence that the more scripts they can successfully encode, the higher the rewards, and instead of seeking to unify scripts, they are incentivized to see differences and encode scripts separately.

In need of reform

We must not lose sight of why we need codes, i.e. so that we can build systems that serve people in their own languages and writing. It is not to provide a convenient database of characters for people who study ancient scripts. One important principle of software engineering is the "separation of concerns", as in, for example, the separation of form from function which allows one to focus on what the document says, and only later to format it, by inserting such secondary characteristics like bold, italics, and size and choice of font.

In this discussion on Nepal Bhasa writing styles, one sees emerge a possible proliferation of encodings for scripts which are essentially the same. If this proliferation continues we will end up with a situation not unlike the one encountered in Asia with hack fonts and hack encodings. People would need to possess the same font in order to share documents across the Internet. While it had seemed that Unicode could save South Asian languages from hack fonts, it now seems likely to perpetuate the same situation with hack Unicodings. Newars using a font in the Newa code block cannot communicate with Newars using a font in the (proposed) Bhujjimmola code block.

Prior to Unicode it was common to standardise encodings and keyboards together as seen in ISCII (BIS 1991), but Unicode has departed from this practice. Separation for the purposes of standardisation is beneficial. But in other ways Unicode has brought together very many standards into a monolith. The Unicode standard consists of three parts: the core specification and code charts which assign characters to code points (which we have focused on), various annexes, and the Unicode Character database (UCD). Whistler and Freytag (2008) observe about the UCD:

> Other character set standards leave it to the implementer, or to unrelated secondary standards, to assign character semantics to characters. In contrast, the Unicode Standard supplies a rich set of character attributes, called properties, for each character contained in it. (section 2.1)

The Unicode character set is already a combination of many standards; adding the character properties simply compounds that situation. Has

Unicode gone too far? Is it time to return to many smaller though inter-related standards? Modular software is best practice in software development. Should we also be looking to a best practice in modular standardisation?

Until around 1980, the encoding of scripts for computer systems was separate from the encoding of scripts for library archival and antiquarian work. This led to duplication of effort and the committees were combined. Perhaps what I have documented above is an unintended consequence of that merger, and some separation may need to be reinstated.

5.2 Computerising your language

If you are a user of a small language not yet computerised, what are you to do?

Converting handwriting to computer writing

When writing moved from pen to keyboard, the secondary aspects of the writing had to be made explicit. In Roman handwriting, the selection of small or capital letters was implicit, but with typewriters this became explicit using the shift key. Roman writing for computers was based on manual typewriters with the input keys, internal computer codes, and output characters in one-to-one correspondence. Today we can use key-mapping software to make any key press or a combination and sequence of key presses to produce a particular internal code. With open-type fonts, the character can be rendered on a screen or in print dependent upon a sequence of internal codes.

In most Brahmi-derived Asian handwriting, all consonants have an implicit 'a' sound, which can be overridden by the diacritic for a different vowel. The diacritic virama suppresses the implicit 'a' at the end of a word. If a cluster of consonants follow each other without vowels in between, a special conjunct ligature can be made. There are too many conjuncts to assign each of these a separate key on a keyboard, so we need to distinguish between consonants that are part of a conjunct, from consonants which do not form part of a conjunct. One possible solution is that the vowel should always be explicit, but this does not appear to have been considered historically; instead, to signal a cluster either the implicit vowel must be suppressed with a special character like a shift, or a special cluster form of the consonant must be chosen. Most Brahmi scripts, such as Devanagari, use the virama to signal a conjunct, but some writing encodes two forms of consonants, a stand-alone form and a second 'sub-joined' conjoining form.

So an important step in moving from handwriting to computer writing must be to decide how computer writing will be achieved, giving the principles for the layout of characters on the keyboard and defining the

sequences in which these must be typed. Design should at least outline the keyboard at the same time it determines the repertoire of characters to be encoded.

Alternatively it might be best to use some other input method, like stroke sequence recognition as used in handwriting on touch screens.

Already established writing

My advice would be to go for Unicode, but with circumspection. Gather lots of examples of the writing, then write a proposal to the Unicode Technical Committee and ISO, exhibiting and explaining the basic character set. Model the proposal on some existing successful proposal such as that of Pandey (2011e) for Tirhuta, avoiding local theories of writing and language. Find a champion within the UTC/ISO community, but beware of the partial expert.

If this does not succeed, or as an interim measure, go for a local national standard, choosing a Unicode compatible encoding in the Private Use Area to enable you to use all the technology available for Unicode. Note that there are drawbacks to this, because every person in the user-community needs to know your encodings, if only indirectly through a font.

Unwritten languages

Follow the advice of UNESCO in Robinson and Gadelii (2003), making it phonetic, if possible exploiting some existing locally established writing, but adjusting it for your specific language with new characters as needed. The linguistic community may well want to make it visually distinct even though they are covertly following the same underlying principles. However I would advise avoiding the Indic model with its complex rules, preferring something truly alphabetic like the Roman system with its simple rules, or the sophisticated Hangul system.

Once the writing has been created, ensure that it is actively used, for example in schools and the media, and then proceed to Unicode as in the section above.

An alternative approach to unwritten languages would be to look for a picture based language, inspired by emoji characters and the Chinese writing system, and by the iconography developed by John Roscoe (2013).

Stay with speech

With the advances in speech technologies and the decreasing cost of such technology, it may be just as easy to use speech, as is done daily with mobile phones. If speech is accompanied by pictures it can communicate richly, as described by Nyiri in 2003.

However, part of the motivation for computerising a small language would be to access knowledge using translation paths from international languages. Researchers are developing speech-to-speech translation, but at the moment this is based on writing-to-writing translation with speech recognition (ASR) at the front end and speech generation (TTS) at the back end. I do not know of any attempts to create direct speech-to-speech systems, so it seems we cannot escape from writing. But maybe the "writing" of our small language need not be intelligible to humans, just to the machine for the purposes of translation. More research is needed.

6. Where do we go from here?

While I recognise the importance of standardisation to enable interworking, we are left with serious concerns about the standardisation process as a means of meeting the needs of small communities. There is a deep social injustice in the current situation that favours a few hundred scholars of an extinct writing over the interest of hundreds of thousands in a community of living users of a language. We need look no further than the travails of Nepal Bhasa for an example of this.

We need to reform the coding standards. Perhaps it is time to revert to the former division between encodings for software systems and living users, and for libraries and antiquarian users, and to unbundle the Unicode collection of standards into interrelated modular standards.
Meanwhile small linguistic communities should move ahead as best they can following the guidance of section 5.2.

Bibliography

- **Anderson, Deborah** (2012). Liaison Report, Script Encoding Initiative, UC Berkeley. Document ISO/IEC JTC1/SC2/WG2 N4220 2012-02-12.
- **—** (2014a). *Comparison between Newar and Nepaalalipi proposals (L2/12-003 and L2/14-086)*. http://www.unicode.org/L2 /L2014/14220-newar-nepaalalipi-compare.pdf. (consulted 24/09/2014).
- **—** (2014c) Recommendations to UTC from Script Meeting in Nepal Unicode Technical Committee document L2/14-253 6 October 2014 http://www.unicode.org/L2/L2014/14253-rec.pdf (consulted 24/09/2014).
- **Bal, Bal Krishna; Srishtee Gurung and Pat Hall** (2006) "Towards Universal Access to ICTs in Nepal." Paper presented at *Computer Society of India conference*. (Kolkata, India, November 2006).
- **BIS** (1991) *Indian Standard Indian Script Code for Information Interchange – ISCII*. IS 12194: 1991. Bureau of Indian Standards, New Delhi.
- **Cahill, Michael and Elke Karan** (2008) *Factors in Designing Effective Orthographies for Unwritten Languages,* SIL Electronic Working Papers 2008-001, February 2008.

- **Czyborra, Roman** (1998) *The ISO 8859 Alphabet Soup*. http://czyborra.com/charsets/iso8859.html (consulted 20.01.2015).

- **Everson, Michael** (2000a) *Newari code table and names list* http://www.evertype.com/standards/tai/newari.pdf (consulted 15.2.2012).

- **—** (2000b) *Nepali* http://www.evertype.com/standards/tai/nepali.pdf (consulted 20/1/2015 and changed type extension remove '.html').

- **—** (2009a) *Preliminary proposal for encoding the Rañjana script in the SMP of the UCS*, International Organisation for Standardisation document ISO/IEC JTC1/SC2/WG2 N3649 http://www.evertype.com/formal.html or http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3649.pdf (consulted 15.02.2012).

- **—** (2009b) *Roadmapping the scripts of Nepal.* International Organisation for Standardisation document ISO/IEC JTC1/SC2/WG2 N3692, http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3692.pdf or http://www.evertype.com/formal.html (consulted 2012/02/15).

- **Everson, Michael, Rick McGowan, Ken Whistler, and V.S. UMAmaheswaran** (2012) *Snapshot of Pictorial view of Roadmaps to BMP, SMP, SIP, TIP and SSP* Document ISO/IEC JTC 1/SC 2/WG 2 N4186 2012-02-16 http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4186.pdf (consulted 10.05.2015).

- **Hale, Austin and Kedar P. Shrestha** (2005) *Newar*. Volume 256 in Languages of the World/Ma9terials. Munich and Newcastle: Lincom Europa.

- **Hall, Pat (**2012a) Comparison of coding proposals for Nepal Lipi 2012-3-3 Privately distributed.

- — (2012b) Proposal to Encode the Newar Script in ISO/IEC 10646 2012-11-29 unpublished working draft.

- **—** (2012c) *Proposal to Encode Nepal Himalayish Scripts in ISO/IEC 10646*, ISO/IEC JTC1/SC2/WG2 N4347 http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4347.pdf (consulted 11.10.2013).

- **Hall, Pat, Bal Krishna Bal, Sagun Dhakhwa, and Bhim Narayan Regmi** (2014) Issues in Encoding the Writing of Nepal's Languages. pp CICLING conference 2014, pro pp 52-67 in A. Gelbukh (Ed.): *Computational Linguistics and Intelligent Text Processing*, proceedings of CICLing 2014, Part I, LNCS 8403. New York: Springer.

- **Lewis, M. Paul and Gary F. Simons** (2010) Assessing Endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistique (*RRL), Vol. LV, No. 2. pp 103-120. http://www.lingv.ro/RRL-2010.html (consulted 08.10.2013).

- **Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig** (eds.) (2013) *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com; specific page on Nepal, http://www.ethnologue.com/country/NP/languages (consulted 14.05.2015).

- **Manandhar, Devdass, Samir Karmacharya, Bishnu Chitrakar** (2012) *Proposal for the Nepaalalipi script in the UCS*, 2012-02-05, ISO/IEC JTC1/SC2/WG2 N4322, International Organization for Standardization. http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4322.pdf (consulted 14.05.2015).

- **—** (2012b). *Proposal for the Nepaalalipi script in ISO/IEC 10646.* UTC document L2/12-349 2013-12-31.

- — (2013). *Proposal to encode Ranjana Script in ISO/IEC 10646* Unicode Technical Committee document L2/13-243 31 Dec 2013 http://www.unicode.org/L2/L2013/13243-ranjana.pdf (consulted 14.05.2015).

- **McGowan, Rick and Michael Everson** (1999). Unicode Technical Report #3: Early Aramaic, Balti, Kirat (Limbu), Manipuri (Meitei), and Tai Lü scripts ISO/IEC JTC1/SC2/WG2 N204 1999-07-20.

- **Michailovsky, Boyd and Michael Everson** (2002). *Revised proposal to encode the Limbu script in the UCS.* Document ISO/IEC JTC1/SC2/WG2 N2410, International Organization for Standardization. http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2410.pdf (consulted 14.05.2015).

- **Noonan, Michael** (2003). *Recent Adaptations of the Devanagari Script for the Tibeto-Burman Languages of Nepal*. http://archiv.ub.uni-heidelberg.de/savifadok/202/1/Recent_Adaptions_of_Devanagari_Script.pdf (consulted 01.05.15).

- **Nyiri, Kristof** (2003). "Pictorial Meaning and Mobile Communication" [English translation of (2002f)], in Kristóf Nyíri (ed.). *Mobile Communication: Essays on Cognition and Community*. Vienna: Passagen Verlag, 2003, 157–184.

- **Pandey, Anshuman** (2011a). *Preliminary Proposal to Encode the Jenticha Script in ISO/IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N3962 2011-01-25 http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3962.pdf (consulted 10.2.2012).

- — (2011b). *Preliminary Proposal to Encode the Tikamuli Script in ISO/IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N3963 2011-01-25, http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3963.pdf (consulted 10.02.2012).

- — (2011c). *Introducing the Khambu Rai Script*. Document ISO/IEC JTC1/SC2/WG2 N4018 2011-04-13, http://www.anshumanpandey.com/ and http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4018.pdf (consulted 10.02.2012).

- — (2011d). *Introducing the Khema Script for Writing Gurung*. Document ISO/IEC JTC1/SC2/WG2 N4019 2011-04-13 http://www.anshumanpandey.com/ and http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4019.pdf (consulted 10.02.2012).

- — (2011e). *Proposal to Encode the Tirhuta Script in ISO/IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N4035 2011-05-01 http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4035.pdf (consulted 10.02.2012).

- — (2011f) *Introducing the Magar Akhar Script*. Document ISO/IEC JTC1/SC2/WG2 N4036 2011-05-01 http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4036.pdf (consulted 10.02.2012).

- — (2011g) *Introducing the Kirat Rai Script*. Document ISO/IEC JTC1/SC2/WG2 N4037 2011-05-01 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4037.pdf (consulted 10.02.2012).

- — (2011h). *Preliminary Proposal to Encode the Prachalit Nepal Script in ISO.IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N4038 2011-05-03 http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4038.pdf (consulted 10.02.2012).

- — (2011i). *Proposal to Encode the Jenticha Script in ISO/IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N4028 2011-05-31 http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4028.pdf (consulted 10.2.2012).

- — (2011j). *Introducing a Script for Writing Dhimal*. Document ISO/IEC JTC1/SC2/WG2 N4140 2011-09-29, http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4140.pdf (consulted 10.2.2012).

- — (2012a). *Proposal to Encode the Newar Script in ISO.IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N4184 2012-01-05 http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4184.pdf (consulted 10.2.2012).

- — (2012b). *Standard Nomenclature for the Newar Script: Considerations for International Standards*. Private distribution, 26.06.2012.

- — (2014a). *Specimen Showing Representation of Murmured Consonants in the Newar Script Unicode Technical Committee* document L2/14-290 28 Oct 2014 http://www.unicode.org/L2/L2014/14290-newar-murmured.pdf (consulted 14.05.2015).

- — (2014b). Introducing the Bhujinmol Script Unicode Technical Committee document L2/14-283 28 Oct 2014 http://www.unicode.org/L2/L2014/14283-bhujinmol.pdf (consulted 14.05.2015).

- **Regmi, Bhim Narayan** (2008). Developing a Devanagari-based multi-language orthography for Nepalese languages. In *Second International Conference on Language Development, Language Re- vitalization, and Multilingual Education in Ethnolinguistic Communities*, Bangkok, July 1-3.

- **Regmi, Bhim Narayan, Bhim Narayan Regmi, Dan Raj Regmi, Manju Acharya, Hari Narayan Mahato, and Balaram Lamichhane** (2012). *Typological Study of the Languages of Nepal. Report Submitted to Second Higher Education Project University Grant Commission*, Nepal (in Nepali).

- **Robinson, Clinton and Karl Gadelii** (2003). *Writing Unwritten Languages*. UNESCO, http://portal.unesco.org/education/en/ev.php-URL_ID=28300&URL_DO=DO_TOPIC&URL_SECTION=201.html (consulted 15.05.2015).

- **Roscoe, John** (2012). *Iconography: A Protocol for Writing with Pictures*. Stavanger University Press.

- **Ross, Hugh McG** (1999). Comment on Proposal for Nepalese Script. downloaded from http://std.dkuug.dk/JTC1/SC2/WG2/docs/n2055.pdf (consulted 27.01.2015).

- **Ross, Fiona** (1999). *The Printed Bengali Character and its Evolution*. Richmond: Curzon Press.

- **Sebba, Mark** (2009). "Sociolinguistic Approaches to Writing Systems Research." *Writing Systems Research*, 1 (1), 35-49.

- **Shakya, Rabison** (2002). *Alphabet of the Nepalese Script*. Patan, Nepal: Motiraj Shakya and Sanunani Shakya.

- **Sinclair, Iain** (2013). *Letter in support of N4184 and encoding the Newar script in ISO/IEC 10646* ISO/IEC JTC1.SC2 WG2 N4372 2012-10-12.

- **Toba, Sueyoshi** (1992). *Language Issues in Nepal*, Samdan Books and Stationers, PO Box 2199, Kathmandu, Nepal.

- **Tuladhar, Allen** (1999). *Proposal for Encoding Nepalese script in the ISO/IEC 10646*. http://std.dkuug.dk/JTC1/SC2/WG2/docs/n1999.doc (consulted 15.05.2015).

- **Unicode Technical Committee** (2012a). *On the encoding of the "Nepaalalipi" / "Newar" script*, UTC document L2/12-200, 11 May 2012.

- — Unicode Glossary**.** http://www.unicode.org/glossary (consulted 15.05.2015).

- — (2012b). Minutes of UTC Meeting 133. http://www.unicode.org/consortium/utc-minutes/UTC-133-2012011.html (consulted 15.05.2015).

- — (2014). Draft Minutes of UTC Meeting 141, 27th to 30th October 2014. http://www.unicode.org/L2/L2014/14250.htm (consulted 23.01.2015).

- **Whistler, Ken and Asmus Freytag** (2008). The Unicode Character Property Model Unicode Technical Report #23, http://unicode.org/reports/tr23/ (consulted 15.05.2015).

- **Whistler, Ken** (2014a). Rationale for Atomic Encoding of Murmured Resonants in Newa. Unicode Technical Committee document L2/14-281 27 October 2014 http://www.unicode.org/L2/L2014/14281-newa-atomic.txt (consulted 15.05.2015).

- — (2014b) Towards a Consensus Encoding for Newa Unicode Technical Committee document L2/14-285 4. December. http://www.unicode.org/L2/L2014/14285-newa.pdf (consulted 15.05.2015).

Biography

Patrick Hall has alternated between university and industry, concerned with the development of large software systems. He spent three years in Saudi Arabia in 1978-80 working on Arabic software, during 1993/4 led the EU funded Glossasoft project applying computational linguistics to the globalisation of software, and then ran several projects in South Asia, living from 2005 to 2010 in Nepal gathering corpora for Nepali.
email: pavhall@ltk.org.np OR p.a.v.hall@btinternet.com

[1] '*Newari*' is not a word in Nepal Bhasa or Nepali and the Newars feel offended by its use, even in English where it would be a legitimate neologism and has been used for many years. Ethnologue lists this as an alternative name, but in the latest edition as pejorative, which is overstating the position. The preferred term '*Nepal Bhasa*' is also problematic, since this means the language of Nepal, which was true historically when 'Nepal' denoted the Kathmandu valley, but is no longer true today with 'Nepal' denoting the whole country.

[2] This proposal included an Annex not available online, but since I was a co-author, this can be obtained from me on request.

[3] Suggested in private communication by Boyd Michailovsky who acknowledged that it would need to be properly researched.

[4] I searched for a term that would mean writing with a different visual design used for writing the same language, just as European languages might be written in a cursive style or block printing style, in a serif or a sans-serif style, or German which might be written in a normal style or the Gothic style. The various books on writing systems do not make this distinction with a consistent terminology, so I chose 'style' as a term that

captured what I intended but was not part of the official Unicode terminology and not listed in the Unicode Glossary.

[5] Hack fonts and hack encodings: when personal computers emerged during the 1980s they only catered for the major languages and writing of the developed world. Technical experts from minor languages writing in complex scripts like Arabic and Devanagari adapted the existing 8-bit codes and 48-key keyboard for their scripts by deciding which characters in their writing systems would be assigned to which keys. In some cases they would follow an existing typewriter convention, in other cases they would look for phonetic similarities to a language using the Roman system and assign characters accordingly — so for example the Arabic alif and the Devanagari A vowel might be assigned to the A-key. The characters of the alphabet then take the encoding of their assigned key, and a font is constructed using these accidental *hack codes*. If two font developers use different key assignments they end up with different encodings, with the consequence that communicating persons must both use the same font. In Nepal in 1997 there were several such *hack fonts*, all with different key assignments and hence different hack codes, but over time most people have migrated to just one of these, Preeti, driven by the choice made in dominant newspapers, and still in use today (but diminishing).