Problems with Unicode for Languages Unsupported by Computers

Pat Hall, Language Technology Kendra, Patan, Nepal
pavhall@ltk.org.np

Abstract

If a language is to be used on the Internet it needs to be written and that writing encoded for the computer. The problems of achieving this for small languages is illustrated with respect to 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 the language of the Newar people, a language which has a mature written tradition spanning more than one thousand years, with several different styles of writing, and yet has no encoding of its writing within Unicode. Why has this happened? I explore this question, looking for answers in the view of technology of the people involved, in their different and possibly competing interests, and in the incentives for working on standards. I also explore what should be done about the many other unwritten and uncomputerised languages of Nepal. We are left with serious concerns about the standardisation process, but appreciate that encoding is critically important and we must work with the standardisation process as we find it..

Introduction

If we are going to make knowledge written in our language available to everybody in their own languages, those languages must be written and that writing must be encoded in the computer. Those of us privileged to be native speakers of a world language such as English or Spanish may think this is straightforward, but it is not, as will be explained here.

There are nearly 7000 languages world-wide, the current authority on this is Ethnologue (Lewis 2009), who reports the distribution of languages between continents as in Table 1.

Table 1. Distribution of living languages (Lewis 2009)

Area	Living languages		Number of speakers			
	Count	Percent	Count	Percent	Mean	Median
Africa	2,110	30.5	726,453,403	12.2	344,291	25,200
Americas	993	14.4	50,496,321	0.8	50,852	2,300
Asia	2,322	33.6	3,622,771,264	60.8	1,560,194	11,100
Europe	234	3.4	1,553,360,941	26.1	6,638,295	201,500
Pacific	1,250	18.1	6,429,788	0.1	5,144	980
Totals	6,909	100.0	5,959,511,717	100.0	862,572	7,560

Ethnologue reports for each of these whether or not the language is written to the extent of having had the bible translated into it – Trosterud (1999) analysed the Ethnologue entries for all languages, with his results shown in Table 2.

Table 2. Number of languages with bible translations (Trosterud 1999)

Published bibles	Number of languages
Complete Bibles	320
New Testaments only	801
Bible portions	919
TOTAL	2040

2040 languages is just under one third of all languages. Trosterud points out that having bits of the bible written in some script that was invented for that translation does not indicate that the language is in any real sense written. Other factors are much more important, as seen in UNESCO's (2003) linguistic vitality factors 5 and 6;

5 Response to New Domains and Media; it is important that the language is used not just in newspapers but also in films, on TV, in the computer and on the internet;
6 Materials for Language Education and Literacy; the language should be taught in schools, either as a second language or ideally as a first language used as the medium for education.

By looking at the size of the 2000th language in the Ethnologue list Trosterud suggests a figure of 16,000 speakers as a threshold for languages that we could reasonably expect to be literate or be helped to become literate.

This suggests that if knowledge is to be made widely available across the Internet, we need to support languages of at least 16,000 speakers, and these languages need to be actively written with their methods of writing encoded in the computer. 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 as yet does not have its method of writing encoded for the computer.

We are left wondering why Nepal Bhasa has not yet been encoded. I discuss three factors relevant to the encoding of writing in Section 4: lack of awareness of the full power of software technology, the different and conflicting interests of the users of encodings, and perverse incentives that favour historical interests in dead languages over the interests of living users of the writing. There is a deep social injustice in the current situation that favours a few hundred of scholars of an extinct writing over the interest of hundreds of thousands in a community of living users of a language. Finally in section 5 I discuss what might be done in improve the current situation.

2. Nepal 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 in from both directions, those from the north bringing with them languages of the Tibeto-Burman family, while communities migrating from the south mostly brought with them languages of the Indo-Aryan family.

2.1 Nepalese languages

Once in Nepal communities remained completely isolated by steep valleys and high mountains and by thick forest, leading to the evolution of many distinct languages, given as 92 in the 2001 census but now put by Ethnologue at 124 distinct living languages, though this increase in number seems mostly related to distinguishing dialects within larger groups previously thought to belong to a single linguistic community. Ethnologue's linguistic map for Nepal, reproduced in Figure 1, shows the hotchpot of languages scattered across the country.

If we take Trosterud's suggestion that at least those languages with more than 16,000 speakers should be written, we find that we should expect all languages down to and including Dhimal should be written; this is 28 languages, just under one third of the languages, in line with the proportion in the population of world languages as a whole. Table 3 lists these 28 languages plus two others, with relevant

characteristics extracted from Ethnologue. Note that 8 of them have much larger populations across the border in India, with one of these, Maithili, the second largest language of Nepal. This leaves 20 Nepalese languages, only one of which, Nepali, is used in written form in all walks of life and can be considered fully literate; however most of them have at least some limited use in writing.

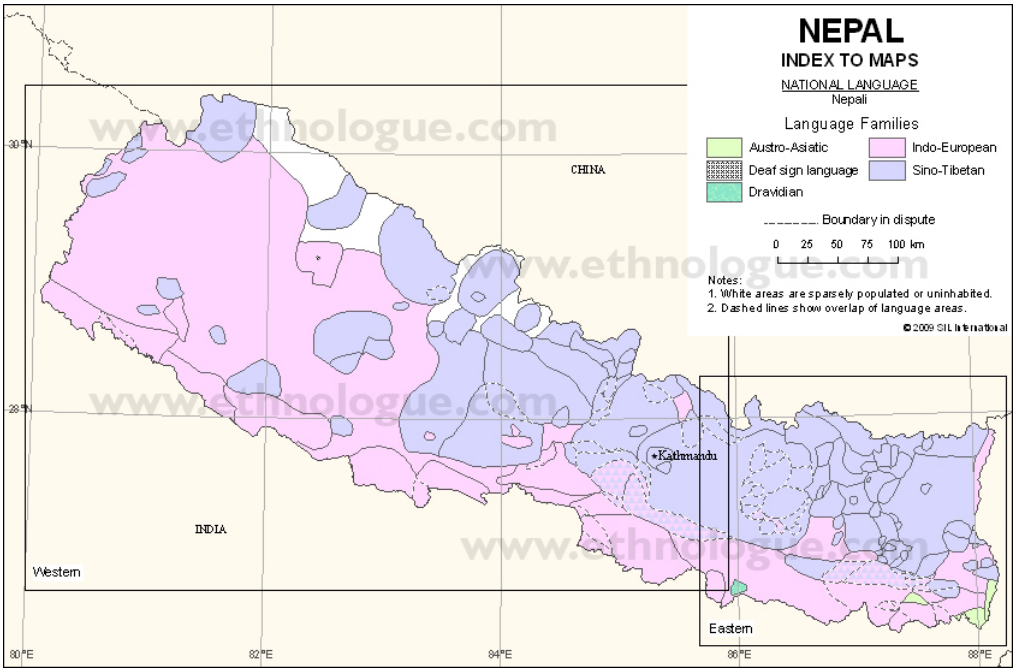

Figure 1. The Ethnologue linguistic map for Nepal, showing the boundaries between linguistic communities.

Table 3. Number of mother tongue speakers of the major languages of Nepal.

Speaker numbers in Nepal are those of the 2001 census – usually Ethnologue reports these, but sometimes reports significantly different figures.

Key to column entries:
Script – D= Devanagari, T = Tibetan, O = other
Bible – from Ethnologue, usually the result of linguistic work by SIL who also produced a dictionary and grammar and chose the Devanagari script: F=full bible, NT = new testament only; P = portions only available.
Literacy – blank, no written use; L = limited, maybe a language description and a small newspaper, M = moderate, some newspapers or similar, and radio programs; F = full, with literature and everyday written use.

Rank	Mother Tongue	Type	Number		script	bible	literacy
			in Nepal	in India/Other			
	Total		***22,736,934***				
1	Nepali	IA	11,053,255	2,765,000	D	F 1914-2004	F
2	Maithili	IA	2,797,582	31.900.000	D, O		F
3	Bhojpuri	IA	1,712,536	36,836.000	D, O	NT 1999-2006	L
4	Tharu (Dagaura/Rana)	IA	1,331,546	224,000	D		M
5	Tamang	TB	1,179,145	14,000	D, T	P 2005	M
6	Newar	TB	825,458	14.000	D, O	NT 1986	M
7	Magar	TB	770,116	71,700	D	NT 1991	
8	Awadhi	IA	560,744	37,700,000	D	F 2005	L

9	Bantawa	TB	371,056	19,800	D		L
10	Gurung	TB	338,925	33,000	D	NT 1982	L
11	Limbu	TB	333,633	87,000	D, L, O	P 1996-2003	M
12	Bajjika	IA	237,947		D		
13	Urdu	IA	174,840	10,700,000		F 1843-1998	F
14	Rajbansi	IA	129,829		D		L
15	Sherpa	TB	129,771	31,500	D, T	P 1975, NT 2000	L
16	Hindi	IA	105,765	180,000,000	D	F 1818-2000	F
17	Chamling	TB	44093				L
18	Santhali	AA	40,260	6,117,000	O. D, B, L	F 1914-2005	M
19	Chepang	TB	36,807		D	NT 1993	L
20	Danuwar	IA	31,849		D		
21	Jhangar/ Dhangar / Kurux	D	28,615		D	P 1977	
22	Sunuwar	TB	26,611		D	NT 1992	L
23	Bangla	IA	23,602	110,000,000	B	F 1809-2000	F
24	Marwari (Rajsthani)	IA	22,637	5,600,000		NT 1820-21	L
25	Manjhi	IA	21,841	20,400			
26	Thami	TB	18,991	800			L
27	Kulung	TB	18,686		D		L
28	Dhimal	TB	17,308	450	D		L
29	Angika	IA	15,892	725,000	D		L
...	...						
50	Lepcha/ Lapche	TB	2,826	48,000	O, T	NT 1989	M

2.2 Nepalese writing

However only four of these purely Nepalese languages have any significant tradition of being written:

- Nepali, historically known as Khas, Parbatiya and Gorkhali, with 11,053,255 speakers in 2001, has been written in Devanagari, the script used across north India and in particular for Hindi, for around 300 years.
- Newari, with 825,458 speakers in 2001, is known as Nepal Bhasa within the linguistic community, and has been written for over a thousand years in a number of scripts.
- Limbu with 333,633 speakers in 2001, has a traditional script Sirijanga which was probably derived from Lepcha writing (Omniglot 2010). It is claimed to have been invented in the 9th century and then revived in the 17th century by Te-ongsi Sirijonga, and then revived again in 1925 when it was formally named "Sirijanga".
- Lepcha (also known as Rong), with 2,826 speakers in Nepal but 48,000 in Sikkim in India, is written in a script evolved from the Tibetan script, which tradition claims was devised in the 17th or 18th centuries (Wikipedia 2012b).

Ethnologue only reports limited literacy for Newari and Limbu, not surprising since these languages were suppressed by successive Nepalese governments from the late 18th century onwards until 1990. While the writing of Limbu and Lepcha 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 their regime by the Gorkhas in the mid 18th century. Note that cross border languages, and particularly Maithili and Bhojpuri, also have their own mature literature and may be written in their own distinctive script; for Maithili the script is known as Mithilaksha or Tirhuta, for Bhojpuri it is Kaithi.

Indic writing including Devanagari and Bengali has been printed in movable type since around 1800, with the type evolving and being simplified over the centuries (see for example 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). Work had been proposed to include Devanagari within the then established standard for computers, ISO 8859 (Wikipedia 2012), as part 12, but this work was abandoned expecting to adopt ISCII's codes into ISO 8859. However ISO 8859 was in turn superseded by Unicode, which included a code block for Devanagari and other major Indic scripts from the start, with the code blocks adapted from a 1988 version of ISCII (Unicode 1990). One significant difference between ISCII and Unicode was that in ISCII all the scripts of India had been unified within a single table, with the different scripts selected by appropriate font, whereas in Unicode these were dis-unified into separate code blocks.

The encoding of Limbu was added to the Unicode Standard in April 2003 with the release of version 4.0. Limbu was introduced to the standardisation process by McGowan and Everson in 1999, and a proposal was written jointly by Boyd Michaelovsky and Michael Everson in 2002. Michaelovsky is a linguist who has done considerable field research among the Limbu in Nepal learning about their writing in context, appealing in the proposal to both examples of writing and to the phonology of the spoken language. Even so there have been some discussions since then about missing characters, and in 2011 Pandey proposed two additional composite characters, though there is a case for introducing the virama instead.

The encoding of the Lepcha script was initiated by Michael Everson and others within the Unicode technical Committee in 2003, and formally proposed in 2005 (Everson 2005), finally being added to the Unicode Standard in April 2008 with the release of version 5.1. Primary sources of knowledge about Lepcha writing in the Everson document are from two academic texts from the late 19th century and several texts from the 1970s, with copious samples of writing taken from these texts included in the appendices, plus reference to two experts consulted, a linguist in Leiden in the Netherlands, and a typographer with Xenotype in the US.

While the writing of Lepcha and Limbu have followed a normal path to standardisation – an introduction of the script to the standardisation community, followed by a full proposal, and then agreement within the ISO and Unicode committees, leading to inclusion in the next version of the standard, Newar writing has not had such a smooth passage, as will be discussed in section 2.4. But before that I briefly discuss other proposals for writing languages of Nepal.

2.3 Other languages and their encoding

Field linguists aiming to document the languages that they study, and members of the Summer Institute of Linguistics (SIL), have for many years improvised a means of writing the language, usually based on Devanagari. Michael Noonan (2003) has given a very thorough analysis of some of these, relating the choices made to the underlying phonologies of the languages.

When the Indian constitution first scheduled its official languages Maithili was viewed as a dialect of Hindi, a view that was vigorously contested and eventually led to the inclusion of Maithili as a distinct scheduled language in 2004, though still written in Devanagari. Their traditional style of writing, Mithilaksha/Tirhuta, was treated as an exotic for use in wedding invitations and similar, though discussions have been made as to whether it could be unified with Bangla or with Devanagari. In 2008 a Unicode compliant Mithilaksha font Janaki was produced in Nepal, mapped to the Devanagari code block, implicitly assuming a unification with Devanagari for the advantage that existing documents encoded in Devanagari could be rendered in Mithilaksha by a simple change of font. Then in 2011 Pandey proposed a separate encoding of Tirhuta, arguing briefly and inadequately that it could not be unified with Bengali, but not discussing the situation with respect to Devanagari.

A large proportion of Nepal's languages are not yet written, though linguists and anthropologists have written fragments of many languages using extensions of Devanagari (Noonan 2003). Some language activists have created their own distinctive writing, with proposals that have reached discussion towards standardisation (Anderson 2012) – the languages are Sunawar (Pandey 2011a, -b, -i), Bantawa (Pandey 2012c, -g), Gurung (Pandey 2012d), Magar (Pandey 2012f), and Dhimal (Pandey 2012j). Much of the drive for the writing of several of these languages seems to come from Sikkim 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 2.

Mark Turin

Figure 2. The Sikkim Herald in 11 languages (with permission from Mark Turin.)

All of the scripts or writing styles for these languages are seen as candidates for separate standardisation, apart from Magar who claim to write their language in Brahmi (which they call Akkha) and thus Pandey concludes:

> Until additional research provides information that clearly differentiates it from Brahmi, Magar Akkha should be considered a variant of the latter and unified with it.

In the discussion about the Tikamuli writing for Sunawar, Pandey notes:

> It has no genetic relationship to other writing systems, although it has similarities to the Limbu (Sirijonga) and Lepcha (Rong) scripts.

What is meant by a "genetic relationship" is not clear, there will certainly have been contact between the linguistic groups with the diffusion influences that then take place. What is not being considered for these, apart from Magar, is unification with any other Unicode code blocks.

3. The Newar's Nepal Basha and Nepal Lipi

Figure 3. Newar writing from the Golden Temple in Patan, Kathmandu valley, Nepal.

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. They call the Kathmandu valley "Nepal", their language "Nepal Bhasa" the language of the Kathmandu valley or Nepal, and their writing of it "Nepal Lipi" or "Nepaalalipi", the writing of Nepal. In the temples around the valley you can see their writing carved into stone or wood, or embossed in brass or other metals, as seen in Figure 3.

3.1 The styles of writing Nepal Bhasa

There are two distinct styles of writing in Figure 3, an ornate style with many long downward strokes which they call "Ranjana", and a more rounded style which they call "Prachalit". While these are not the oldest examples of their writing, very old examples can be seen in the Patan Museum nearby.

Other styles of writing have also appeared. Rabison Shakya (2002) in his book on how to write Nepal Bhasa identifies a third style which he refers to as Bhujimmola, while Hemraj Shakyavansha (1985) identified 9 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.

Figure 4 shows part of a contemporary Newari newspaper with the headline in Ranjana and body in Prachalit.

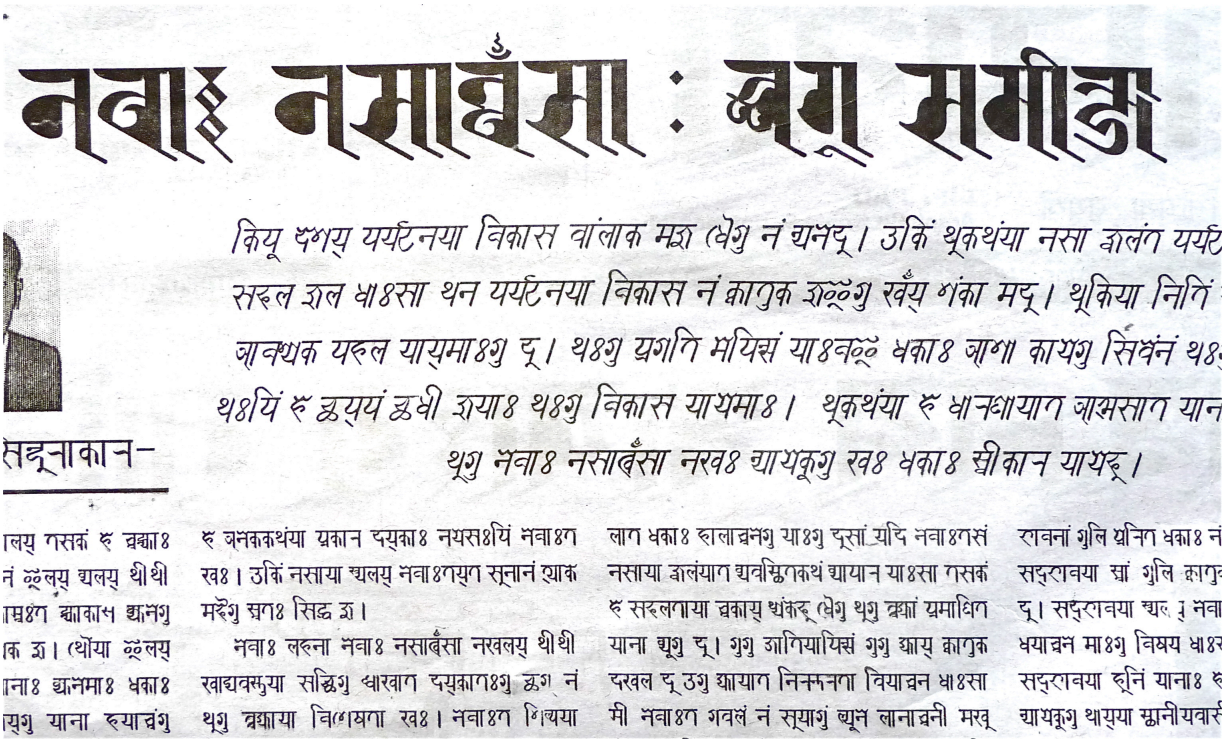

Figure 4. Newspaper with Ranjana headline, and Prachalit body text.

Much current writing of Nepal Bhasa is done in Devanagari, though this is not satisfactory. Hack fonts are available for both Ranjana and Prachalit, based on eight bit encodings determined by the key on the keyboard to which the character has been assigned for input. What is needed is a proper encoding in Unicode of Nepal Lipi for which open type fonts can then be produced.

Around 1998 a proposal was submitted to Unicode by a committee in Nepal for a distinct encoding for the writing of Nepali, to include three common consonant compounds (conjuncts) - <tra>, <ksha>, and <gya> - that collate separately and are taught as part of the basic alphabet in Nepal. This proposal was rejected on the basis that the three conjuncts distinguished did not define a writing system difference and should continue to be treated as conjuncts with the collation differences handled through collation algorithms. This seemed fair enough, and after this software was developed for the Nepali language working with the Unicode Devanagari code block (eg. Bal, Gurung and Hall 2006).

3.2 Early attempts to encode Nepaalalipi

In 2001 the Unicode expert Michael Everson posted a proposed code block (Everson 2001) which he named "Newari" and illustrated with graphic characters from Ranjana, and also about the same time he posted a draft code block named "Nepali" and illustrated this with graphic characters from Prachalit (this posting is no longer available); what is singular about these drafts was their inclusion of the same three conjuncts, <tra>, <ksha>, and <gya>, as in that earlier proposal for Nepali, presumably influenced by this.

Though the Newar community already had 8-bit hack fonts for their writing (the book by Shakya (2002) could be viewed as a description of his 8-bit hack fonts), a number of Newar activists began to take an interest in Unicode standards for Newar writing, with ideas expressed as proposed code blocks. A meeting of the Nepal Lipi Guthi in July 2008 explored the ideas that each style of writing should be separately encoded and that the actual shapes of the characters should themselves be standardised.

I became involved at about this point, and studied the book by Shakya (2002) to find out more about the writing of Nepal Bhasa, and how it differed from Devanagari which I knew was often used to write Nepal Bhasa. Shakya gives small tables of the basic characters of the writing, for Ranjana, for Prachalit, and for

Bhujimmola, but to my surprise these were not equivalent, the Prachalit tables had significantly more characters, as can be seen in Figure 5 for the consonants of the three writing styles.

Figure 5. The consonant characters of the Ranjana, Prachalit and Bhujimmola styles as tabulated by Rabison Shakya (2002).

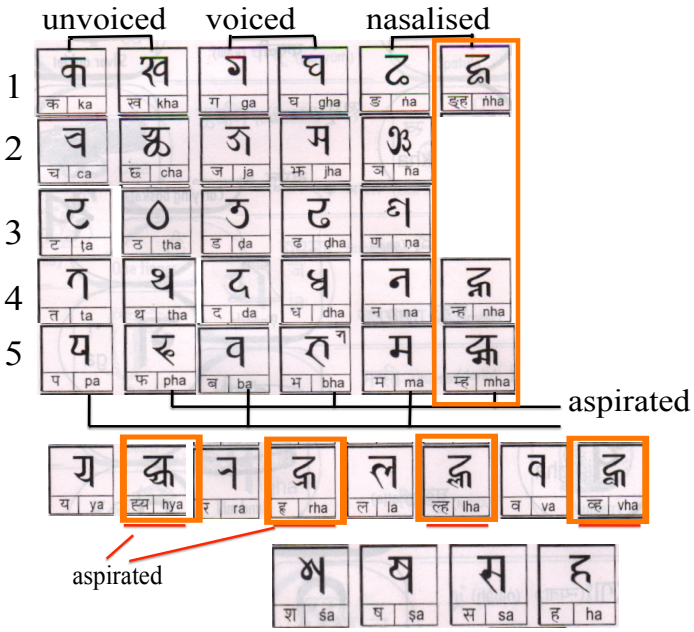

Figure 6. Prachalit in Vargs, highlighting the "extra" aspirated consonants.

The differences in vowels are due to the composition of a vowel with a diacritic, but with consonants something deeper seemed to be happening – Figure 6 rearranges the consonants of Shakya's Prachalit table to show the 7 "extra" consonants in the groups or vargs used in South Asia to analyse the organisation of a script. All of these extra consonants look as if they are a combination of two more basic consonants, so we must ask ourselves if they represent anything special – and we find from Hale and Shrestha (2005) that they are phonologically distinct aspirated or breathy consonants, different from a consonant followed by a "ha". Further, they are written as if the "ha" came first, so these glyphs must be viewed as distinct letters of the spoken Newar alphabet. We will discuss these further later.

In 2009 the Script Encoding Initiative at Berkeley University funded Everson to encode Ranjana. The starting document was a paper to the ISO WG2 committee (Everson 2009a) which advocated:

> Since Rañjana is visually and structurally similar to the Lañtsa & Wartu scripts used for Buddhist Sanskrit documents in Tibet (China), Bhutan, Mongolia, Nepal, Sikkim & Ladakh (India) it has been considered would be practical to merge these two scripts (Lañtsa and Wartu) with Rañjana for encoding purposes. (p1)

An email discussion list was created, which was very productive with many heated exchanges, but no consensus emerged. A very strong lobby of Newars wanted Prachalit to be encoded first, as the style of writing used in the daily life of the Newars, while I personally argued that Ranjana and Prachalit were equivalent, just different visual expressions of the same underlying writing to support the language Nepal Bhasa, like fonts; but I was roundly rebuked because "Unicode did not encode languages, Unicode encoded writing".

After much debate, Everson posted a second document (2009b) which began:

> Analysis of the Nepalese scripts is slightly complicated because different font styles have different names, and some writers equate name with script identity. Analysis of the chief features suggests that a number of these can be unified on structural grounds. (p1)

and then after a discussion of the visual appearance of the various writing styles, 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. (p2)

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, though the level of agreement from that meeting was unclear, and no record of the meeting was produced. It seemed that if an encoding were to be produced, it would cover all the variants for writing Nepal Bhasa within a single code block.

3.3 Recent encoding proposals

So it was a surprise when a proposal to encode Prachalit was circulated (Pandey 2011), soon followed by a proposal to encode "Newar" (Pandey 2012a). This latter proposal evolved from Pandey's earlier Prachalit proposal, replacing it, with a very similar draft code block but having 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 "Nepal Lipi" or "Nepaalalipi", their traditional collective name for their scripts. This choice was argued against in correspondence with the Script Encoding Initiative, an organisation very close to the Unicode Consortium, on the grounds that "Lipi" translates as "script" and names of code blocks cannot include the word "script". This denial of their traditional name for the code block has considerably offended the Newar community, who know that the two Japanese syllabaries Hiragana and Katakana have code blocks with those names that contain "gana" and "kana" which mean "script". This objection was further compounded by a 35 page document from Pandey (2012b) further arguing that "Newar" was the correct choice. This later document is very thoroughly researched, but seems to miss the point that the purpose of the code block proposed is to support more than the writing of the Newar language, having in mind to add additional characters later to meet the needs of other languages of Nepal.

Pandey's Newar proposal (2012a) also included a discussion of a number of "additional consonantal forms", shown in Figure 7.

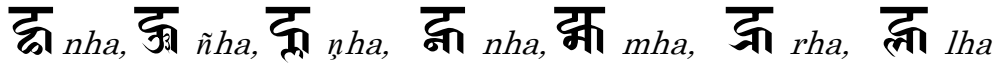

Figure 7. Newar characters – are these consonants or conjuncts?

These are those extra aspirated consonants included in Shakya's Prachalit table and analysed earlier. Pandey states 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 the pronunciation is that the breathy "ha" follows, then, so Pandey argued, the "ha" should correctly be written at the bottom.

At around the same time as Pandey's proposal, Devdass Manandhar and colleagues (Manandhar 2012a) submitted an alternative proposal to the Unicode Technical Committee (UTC) that described the Newar writing from a Newar perspective; this proposal included the standard ISO WG2 form and should then have been automatically submitted to the ISO WG2 registry, but this was not done, though a direct submission by Manandhar has now placed it in the ISO WG2 registry (Manandhar 2012b). At their May meeting the UTC considered both proposals, giving a very short response to the authors concerned (Unicode Technical Committee 2012). There was no discussion at the UTC July meeting, so there it rests.

I have analysed the differences between the two proposals, and found strong agreement in many aspects of the proposals. Manandhar and colleagues were concerned about duplicate ways of encoding long vowels and about collation order based on code sequence; neither of these are issues for the encoding and are handled in other ways, and thus can safely be ignored.

However they do include those extra breathy consonants, and have subsequently found examples of their use from the end of the 18th century. While it is clear that the breathy consonants are present in the spoken language, why are they written in this way? The fact that Hale and Shrestha found them as phonemes distinct from the sequence of the consonant followed by <ha> shows that there must be contrasting pairs of words with different meanings with the breathy consonant and with the consonant followed by <ha>. So to avoid ambiguity in the writing, the breathy consonant needs to be written differently, and writing them in the wrong order is one way to achieve this. Equally well we must assume that the sequence <ha> followed by the consonant never occurs, though this needs checking. Noonan (2003) reports a similar problem in the writing of Chantyal, leading to a similar convention in the writing of Chantyal using Devanagari. An alternative possible explanations is that at one point in the past when the orthography of Nepal Bhasa was established, pronunciation was as written, but with the passage of time pronunciation changed and that what we now see is evidence for a need for a spelling reform, similar to that carried out several times for Dutch and recently for German. Members of the Nepal Lipi Guthi have discussed the need for new glyphs to be designed for these characters, and they have been assured that if this was done and the glyphs used in a number of documents, the case would be much stronger.

What is now needed is a composite proposal that meets the constraints of ISO WG2 and the UTC and the approval of the user communities as represented by those who have brought forward the proposals. The two major stumbling blocks are those "extra" aspirated consonants shown in Figures 6 and 7 and discussed above, and the name of code block. There could be considerable advantage in including extra characters for some of the other Tibeto-Burman languages currently being proposed for separate encoding.

It does look as if any new proposal should include Ranjana, and the example seen in Figures 4 shows the use of both Ranjana and Prachalit writing styles in the same document. However Everson has most emphatically declared that Ranjana must be encoded separately, and slots for both Newar and Ranjana appear in the forward planning Roadmap of Unicode (Everson et al 2012), and in the forward planning of the Script Encoding Initiative (Anderson 2012).

4. Why these standardisation delays?

We need to understand why have these delays in standardising the encoding of the writing of Nepal Bhasa have taken place, so that the standardisation can move forwards. It appears that there are three critical factors at work here:

- lack of appreciation of the nature of the technologies now being deployed;
- divergent interest groups who focus on their own concerns, largely ignoring or even denigrating the interests of other groups;
- perverse incentives that favour some decisions over others.

I will pick these critical factors up in turn and discuss them in detail in the context of the interdisciplinary nature of the encoding of writing.

4.1 Powerful technologies for flexible use

Originally the facilities for writing using computers were very crude, very close to old-fashioned manual typewriters only capable of creating texts in a single size and font, with the input keys, internal computer codes, and output characters in one-to-one correspondence. Today you can use key-mapping software to make any key press or combination and sequence of key presses produce a particular internal code, and with open-type fonts make the character rendered on a screen or in print depend upon not just the current code but a complete sequence of internal codes.

When writing moved from pen to keyboard secondary aspects of the writing had to be made explicit. So for hand-writing languages that used both small and capital letters within the Roman alphabet, the decision to write one or the other was implicit, but now had to be indicated explicitly: this was initially achieved by a shift key, linked to the mechanical systems of typewriters and since carried through into keyboards but not into current internal coding systems which now include both small and capital letters.

In Asian writing derived from the Brahmi abugida system, all consonants have an implicit “a” sound, which can be overridden if the diacritic for a different vowel is present, and is suppressed if consonants are written as clusters or conjuncts. There is also an explicit diacritic called the virama or halanta used to suppress the implicit “a” at the end of a word. While the basic alphabet may only include between 50 and 100 distinct characters, there may be as many as a thousand consonant clusters with visually distinct written forms; in many cases these are created from reduced forms of the basic consonant joined together either horizontally or vertically. Mechanical typewriters were very limited in the form of the clusters that could be typed, and the horizontal half forms of the glyphs were used and may have been explicitly invented for the typewriter because they could be typed in sequence to form a cluster. With the coming of computers and open type more sophisticated strategies can now be used; most Brahmi scripts, such as Devanagari, use a virama followed by a vowel to signal a cluster. For some languages, notably Tibetan, a different strategy was used, building on the half consonants but now using the superior technology of open type these can be rendered as vertical stacks as is done in handwriting, with consonants having in effect two forms, a form with the implicit “a” and a second “sub-joined” form without. These extra actions to make explicit what is implicit in handwriting affect the input of text, and need not have been carried through in identical fashion into the internal codes but unfortunately were, and thus most Brahmi code blocks are described as “virama models” while Tibetan follows a “subjoined model”. Manandhar’s proposal for Nepaalalipi has elements of both models and could go either way, though the UTC has advised that it should be a virama model. Internal codes need not have followed either of these approaches focused on surface features of the writing, and could instead have been more abstract with the internal coding model discarding the abugida and using a deeper and simpler alphabetic model with all vowels explicit, with input mappings and output rendering converting between the surface form for external communication and the semantic internal form.

These facilities for inputting, encoding, and outputting text are used in the context of typographical software, such as WYSIWYG document editors like Microsoft Word or XML with document formatting DTDs like docbook or TEI. The visual appearance of the text, the relative sizes and layout, are an important second order aspect of communication using writing that need to be controlled. For example, I can choose to make my headings in a different font and different size from the body, as I have done in

this document, only making that decision after I have completed the document. Originally everything was in the sans serif font Helvetica, but later I changed the body text is in the serif font Times New Roman. Looking at examples of Newar writing in newspapers such as that in figure 4 above it seemed that similar typographical flexibility would benefit the community, but when I raised this on a discussion list, that having prepared text in say Prachalit you might want to change parts of this to Ranjana using a font change, I got a robust and scathing negative response from the Unicode expert leading the discussion. It seemed he just did not know enough about the technology and its possible use.

An issue often conflated with encoding is collation. It is important to be able to compare two segments of text to determine whether they are equivalent, and if not which should come first when being sorted alphabetically. Because the characters of the writing are encoded as numbers, it is tempting to think that the numerical order of these codes should determine the collation sequence, but this wouldn't always work. For example, are capital letters and small letters equivalent when sorting a list of words or searching for a match on the Internet – clearly sometimes they are and sometimes they are not. Thus the practice in software is to define collation separately from coding, possibly using some intermediate "normalized" form of the text. Manandhar's draft for Nepaalalipi has been careful to place its characters in the "correct" sequence for sorting, but the numerical sequence of characters within a code block can be arbitrary, and related characters can be grouped to facilitate communication and understanding. Further Manandhar's draft has avoided duplicate ways of representing characters, for example expecting long vowels to be indicated by a short vowel followed by the lengthening character visarga – this is not necessary, short and long vowels and the lengthening character can all be included, as recommended by the UTC.

4.2 Living languages versus dead texts

What exactly does Unicode standardise? The Unicode Consortium on its website http://www.unicode.org claims

> Unicode provides a unique number for every character (on page WhatIsUnicode)
> Unicode enables people around the world to use computers in any language (on home page)

and then in its FAQs elaborates:

> Unicode encodes scripts *for* languages, rather than languages per se
> Unicode Standard encodes characters on a per script basis
> the Unicode Standard does not encode *scripts* per se

Unicode clearly is concerned with languages while encoding the characters used in their writing. However when I have raised concerns about language having in mind the needs of communities of living users, I have been told firmly that Unicode does not encode languages but encodes writing. But writing for what purpose? In the debate about those extra characters used for writing the Newar's language Nepal Bhasa I appealed to the phonology of the language as currently spoken, to be told that these characters could only be encoded if they have a visually distinct shape that has been widely used over some significant period. I have been told that the need for new distinct graphical forms has been discussed within the Nepal Lipi Guthi, but no design work has yet been undertaken.

This insistence on referring to the written record rather than the needs of contemporary users biases the standardisation process towards the users of encodings of antiquarian scripts, such as historians and librarians. A historian interested in linguistic communities from the past only has antiquarian documents to study and will quite reasonably focus on the visual similarities of the writing across different manuscripts. This has led to claims like that quoted earlier from Everson (2009a) for unification of Ranjana and similar looking scripts in Tibet and Bhutan. It also leads to the statement in Pandey's proposal for Newar/Prachalit (2012a)

the Newar script is being promoted as the written standard by various Newar organizations. The script is also used for writing Sanskrit and Nepali. It was used historically for writing Maithili, Bengali, and Hindi.

In his view it is the visual similarity of the scripts that unifies the script across languages. This contrasts with the statement in Everson's document (2009b) which focuses on visual differences:

> **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. At this stage, at least we can distinguish:
> **2.1 PRACALIT flat-headed script,**
> **...**
> **2.2 BHUJIṀMOL curve-headed script,**
> **...**

Two scripts are seen as different because of minor differences in the way particular characters are written. These distinctions rest simply upon the surface visual appearance of the writing and do not rest upon the deeper semantics for writing a particular language.

These arguments based on visual appearance just do not make sense when we contemplate how we use our own Roman or Latin writing. We could write a given text in the Times New Roman font or a Fraktur or Blackletter font: it would be the same text though looking significantly different, both written in the Roman writing system.

If we now consider the writing needs of languages that are currently unwritten, the Unicode focus on writing make it impossible for languages that are currently only spoken. UNESCO (Robinson and Gadelii 2003 - see also Cahill and Karan 2008) provides guidance on how to create a writing system for such languages, starting with a phonological analysis after which the required phonemes would be given graphical form, probably as extensions of some locally dominant script. Several of Nepal's languages have in effect undergone this procedure by field researchers and bible translators, to be written in extensions of Devanagari (Noonan 2003). The key step is that first step of phonological analysis, and by making a phonemic inventory of a group of related languages, such as the Tibeto-Burman languages of Nepal (partially given by Noonan 2003), we should be able to create a single code block that met the needs of all these languages. The trouble is that Unicode requires established written forms before any codes can be assigned. My contention is that this could happen now based on the phoneme inventory.

4.2 Perverse incentives

Coding proposals are written by a surprisingly small set of people. These proposals are listed in the document registers on the ISO WG2 website (2012a, b, c), Table 4 shows the number of proposals and other standards documents authored or co-authored by some of the star Unicode authors.

It clearly helps in the proposal writing process to have people involved who are experienced and know what needs to be done, so the top four people would be valuable assets to Unicode and ISO WG2. They receive special accolades from Unicode, and in one instance an author was featured in the New York Times (Erard 2003). Some of the authors get paid for their contribution through the Script Encoding Initiative in Berkeley who in turn receives funding from the US government. 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, the writers of proposals are incentivized to see differences and encode scripts separately. I have heard other stories where the financial incentives have led the authors of alternative encoding proposals for a particular script to behave as competitors rather than seeking to collaborate to produce the best encoding for the community of users.

Table 4. Authorship of ISO WG2 proposals

Person	number of documents authored in period			total
period	Sept 08 – Oct 10	Oct 10 – June 11	June 11 – Feb 12	
star 1	42	41	15	98
star 2	19	41	21	81
star 3	9	20	15	44
star 4	19	17	4	40
other 1	9	5	0	14
other 2	6	7	1	14
other 3	2	5	0	7

This works strongly against the interests of the linguistic communities of Nepal:

- different styles of writing Nepal Bhasa are likely to be encoded differently rather than using a single encoding with the different styles captured by fonts, and
- the different styles created recently for writing smaller languages like Bantawa, Dhimal Gurung , Magar, and Sunuwar could be unified with Nepaalalipi, with the benefits of sharing fonts, but the incentives seems unlikely to allow this to happen

5. Where do we go from here?

It seems clear that some reorientation of the encoding standardisation process towards languages and the living users of those languages is necessary if we are prevent the needless proliferation of encodings for scripts which are essentially the same. If this proliferation continues we will end up with a situation not unlike that encountered in Asia with hack fonts and their accidental encoding requiring that if people are to share documents across the internet they must all possess the same font with it own unique encoding. Unicode had seemed the means of saving South Asian languages from hack fonts, it now seems likely to perpetuate the same situation with hack encodings.

Meanwhile those of us concerned about the writing of Nepal Bhasa and its community of users living today in Nepal and in the diaspora must aim to submit a new proposal uniting the best of the previous proposals, hoping that the wider international community in ISO WG2 will understand our position. It could be beneficial to also unify with those other scripts of Tibeto-Burman languages Bantawa, Dhimal Gurung , Magar, and Sunuwar currently proposed for separate encoding, as well as those currently written in augmented Devanagari described by Noonan (2003): Chantyal, Sherpa, Tamang and Thangmi.

Beyond this I am concerned that hack encodings may already have proliferated within Unicode, having in mind the N'Ko writing system of West Africa that looks so like a variant of Arabic writing from which has clearly been derived. How many more are there like these, should a major review be undertaken?

As to the fate of languages that are currently not written, where the rational approach should be to encode their phonemic inventory, leaving to font designers the graphical appearance that might look like established writing systems, well we will just have to hope.

References

Anderson, Deborah (2012) *Liaison Report, Script Encoding Initiative, UC Berkeley*. Document ISO/IEC JTC1/SC2/WG2 N4220 2012-02-12

Bal, Bal Krishna; Srishtee Gurung and Pat Hall (2006) Towards Universal Access to ICTs in Nepal *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

Coulmas, Florian (1989) *The Writing Systems of the World*. Blackwell

Crystal, David (2000) *Language Death*. Cambridge University Press.

Diringer, David (1962) *Writing* Thames and Hudson

Erard, Michael (2003) For the World's A B C's, He Makes 1's and 0's, New York Times September 26, 2003

Everson, Michael (2000) *Newari code table and names list* http://www.evertype.com/standards/tai/newari.pdf (accessed 2012/02/15)

Everson, Michael (2005) Proposal for encoding the Lepcha script in the BMP of the UCS, ISO/IEC JTC1/SC2/WG2 N2947R 2005-07-04

Everson, Michael (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 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3649.pdf (accessed 2012/02/15) also from http://www.evertype.com/formal.html

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

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 Downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4186.pdf

Gaur, Albertine (1992) *A History of Writing*. Revised Edition. Cross River Press.

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

Hartell, Rhonda L (editor) (1993) Alphabets of Africa. UNESCO-Dakar Regional Office and Summer Institute of Linguistics, Dakar

ISO WG2 (2012a) SC2/WG2 partial document register (N3505-N3948) Document N3800 on http://std.dkuug.dk/jtc1/sc2/wg2/ accessed 29/8/2012

ISO WG2 (2012b) SC2/WG2 partial document register (N3800 – N4116) Document N3800 on http://std.dkuug.dk/jtc1/sc2/wg2/ accessed 29/8/2012

ISO WG2 (2012c) SC2/WG2 partial document register (N4046 – N4255) Document N4100 on http://std.dkuug.dk/jtc1/sc2/wg2/ accessed 29/8/2012

Kasah, Sharad Junior (2009) Paper distributed at meeting of the Nepal Lipi Guthi in Kathmandu, March 2010.

Lewis, M. Paul (ed.), 2009. *Ethnologue: Languages of the World, Sixteenth edition*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com/

Manandhar, Devdass, Samir Karmacharya, Bishnu Chitrakar (2012a) Proposal for the Nepaalalipi script in the UCS, submitted 2012-02-11, UTC document L2/12-120

Manandhar, Devdass, Samir Karmacharya, Bishnu Chitrakar (2012b) Proposal for the Nepaalalipi script in the UCS, 2012-02-05, ISO/IEC JTC1/SC2/WG2

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. Downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2410.pdf

Nettle, Daniel and Suzanne Romaine (2000) Vanishing Voices. Oxford.

Noonan, Michael (2003) *Recent Adaptations of the Devanagari Script for the Tibeto-Burman Languages of Nepal*. https://pantherfile.uwm.edu/noonan/www/Papers.html (last accessed 29th May 2010)

Omniglot (2010) Limbu / Karanti alphabet http://www.omniglot.com/writing/limbu.htm (accessed 23 June 2010)

Ostler, Nick (2009) The Alphabetic Principle and its Enemies, Keynote address to the Internationalization and Unicode Conference, 33, San Jose, CA, USA, 14-16 October 2009.

Pandey, Anshuman (2011a) *Preliminary Proposal to Encode the Jenticha Script in ISO/IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N3962 2011-01-25 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3962.pdf (2012/02/20)

Pandey, Anshuman (2011b) *Preliminary Proposal to Encode the Tikamuli Script in ISO/IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N3963 2011-01-25 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3963.pdf (2012/02/20)

Pandey, Anshuman (2011c) *Introducing the Khambu Rai Script*. Document ISO/IEC JTC1/SC2/WG2 N4018 2011-04-13 viewable at http://www.anshumanpandey.com/ and downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4018.pdf (2012/02/20)

Pandey, Anshuman (2011d) *Introducing the Khema Script for Writing Gurung*. Document ISO/IEC JTC1/SC2/WG2 N4019 2011-04-13 viewable at http://www.anshumanpandey.com/ and downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4019.pdf (2012/02/20)

Pandey, Anshuman (2011e) *Proposal to Encode the Tirhuta Script in ISO/IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N4035 2011-05-01 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4035.pdf (2012/02/20)

Pandey, Anshuman (2011f) *Introducing the Magar Akhar Script*. Document ISO/IEC JTC1/SC2/WG2 N4036 2011-05-01 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4036.pdf (2012/02/20)

Pandey, Anshuman (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 (2012/02/20)

Pandey, Anshuman (2011h) *Preliminary Proposal to Encode the Prachalit Nepal Script in ISO.IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N4038 2011-05-03 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4038.pdf (2012/02/20)

Pandey, Anshuman (2011i) *Proposal to Encode the Jenticha Script in ISO/IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N4028 2011-05-31 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4028.pdf (2012/02/20)

Pandey, Anshuman (2011j) *Introducing a Script for Writing Dhimal*. Document ISO/IEC JTC1/SC2/WG2 N4140 2011-09-29 and downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4140.pdf (2012/02/20)

Pandey, Anshuman (2012a) *Proposal to Encode the Newar Script in ISO.IEC 10646*. Document ISO/IEC JTC1/SC2/WG2 N4184 2012-01-05 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4184.pdf (2012/02/20)

Pandey, Anshuman (2012b) Standard Nomenclature for the Newar Script: Considerations for International Standards. Private distribution, 26th June 2012.

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

Rogers, Henry (2005) *Writing Systems. A Linguistic Approach* Blackwell

Ross, Fiona. (1999) *The printed Bengali character and its evolution*, Curzon Press ISBN 0-7007-1135 X.

Sampson, Geoffrey (1985) *Writing Systems. A linguistic introduction*. Stanford University Press

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

Shakyavansha, Hemraj. (1985). *Nepalese Alphabets = Nepāl lipi saṁgraha.* Seventh Edition.

Trosterud, Trond (1999) How Many Written Languages in the World? Foundation for Endangered Languages http://www.ogmios.org/ogmios_files/117.htm (14 Aug 2012)

UNESCO (1998) *Universal Declaration of Linguistic Rights* Downloaded on 24/10/2010 from http://www.linguistic-declaration.org/versions/angles.pdf

UNESCO (2003) *Language Vitality and Endangerment*, UNESCO, Paris

Unicode Consortium (1990) *The Unicode Standard. Worldwide Character Encoding. Versn 1.0, Volumes 1 and 2*. Addison Wesley.

Unicode Consortium (2001) *Roadmaps* ISO/IEC JTC1/SC2/WG2 N2383 downloadable from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2383.pdf accessed 2012/02/14

Unicode Consortium (2007) *The Unicode 5.0 Standard.* Addison Wesley

Unicode Consortium (2012) *Roadmap to the SMP* http://www.unicode.org/roadmaps/smp/

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

Vajrachaarya, Suwarn (2009) *Towards Encoding Nepal Scripts: A Report and Preliminary Proposal*. Distributed at meeting of the Nepal Lipi Guthi in Kathmandu, March 2010.

Wikipedia (2012a) *ISO 8859*. Accessed 18th August 2012

Wikipedia (2012b) *Lepcha Alphabet*, Accessed 19th August 2012

Yadava, Yogendra P. (2003) Language Chapter 4.*Population Monograph, volume 1,* Kathmandu: Central Bureau of Statistics.2003 and UNFPA

Yadava, Yogendra P. (2007) Linguistic Diversity in Nepal Perspectives on Language Policy International Seminar on *Constitutionalism and Diversity in Nepal***,** CNAS, 22-24 August,