Bandhu Collection Project: Working Paper 1

Restoring and Standardising an Early Digital Dataset

Andrew Hardie
Department of Linguistics and English Language
Lancaster University

Introduction

The Bandhu Collection Project (BCP), funded by the British Academy[1] from January 2006 to January 2007, aimed to retrieve for the use of contemporary linguists a digital resource from the very earliest days of linguistic data on computer: the Bandhu Collection.

The nature of the dataset

In the early 1970s, a collection of spoken Nepali text was gathered by Professor C. M. Bandhu of Tribhuvan University, Kathmandu. The texts, amounting to approximately 43,500 words[2] of transcriptions, represent the Nepali of a number of villages in different districts of Nepal. The Bandhu Collection is small by the standards of modern corpora, though large by the standards of the time. While originally produced in digital form, the encoding and formatting of the data was not standardised and until 2006 the Collection existed only as a set of duplicated hard copies.

The data consists of thirty-five recordings of various lengths, recorded and transcribed in computational form. The majority are single-speaker narratives of various sorts (stories, anecdotes, and so on) but a minority of texts are interviews or discussions with multiple speakers. The transcription was not in Devanagari, the alphabet used to write Nepali, as there was no computer support for Devanagari in the early 1970s. Instead, a system of transcription was used where a subset of the upper-case Latin alphabet was used to represent the phonemes of Nepali. This meant, for instance, that X was used to represent the sound normally represented in transcriptions of Nepali as *a*. Utterance breaks are indicated in the data; morpheme breaks within each orthographic words are indicated by hyphens. Alongside the transcriptions, some information about each text and speaker has also been preserved.

The Bandhu Collection's value from an ethnographic and linguistic point of view has always been considerable. However it has recently become of even greater potential utility with the advent of the Nepali National Corpus (NNC), developed in the period

[1] Grant reference: British Academy Small Research Grant SG-42148.

[2] Word counts in Nepali are necessarily approximate because of the very large number of enclitics in Nepali, which may or may not be counted as separate words, depending on the current purpose. The estimate of 43,500 words in the Bandhu Collection is base don counting enclitics separately.

2005-2007 by the EU-funded Nelralec / Bhasha Sanchar project[3]. As part of the NNC's design process, standards for machine-readable corpora in Nepali have been developed. The opportunity therefore exists to bring the Bandhu Collection into line with the much larger corpora being developed today, allowing researchers to access and analyse the Bandhu Collection using the tools developed for the NNC, and, furthermore, to make comparisons between the smaller, older dataset and the larger, newer dataset.

Redigitisation

The first task was to make the Bandhu Collection available in digital form again, in a format compatible with contemporary standards for electronic documents.

The fastest way to digitise text with relatively clear and regular letter-shapes – such as computer printouts, which is the form in which the Bandhu Collection survives – is by scanning the text and using optical character recognition (OCR) software, followed by careful post-editing to amend incorrectly-read characters. This was the first approach attempted in the course of re-transcribing the texts. However, the quality of the hard-copy reproduction of the Bandhu Collection to which this project had access was not sufficiently high for this to be viable. This was due to interference in character recognition from such factors as: horizontal rules on the paper the texts were originally printed out on; blurring and darkening from repeated photocopying; and, most significantly, a ghost-image on each page of the text of the following page (picked up in photocopying due to semi-translucent paper). The state of the data is exemplified in

Since scanning and post-editing was not practicable, we instead re-typed the dataset. For quality control, this was done in two stages. First, a non-Nepali speaker with no knowledge of the transliteration scheme typed the texts. This was necessarily done on a letter-by-letter basis as the analyst had no means of recognising the words or understanding any of the meaning. Secondly, another analyst with a basic knowledge of Nepali grammar and an acquaintance with the transcription system re-examined the texts on a word-by-word basis, comparing each line of the typed-up texts to the corresponding line in the original hard-copy, to spot errors of typing and amend them to match the original.

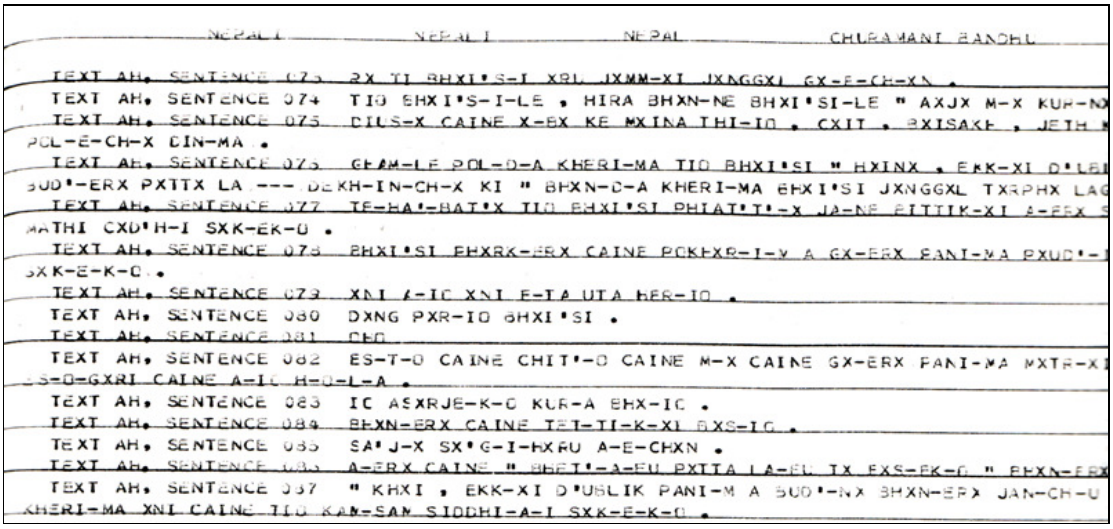
This approach is highly advantageous because the mistakes likely to be made by a reader/typist unfamiliar with Nepali, and the mistakes likely to be made by a reader/typist familiar with Nepali, are very different. The first analyst made very few mistakes, on the whole (around 350 typing errors through the whole dataset), and these tended to involve the selection of an incorrect choice from a pair of similar letters such as B and E, K and X, or I and T; these pairs of letters could be at times indistinguishable in the less well-reserved parts of the hard-copy. Since the second analyst was able to pronounce the words and had an awareness of Nepali phonotactics and grammar, most such errors were easily spotted. In particular the three most problematic pairs previously were vowel-consonant pairs, making it relatively easy to spot when the wrong one had been transcribed. Grammatical knowledge also helped –

[3] See http://www.bhashasanchar.org .

for instance, –ERX is a verbal suffix, but –ERK (which was typed by the first analyst instead of –ERX on a handful of occasions) is not.

This dual-filter approach means that the initial transcription has a very high degree of faithfulness to the hard-copy from which it was derived. While some human errors are still present – almost inevitably give the size of the dataset – the transcription procedure adopted minimised their incidence.

The readability of some areas of the text has been damaged significantly by repeated photocopying of the printed source (see Fig. 1). However, no more than 30 words in the entire corpus were affected to the extent that no plausible reading of the blurred text can be deduced. These have been rendered as sequences of question marks in the transcription, which can be easily located and replaced in the future if a less distorted hard-copy of the original data becomes available.



```
               NEPALI               NEPALI              NEPAL              CHURAMANI BANDHU

   TEXT AH, SENTENCE 073  BX TI BHXI'S-I XRU JXMM-XI JXNGGXI GX-E-CH-XN .
   TEXT AH, SENTENCE 074  TIO BHXI'S-I-LE , HIRA BHXN-NE BHXI'SI-LE " AXJX M-X KUR-NX
   TEXT AH, SENTENCE 075  DIUS-X CAINE X-BX KE MXINA THI-IO , CXIT , BXISAKH , JETH
POL-E-CH-X DIN-MA .
   TEXT AH, SENTENCE 076  GHAM-LE POL-O-A KHERI-MA TIO BHXI'SI " HXINX , EKK-XI D'UBL
BUD'-ERX PXTTX LA --- DEKH-IN-CH-X KI " BHXN-D-A KHERI-MA BHXI'SI JXNGGXL TXRPHX LAG
   TEXT AH, SENTENCE 077  TE-HA'-BAT'X TIO BHXI'SI PHIAT'T'-X JA-NE BITTIK-XI A-ERX S
MATHI CXD'H-I SXK-EK-O .
   TEXT AH, SENTENCE 078  BHXI'SI BHXRK-ERX CAINE POKHXR-I-V A GX-ERX PANI-MA PXUD'-
SXK-E-K-O .
   TEXT AH, SENTENCE 079  XNI A-IC XNI E-TA UTA HER-IO .
   TEXT AH, SENTENCE 080  DXNG PXR-IO BHXI'SI .
   TEXT AH, SENTENCE 081  CHO
   TEXT AH, SENTENCE 082  ES-T-O CAINE CHIT'-O CAINE M-X CAINE GX-ERX PANI-MA MXTR-XI
BS-O-GXRI CAINE A-IC H-O-L-A .
   TEXT AH, SENTENCE 083  IC ASXRJE-K-O KUR-A BHX-IC .
   TEXT AH, SENTENCE 084  BHXN-ERX CAINE TET-TI-K-XI BXS-IC .
   TEXT AH, SENTENCE 085  SA'J-X SX'G-I-HXRU A-E-CHXN .
   TEXT AH, SENTENCE 086  A-ERX CAINE " BHET'-A-EU PXTTA LA-EU TX EXS-EK-O " BHXN-ERX
   TEXT AH, SENTENCE 087  " KHXI , EKK-XI D'UBLIK PANI-M A BUD'-NX BHXN-ERX JAN-CH-U
KHERI-MA XNI CAINE TIO KAM-SAM SIDDHI-A-I SXK-E-K-O .
```

Fig. 1. Deterioration of text readability in the source document.

Reformatting

The original format of the data was as follows:

```
TEXT AA , SENTENCE 001 /BN/ NAU’ KE H–O .
TEXT AA , SENTENCE 002 /SR/ M–E–R–O NAM SALIGRAM .
TEXT AA , SENTENCE 003 /BN/ SALIGRAM H–O KI ?
TEXT AA , SENTENCE 003 JAT KAMI H–O KI ?
...
```

The text identifier and sentence number within the text are indicated by the leading index on each line of the text. Changes of speaker (all sentence-initial in the example above, but sometimes sentence-medial elsewhere in the data) are indicated by placing the ID code of the new speaker within /forward slashes/. While unambiguous for a human reader, this has the disadvantage that structural markup and text are not clearly separated.

For this reason, the first processing step applied to the redigitised data was to convert the notation of texts and speakers into XML, the most widely-used contemporary standard for textual markup. Rather than mark each line, in the reformatted data opening and closing tags enclose each entire text. Changes of speaker are indicated by (empty) *who* elements, as follows:

```
<bandhuCollection>
<text_id="AA">
<who_id="BN"/> NAU' KE H-O .
<who_id="SR"/> M-E-R-O NAM SALIGRAM .
<who_id="BN"/> SALIGRAM H-O KI ?
JAT KAMI H-O KI ?
...
</text>
...
</bandhuCollection>
```

Each text within the Bandhu Collection, and the dataset as a whole, thus constitute a valid XML document and can be processed as such.

Rendering the text as Devanagari

Although the redigitised data as described so far represents a suitable near-verbatim recreation of the Bandhu Collection with allowances for modern encoding standards such as XML, it has a serious drawback. The computer cannot make any direct comparison between the recreated Bandhu Collection texts and Nepali data created in digital form in the last few years, as the latter is very likely to have been encoded as Devanagari text in the Unicode character set.[4] To allow such comparisons, a procedure was developed whereby each word can be mapped to a Devanagari format identical to the format that the same word has in the Nepali National Corpus (NNC).

The problem here is that the transcription system used in the Bandhu Collection under-specifies a large proportion of Nepali words. That is, it does not contain all the information required to deduce how those words would be written in Devanagari. One simple example of this concerns the vowels *i* and *u*. From its basis in the phonology of Old Indo-Aryan, especially Sanskrit, Devanagari possesses long and short forms of each of these vowels, respectively short इ *i* and उ *u*, and long ई *ī* and ऊ *ū*. However, vowel length is not distinctive in Nepali, so these pairs of letters are pronounced identically. Nevertheless there are some spelling conventions regarding which of each pair to use in different contexts: for instance the passive suffix is usually written as short *i* but the phonologically identical feminine adjective agreement suffix is usually written as long *ī* (although Nepali spelling is not fully standardised). In the Bandhu Collection's transcription scheme, only one letter is used, I – that is, the transcription follows the phonology and not the writing system. So for any given word with I or U in it in the Bandhu Collection, it is unknown whether it corresponds to the long or the short vowel in written Nepali. Similar considerations apply to other letters and letter combinations.

To get around this problem, the automatic process that produced the Devanagari version of the dataset was set up in such a way as to produce an array of possible

[4] See www.unicode.org .

Devanagari forms for each word. So, for instance, a word with both I and U in it would have four possible Devanagari forms automatically generated for it: one with both vowels written as short, one with I short and U long, one with U short and I long, and one with both vowels long. Then, the system looked up each candidate in a word frequency list derived from the entirety of the NNC. The most frequently-occurring of the candidates was selected.

In the process of rendering the text in Devanagari, much spelling variation in the original data was eliminated. For example, the third person singular present tense form of the verb *hunu*, "be", is written in Nepali as छ, i.e. *cha*. However, in the Bandhu Collection, this morpheme is represented as *either* –CH–A *or* –C–A (and likewise with other agreement forms of present-tense *hunu*). Such variation has been removed in the process of transferring the dataset to Devanagari. Since this variation was presumably inserted intentionally to indicate pronunciation variants, this represents a loss of information. However, this is balanced by the advantage of enabling the computer to get an accurate frequency count for the morpheme and to link that result to the corresponding frequency in the Nepali National Corpus. Furthermore, the variation is not ultimately lost, as it is preserved in the Latin-alphabet form of the Bandhu Collection, as well as the XML markup of the Devanagari version.

Likewise, the transliteration process does not preserve the morpheme breaks (indicated by hyphens in the original text). This is because the NNC does not represent morpheme breaks, except for enclitic postpositions and numeral classifiers, which are treated as separate words. To match the NNC's encoding, the morpheme breaks were removed from the Bandhu Collection and enclitics also separated out where appropriate, but again the information is retained in the markup.

Let us now consider an example of the process of generating variants and selecting a Devanagari form. For the word transcribed in the Bandhu Collection as KAMI, two forms were generated, KAMI and KAMi (lowercase "i" was used system-internally to represent the vowel written as long). These transliterate to, respectively, Devanagari कामि *kāmi* and कामी *kāmī*. In the NNC, *kāmi* occurs once, but *kāmī* occurs 177 times. Therefore कामी *kāmī* with the long vowel symbol was selected as the appropriate Devanagari form.

When a word contains multiple instances of I or U, or instances of other transliteration ambiguities, the number of potential Devanagari forms can become quite large. The word NIKL-I-RXN-C-XN exemplifies this. The list of options below shows the internal representation of each variation of NIKL-I-RXN-C-XN, together with the automatically-generated Devanagari equivalent and the frequency of each variant in the NNC. A similar table was generated for each distinct word-form in the Bandhu Collection.

NIKL-I-RXN-C-XN [2]	NIKLIRXNCXN=>निक्लिरन्चन् [0]
	NIKLIRXNCHXN=>निक्लिरन्छन् [0]
	NIKLIRXNCXNX=>निक्लिरन्चन [0]
	NIKLIRXNCHXNX=>निक्लिरन्छन [0]
	NiKLIRXNCXN=>नीक्लिरन्चन् [0]
	NIKLiRXNCXN=>निक्लीरन्चन् [0]
	NiKLIRXNCHXN=>नीक्लिरन्छन् [0]
	NIKLiRXNCHXN=>निक्लीरन्छन् [0]
	NiKLIRXNCXNX=>नीक्लिरन्चन [0]
	NIKLiRXNCXNX=>निक्लीरन्चन [0]
	NiKLIRXNCHXNX=>नीक्लिरन्छन [0]
	NIKLiRXNCHXNX=>निक्लीरन्छन [0]
	NiKLiRXNCXN=>नीक्लीरन्चन् [0]
	NiKLiRXNCHXN=>नीक्लीरन्छन् [0]
	NiKLiRXNCXNX=>नीक्लीरन्चन [0]
	NiKLiRXNCHXNX=>नीक्लीरन्छन [0]

This case includes an example of –C– being replaced by –CH– to standardise the spelling of the present tense of *hunu*, as described above, as well as two instances of I. It is important to note that the version with –C– is still retained on the list of options, so if the version with –C– was predominant in the NNC, it would be selected over the regularised version with –CH–.[5] In other words, what is actually in the NNC has a greater weight in the selection than any preconceptions on the part of the analyst about what forms are normal, typical, or correct.

For some words even more potential forms were generated. For instance, for DI-IA-THI-IO, ninety-two possible Devanagari forms were generated, none of which were found in the NNC.

In cases like NIKL-I-RXN-C-XN, where none of the optional forms are found in the NNC, a fallback procedure was required. Fortunately, it was not necessary to achieve 100% correctness of transliteration here – since the words were not in the NNC, a faulty transliteration would not interfere with a comparative analysis. It was therefore deemed acceptable for less rigorous criteria to be employed.

There were around 1,800 such word-forms, where no instance of the word was found in the NNC (compared to around 6,000 where only one possible Devanagari equivalent was generated, or NNC frequencies allowed a decision to be made by the transliteration program). Of these, the majority occur only once of twice in the

[5] In this case, none of the generated forms were found in the NNC, and the heuristics described below were applied to select a form.

Bandhu Collection. These were resolved by means of a number of heuristics: accepting any form that was the only suggested transliteration for the word even if it did not occur in the NNC, looking in a paper dictionary,[6] preferring the short *i* and *u* over the long *ū* and *ī* where the NNC did not provide an example of the latter, and so on. Some of these heuristics were implemented in the transliteration software, others were applied manually (such as looking in a dictionary). In most cases, one of the automatically generated options was selected. In other cases, a different transliteration was inserted in place of all the options.

A knowledge of Nepali grammar was important for the manual part of this step in the analysis. For instance, Nepali words which end in a consonant are written differently depending on whether they are nouns or verbs. Devanagari consonant characters which are not followed by a vowel are taken to include an inherent vowel, usually transliterated as *a*: so, the single consonant character न represents the syllable *na*. A *halant* diacritic must be added to cancel this vowel: so न् represents the consonant *n*. However, by Nepali convention, *halant* is only written at the end of a word if it is a verb. NIKL-I-RXN-C-XN is a verb (the ending –C–XN is a form of *hunu*, "be") and therefore the options ending in *halant* are correct and the ones ending in न alone are incorrect. This particular heuristic was applied part-manually, part-automatically: all instances of verbs ending in a consonant were dealt with manually, and the remaining instances (the non-verbs) were assigned a form *without* a final *halant* automatically.

Sometimes it was possible to find another inflected form of the root which *was* represented in the Nepali National Corpus. For example, no equivalent of NIKL-I-RXN-C-XN occurs in the NNC but निक्लेर, *niklera*, the equivalent of NIKLERX does. It was therefore possible to deduce that the <I> in the root of NIKL-I-RXN-C-XN is short, not long. Taking this and the other previously mentioned criteria together, the form निक्लिरन्छन्, *nikliranchan*, was selected as the appropriate transliteration for NIKL-I-RXN-C-XN.

Of course, it is possible that some of these last-resort decisions were incorrect and led to the incorporation of forms that are non-native-like. This deficiency in the final Devanagari output is ameliorated by two points. Firstly, as noted above, these are words which never occur in the NNC, and so their presence cannot interfere with any comparative analysis. Secondly, an explicit list of all these decisions is retained in the form of the "override list" resource used by the transliteration program, the application of which is the final processing step. So if and when any such incorrect forms are identified, it will be straightforward to amend the override list and re-run the transliteration program. However, on the other hand the process by which words like NIKL-I-RXN-C-XN have been resolved means that it is impossible to claim that the Bandhu Collection's Devanagari form represents *the orthography that would be produced by a native speaker of Nepali*. Rather, the more limited claim is made that the Devanagari form is an *NNC-compatible orthographic rendering* of the original transcription.

The final output has the following format:

[6] The main reference source used in this process was Schmidt (1993).

```
<bandhuCollection>
<text id="AA">
<who id="SR"/>
<tok orig="BXN-A-IN-CH-X">बनाइन्छ</tok>
<tok orig=".">।</tok>
...
<tok orig="KUR-A-HXRU">कुरा हरू</tok>
<tok orig="BXN-AU'-NX">बनाउँन</tok>
<tok orig="JAN-IA-CH-XN">जान्याछन्</tok>
...
```

Metadata

The original Bandhu Collection in printed form included a typed page of metadata with information on the texts and their speakers. This was also digitised, and reformatted into XML to create a file of stand-off metadata, archived and distributed along with the original-format and Devanagari-format corpus files.

The following excerpts from the metadata file exemplify, respectively, the layout of items of text metadata and the layout of items of speaker metadata.

```
<textInfo id="AF">
      <textType>Interview</textType>
      <topic>How to make shoes</topic>
      <speakerList>BN HR GN A RB</speakerList>
</textInfo>

<speakerInfo id="YN">
      <name>Y.N. Regmi</name>
      <sex>M</sex>
      <age>48</age>
      <caste>Bahun</caste>
      <educationLevel>none</educationLevel>
      <village>Chirtung</village>
      <district>Palpa</district>
</speakerInfo>
```

The original format of the metadata had some human-readable shorthand notations and interlinks which, however, resulted in some of the information being implicit rather than explicit – and thus not easy for a computer to handle. These shorthand conventions were removed in the creation of the XML metadata.

For example, some speakers were not originally assigned a shorthand code (the XML “id”); e.g. if all the utterances in a text were by a single speaker, this was not indicated in the text, but left implicit. During the reformatting to XML, all such speakers were assigned an ID code, and appropriate markers inserted into the texts where they spoke.

Summary and conclusion

The process of digitising the Bandhu Collection data was straightforward, if painstaking, as was the creation of appropriate metadata documenting the resource. However, since present-day Nepali corpus linguistics is based, technologically, on Devanagari Unicode text, as exemplified by the Nepali National Corpus, a direct digitisation of the Collection is not easily exploitable.

The creation of a Devanagari-encoded version of the Bandhu Collection comparable with the NNC was a much greater technical challenge, as documented above. However, this second version is extremely valuable, as comparisons between genre-differentiated sections of the NNC and the spoken narratives in the Bandhu Collection may shed light on the grammatical factors which differentiate different text types and text modalities in the Nepali language.

Reference

Schmidt, RL (ed.) (1993) *A practical dictionary of Modern Nepali*. Delhi: Ratna Sagar.