Building a Corpus:
An Experience of the Nepali National Corpus (NNC)

Krishna Prasad Chalise[1]

Abstract

Nepali National Corpus consists of four parts:

- Written corpus
- Spoken corpus
- Parallel corpus

Written Corpus

The written corpus is further divided into 'core corpus' and 'general corpus'.

*Core Corpus***.** Texts that concur with the date, number and genres of FLOB and FROWN corpora have been collected for this corpus. 500 texts of 15 different genres with 2000 words each published in 1991 or so have been planned to be collected in the core corpus. Of these texts 400 texts have so far been completed. These texts have been adapted to XML mark-up for the purpose of computer processing and words have been tagged with their parts of speech (POS). For this purpose a set of 112 tags has devised. This core corpus is the training corpus for the management of general corpus to be collected.
*General Corpus***.** The general corpus of written texts contains about 14 million words. It includes the digitised (machine-readable) texts available in internet webs and also obtained from authors, publishers, etc. These texts have been automatically annotated for parts of speech and other useful information with the help of computer programs. A tool named 'Font Converter', which converts non-unicode fonts such as Kantipur, Preeti, Jag Himali, etc. into Unicode, has been developed for this purpose.

*Parallel Corpus***.** Nepali-English Parallel Corpus (NEPC) is another major component of NNC. It includes equal amount of texts translated from Nepali into English and vice versa. NEPC will be very much useful in the development of machine translation system and language teaching purposes. It can also be used to develop a practical bilingual dictionary. The initial target in this collection is 100, 000 words. But it is likely that it might be increased up to 1 million words if possible.

Spoken Corpus

Nepali Spoken Corpus (NSC) is one of the activities within Nelralec (Bhasha Sanchar) project. The design of NSC is based on Goteborg Spoken Language Corpus (GSLC). The

[1] *e-mail*: krishna40e@yahoo.com

data are taken from spoken Nepali used in different social activities. The basic assumption of the NSC is that the spoken language differs from written language and it has also different genres as in written language.

NSC contains audio and video recordings from different social activities within the natural setting as much as possible with phonological transcribed and annotated texts, and information about the participants. It is open on its size and number of activities as GSLC but the target is up to half a million words. Twenty three social activities are recorded and transcribed. Total word collected are 2,60,000. NSC has ended on December, 2006.