Documentation on the Nepali National Corpus (NNC)

Version of the document

Author(s)	ELDA
Address	55-57, rue Brillat-Savarin 75013 Paris, France
Date	11/12/2013
Version	1.0

Abstract

The present document describes the Nepali National Corpus produced in 2006 in the framework of the project Bhasha Sanchar (“language communication”), also known as Nelralec, for Nepali Language Resources and Localization for Education and Communication; funded by the EU Asia IT&C programme, reference number ASIE/2004/091-777.

The current version of the Nepali National Corpus (NNC) has been produced in 2006 (Yadava et al., 2008). The present document describes the corpus distributed by ELRA, which includes a monolingual written corpus in Nepali, a spoken corpus in Nepali and an English-Nepali parallel corpus.

Table of content

Introduction

Corpus resources have been developed for Nepali for the first time in the form of the Nepali National Corpus, developed by the Bhasha Sanchar ("language communication") project, also known as Nelralec (Nepali Language Resources and Localization for Education and Communication). The project was funded by the EU Asia IT&C programme, reference number ASIE/2004/091-777.
Partners involved in the project were:

- The Open University, UK (coordinator),
- Madan Puraskar Pustakalaya (the archive library for Nepali), Nepal,
- Central Department of Linguistics at Tribhuvan University, Nepal,
- Lancaster University, UK,
- Göteberg University, Sweden,
- the European Languages Resource Association (ELRA), France.

The current version of the Nepali National Corpus (NNC) has been produced in 2006 (Yadava et al., 2008). The present document describes the corpus distributed by ELRA, which includes a monolingual written corpus in Nepali, a spoken corpus in Nepali and an English-Nepali parallel corpus. Each part of the corpus is described below.

1. The Nepali Monolingual written corpus

The Nepali Monolingual written corpus comprises the core corpus (core sample) and the general corpus. While the core sample (CS) represents the collection of Nepali written texts from 15 different genres with 2000 words each published between 1990 and 1992, the general corpus (GC) consists of written texts collected opportunistically from a wide range of sources such as the internet webs, newspapers, books, publishers and authors.

Given below in tables are the details about the collections with respect to the genres:

Table 1: Core sample framework (based on FLOB/FROWN corpora)

Category Label	Category Title	Number of samples
A	Press: Reportage	44
B	Press: Editorial	27
C	Press: Review	17
D	Religion	17
E	Skills, Trades and Hobbies	38
F	Popular Lore	44
G	Belles Lettres, Biographies, Essays	77
H	Miscellaneous	30
I	Science	80
J	General Fiction	29
K	Mystery and Detective Fiction	24
L	Science fiction	6
M	Adventure and Western	29
N	Romance and Love story	29
O	Humour	9
	TOTAL	500

The primary purpose of the Core Sample was to provide a match to other corpora created from the same sampling frame.

However, there were made some adaptations for selecting genres as all the genres existing in English writings (e.g. science fiction) did not exist in Nepali because of cultural and other East-West differences. Besides, only 398 (instead of 500) texts could be collected for Nepali core corpus since texts from some genres could not be available from the 1991/92 time frame when writings in Nepali were very much restricted and just started broadening with the advent of liberalism after the restoration of democracy in the country.

The collected core corpus is presented in Table 2.

Table 2: Core sample framework (based on FLOB/FROWN corpora)

Category name	**No of files**	**No of words**
A (Press reportage)	33 (44)	66800
B (Press editorial)	23 (27)	46520
C (Press review)	6 (17)	12095
D (Religion)	13 (17)	26412
E (Skills, Trades and Hobbies)	29 (38)	58935
F (Popular lore)	32 (44)	64878
G (Belles Letters, Biographies, Essays)	68 (77)	137873
H (Miscellaneous)	28 (30)	56680
J (Science)	56 (80)	113507
S (Fiction)	110 (126)	220874
Grand total	**398 (500)**	**804574**

The internal structure of the core corpus is as follows:

Press editorial		
Daily:		5
From Kathmandu:	6	
From Outside:	-	
Weekly:		17
From Kathmandu	15	
From Outside:	-	
Half-weekly:		1
From Kathmandu:	-	
From Outside:	1	
Total		23

Religion	
From Book	10
One text translated from Hindi and one text based on Sanskrit	
From article:	3
Total	13
Fiction	
Novel	66
Short story	44
Total	110
Science	
From book	37
From periodicals	19
Total	56
Sub-category	
Science and technology	3
Criticism	20
Anthropology / culture	8
History / Archeology	5
Language and grammar	3
Law / politics	5
Psychology	1
Philosophy	4
Business / economics / administration	6
Unclassified	1

The General Collection (GC) was collected from a wide range of sources such as websites, newspapers, and books, with non-internet texts being gathered in the main directly from publishers and authors. This part of the corpus was intended to allow corpus analyses that depend on a very large corpus.

The written corpus is morphogically-annotated. A parts-of-speech (POS) tagset has been produced within the project : the Nelralec Tagset, described in (Hardie et al., 2005). This is a categorisation system for the manual and automated analysis of morphosyntactic units in Nepali. Tokenisation and lemmatisation for Nepali are further discussed in (Hardie et al., 2011).

2. The Nepali Spoken Corpus

The spoken corpora have been collected from 17 social activities in their natural settings and contain about 260,000 words. These texts are audio-video recordings of the activities with their phonological transcriptions and annotations about the participants but their audio-visual materials can later be transcribed and used for analyzing their paralinguistic and extralinguistic features. The main purpose of this collection is to compare it with the written texts and identify their differences.

The design of Nepali Spoken Corpus (NSC) is based on Goteborg Spoken Language Corpus (GSLC). The 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 recordings from different social activities within the natural setting as much as possible with phonologically transcribed and annotated texts, and information about the participants. Each activity is stored in three files (recording as such in .mpeg, transcription in .txt and recording information in .doc). The total number of words transcribed from these recordings are assumed to be 260,000.

The description of the Nepali Spoken Corpus is provided in the the table below.

Total No. of	**Today**
Recorded Activity types	17
Recorded Activity occurrences (files)	116
Total time (duration)	31 hours 57 min.
Total transcribed words (assumed)	260 000
Total transcribed files	116
Completely checked	0

As can be seen from the table, 116 activity occurrences have been recorded belonging to 17 activity types. For example, the activity type “shopping” has four recorded distinct occurrences and the activity type “discussion” has 15 recorded instances. The total temporal duration of the recorded material is 31 hours and 57 minutes.

3. The English-Nepali Parallel Corpus

The parallel corpus consists of two collections in two genres: national development and computing. Computing texts in both Nepali and English, is about 3 million words, where as national development text is about 966, 203 words.

These corpora can be used as useful resources in developing machine translation system for translating Nepali texts into English and vice versa. They can also be helpful in preparing a Nepali-English/English-Nepali bilingual dictionary, contrastive studies and devising teaching materials for language teaching.

The current release only contains the national development texts. The computing text will be made available as soon as the copyright issues will be solved.

The current release includes:

- texts manually aligned at the sentence level (27,060 English words; 21,756 Nepali words), provided in the tmx format (xml file).
- texts aligned at the document level (617,340 English words; 596,571 Nepali words), provided in raw text and in the original word processing format,
- an additional set of monolingual data in Nepali (386,879 words in Nepali), provided in raw text and in the original word processing format.

References

Yadava, Yogendra P.; Hardie, Andrew; Lohani Ram Raj; Regmi Bhim N.; Gurung, Srishtee; Gurung, Amar; McEnery, Tony; Allwood, Jens; and Hall, Pat. (2008). “Construction and annotation of a corpus of contemporary Nepali”. Corpora 3(2): 213-225.

Hardie, A, Lohani, R, Regmi, B and Yadava, Y (2005). “Categorisation for automated morphosyntactic analysis of Nepali: introducing the Nelralec Tagset (NT-01)”. Nelralec/Bhasha Sanchar Working Paper 2.

Hardie, Lohani, and Yadava (2011). “Extending corpus annotation of Nepali”. Himalayan Linguistics ISSN 1544-7502, Vol 10(1): 151-165.