Nepali Lexicon

Laxmi Prasad Khatiwada, Srishtee Gurung
laxmi@mpp.org.np,srishtee@mpp.org.np

Abstract

This document mainly focuses on the development process of building a lexicon for Nepali Language. It discusses the problems encountered while selecting the entries for lexicon. Various conditions on how words such as homographs, reduplicated words, foreign words, etc are considered in Nepali language context, is described in detail.

1. Introduction

A lexicon is a list of words together with additonal word-specific infromation as in a dictionary. The study of lexicon involves the research on how the vocabulary of a language is structured, the way words are used and stored, the methods of learning words, the origin of words, relationships between words and so on[2].

There were various problems encountered during the analysis and development of Nepali lexicon. The following topics describe the actual problems encountered in various stages. It also suggests a hypothetical approach for solving these respective problems during the building of Nepali lexicon.

1.1. Current State of Nepali lexicon

In the current scenario of Nepal there are a large variety of paper based dictionaries for Nepali language available in the local market. Ironically, electronic version of Nepali lexicon does not exist till date. In order to fulfil the requirement of Nepali users for using Nepali language in computers, an electronic lexicon needs to be developed. NLP Systems like Spell Checking, Syntax Checking, Machine Translation and others require a lexicon to do their proper functioning. In order to meet the above criteria in various NLP systems, a generalized lexicon is a necessity.

1.2. Material Resources

In the current stage, there is a lack of analysis tools for Nepali Language. Ready made corpus of Nepali texts is not available yet. In other words, there are no adequate tools for counting the frequency of words, analysing collocations, concordance of words, etc. Therefore, we had to rely on the paper materials for resources. These resource materials are the authentic dictionaries. And all the selection of lexicon entries is done manually from the dictionaries. A Nepali national corpus is being developed in Madan Puraskar Pustakalaya. With the help of this national corpus, frequency count of Nepali words in different context, will be possible. Other analysis such as collocation and concordance of word could be discovered.

1.3. Lexical Entries

The main problem with the lexical entries is the order for which the entries were to be inserted in the lexicon. Different dictionaries use different collation sequence. This resulted in a problem of having more than one collation sequences for Nepali language. This problem arose because there is no standard collation sequence for Nepali Language yet. However a proposal for a standard collation sequence could be given

Other criteria that were considered during the entries include compound words, homographs, foreign words, POS, reduplicated words and onomatopoeic words. Each of these criteria is discussed in detail in section 3.

1.4. Format of lexicon

The Nepali lexicon is stored both in text and XML format. The entries for the XML format lexicon is done automatically either by using a LexTool or manually by typing in a file, storing in the same format and then converting it to the XML format by using a conversion tool.

This lexicon is the generalized lexicon for Nepali language. The attributes considered for this lexicon are: Root words, Headwords, Pronunciation, Syllable break, Meaning, Parts of Speech, Synonym and idiom. The format and attributes of this generalized lexicon can be changed to fulfil the requirement of a particular NLP system. More detail about the XML format of lexicon is given in section 3.

2. Discussion

2.1. Current State of Nepali Lexicon

Lexicons are distributed according to several features. These may be subject specific, style specific etc. The Nepali Lexicon intended for development falls under the category of mono-lingual lexicon.

The distribution of this lexicon is shown in appendix A. Till date there are about 11,000 words or entries in the lexicon. Linguist did the analysing part of the language by collecting the entries from the available resources i.e. dictionaries and compiling them; while the typist typed the entries selected by the linguist. We had one linguist and one typist dedicated for this task.

2.2. Material Resource

The following is the manual procedure for the selection of entries or words in the lexicon

- Collection of various Nepali Dictionaries
- Selecting common words from Dictionaries
- Getting proper annotations

Material collection is an important phase of the lexicon development process. It largely depends upon the category of the lexicon. As the intended lexicon is of general type, the initial inventory of words is derived from sources like dictionaries and survey of the vocabularies commonly found in grammatical text books. Especially the dictionaries referred for the word selection are: Brihat Nepali Sabdakosh Practical Nepali Sabdakosh Sajha Sabdakosh and Technical Nepali Sabdakosh

2.3. Lexical Entries

2.3.1. Collation Sequence

Madan Puraskar Pustakalaya has proposed a Nepali Collation sequence for the sorting utility in localizing Linux in Nepali language. This proposal is under consideration for standardization. The same order is used for lexicon entry too. One of the dictionaries which follow the same order is Nepali Brihat Shabdkosh.

Entries in the lexicon will be arranged in the following order shown as below:

ँ, ं, ः, अ, आ, इ, ई, उ, ऊ, ऋ, ए, ऐ, ओ, औ
क, ख, ग, घ, ङ
च, छ, ज, झ, ञ
ट, ठ, ड, ढ, ण
त, थ, द, ध, न
प, फ, ब, भ, म
य, र, ल, व, श, ष, स, ह

2.3.2. Infinitive form

Infinitive form is based on traditional grammar citation, because नु is the highly common factor of TAM (tense, aspect and model) paradigm. यो is the past form of the verb. यो is one element among the paradigm of tense, aspect and model.

2.3.3. Compound words

Proper compound words are taken as a citation form. A group of two or more words with a different and single semantic unit are taken as compound form and are given a different entry.

Such as लामीडाँडा, मामाघर, कुन्जीपाटी, Onomatopoeic words like खत्र्याक्क, झम्ल्याङ्ग etc and reduplicated words having reduplicated form with or without merging orthographic presentation such as कुपुकुपु, किरामिरा, कथाकुथुङ्ग् are all given a single entry.

2.3.4. Homographs

Words with same orthographic representation but different grammatical class with different meaning is cited as separate entry.

नहर: ना. तासको नवौं पत्ती
नहर: ना. पानी बग्ने कुलो

2.3.5. Foreign terms

Foreign terms that common user uses as a Nepali word is listed as headwords, e.g.

बाइन्डर:ना. बाँध्नका लागि प्रयोग गरिने वस्तु
बाइनरी: ना. द्वि-आधारी अङ्क

Some rules have been followed to manage such foreign terms in Nepali.

- **Transliteration** No transliteration is needed among languages sharing the same alphabet, such as German and English. Proper transliteration has been done between English - Nepali terms as reflected in the pronunciation of words.

- **Borrowing.** Borrowing occurs when the target language has no equivalent words for the source language words. In some cases the source language equivalent may be translated using an expression not of the target language but of the adjacent/familiar local language.

- **Place** names may be presented with the spelling in which they are spelled customarily in the target language. However, the names of the organizations may be presented in transliteration.

- **Trade** names in use may be presented as they are. In the translation of the material meant for popularization of science, or meant for giving instructions to consumers, trade names as well as their synonyms in the target language may be indicated.

2.3.6. Synonym

Synonyms have been specified with cross-references. These related words are looked up in thesaurus. To find proper mosaic of lexical field is very difficult. In other word sharp demarcation of lexical field is very confusing within the process of making thesaurus. On the process of thesaurus entry relative value or position of the citation word has been kept consciously.

A sample here is given to show how to determine the difference of synonymic word according to the lexical field, by part of speech (POS) tagging.

- POS and grammatical tag convention is shown in the appendix B.

3. XML Format of lexicons

This XML format is for the generalized Nepali Lexicon. As it is clearly seen through the tags that the tag <category> tells us the entries being inserted in the lexicon is of type *verb*. The tag <attributes _list> lists all the attributes that are being considered currently for the Nepali Lexicon.

<entries> tag lists each entry for the category Verb and the corresponding value for each of the attributes for the word "उकास".

```
<lexicon>
<header>
<language>Nepali</language>
<category>Nepali verb</category>
<attributes_list>
<attribute>Rootword</attribute>
<attribute>Headworrd</attribute>
<attribute>Pronounciation</attribute>
<attribute>Syllablebreak</attribute>
<attribute>Meaning</attribute>
<attribute>Partsofspeech</attribute>
<attribute>synonym</attribute>
<attribute>idiom</attribute>
</attributes_list>
</header>
<entries>
<entry>
<Rootword>उकास्</Rootword>
<Headworrd>उकास</Headworrd>
<Pronounciation>उकास्</Pronounciation>
<Syllablebreak>उ कास्</Syllablebreak>
<Meaning>१.तलतरि रहको जस्तुलाई माथतिरि तान्नु
२.अप्ठ्यारो परेकाबेला छुटकारा दलिाउनु</Meaning>
<Partsofspeech>क्रयिा</Partsofspeech>
</entry>
</entries>
</lexicon>
```

4. Results

There are about 11,000 words or entries in the lexicon. These entries took about 350 days. This concludes that on an average 30 entries were inserted in the lexicon in a day. We had one linguist working full time and one typist working part time for the lexicon entries.

5. Conclusion

The generalized Nepali lexicon may be modified according to the need of a particular NLP system. For example this lexicon is modified according to the framework of Hunspell i.e. the spell checker in Open Office. Therefore the format and attributes of generalized lexicon is change according to the needs of Nepali spell checking in Open Office.

Since there is unavailability of resources for analyzing the language, the generalized Nepali lexicon developed may not be optimal. Even though, it is currently being used for spell checking and thesaurus in Open Office.

6. References

Following are the dictionaries that were referred for the selection of words or entries for the lexicon.

[1] Ghimire M.et al., *Brihad Nepali Sabdakosh*, Rayal Nepal academy, Nepal, 5th Reprint, 2004

[2] http://en.wikipedia.org/wiki/Lexical

[3] Adhikari H., *Prayogatmak Nepali Sabdakosh*, Bidthyarthi Prakashan, Nepal, 2005

[4] Dixit N.A. ,*Sajha Sabdakosh, SajhaPrakashan,* Nepal, 5th Reprint 2001

[5] Bigyan Nepali Sabdakosh, Royal Nepal Academy, Nepal, 2002.

Appendix A

The distribution of Nepali Monolingual Lexicon can be perceived through the following figure:

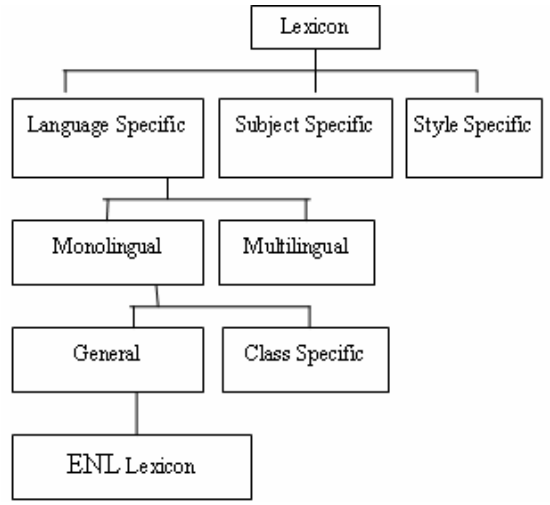

Figure 1: Distribution of Nepali Monolingual Lexicon