Nepali Lexicon Development

[1]Sanat Kumar Bista, [1]Birendra Keshari
[2]Laxmi Prasad Khatiwada, [2]Pawan Chitrakar, [2]Srihtee Gurung

[1] Information and Language Processing Research Lab
Kathmandu University, Dhulikhel, Kavre, Nepal
{sanat,birendra}@ku.edu.np

[2]Madan Puraskar Pustakalaya
Lalitpur, Nepal
laxmi@mpp.org.np

Abstract

This report on Nepali Lexicon Development highlights the process followed in the development of Nepali Lexicon for the Nepal component of the PAN localization project. The scope of this development initiative lies within its use in spell checking system for Nepali. The target is to have a minimum measure of twenty five thousand words with at least four hundred noun headwords, one hundred and fifty verbs, one hundred adverbs and adjective and all the pronouns of Nepali grammar in the entries of lexicon database. This report also describes briefly the tool used in development, and some problems and issues faced during the lexicon development process.

1. Introduction

Lexicon is the heart of any natural language processing systems. All NLP Systems like Spelling Checker, Syntax Checker, Machine Translation and others require a lexicon to work. The format of the lexicons for these systems may differ according to their specific need but the basic concept of Lexicon remains the same for any NLP systems.

A lexicon may be defined as a list of words with additional word-specific information, i.e. a dictionary. The word has a Greek origin, which means vocabulary. The study of linguistics involves several issues like what words are, the structure of the vocabulary in the language, the methods of using and storing words, the way words are learnt, the origin of words, relationships between words , the methods of words creation etc. [2].

Linguistics, however, has a more specialized definition of the lexicon as it deals with lexemes to actualize the words. Lexemes are formed with the help of morpho-syntactic rules. Being constrained on a set of certain principles, say for instance grouping all words belonging to a certain category, the vocabulary of the language is built. Then for generating certain new words (both simple and complex), lexical rules are used. For example, the suffix -able which clearly applies to transitive verbs like “measure” to get a new word, measurable does not necessarily apply to some other verbs like “sing” and “dance” [2]. Two important terminologies worth introducing here are Lexicology and Lexicography. Lexicology is the science of dictionary compilation and Lexicography the art of lexicon building [1]. Some linguistic aspects of concern to a Lexicographer are:

a. Linguistic information such as phonemic, morphological, and semantic analysis.
b. Language, context relative as well as grammatical information of that Language.
c. World knowledge and extra linguistics information such as cultural conventions, date, and currency format etc.

Lexicons are distributed according to several features. It can be language specific, subject specific, style specific etc. The *Nepali Lexicon* intended for development falls under the category of **Language Specific Lexicon** designed to suit those readers who intend to use Nepali Language in writing texts. The following figure categorizes the lexicon distribution and also shows the position of Nepali Lexicon in it.

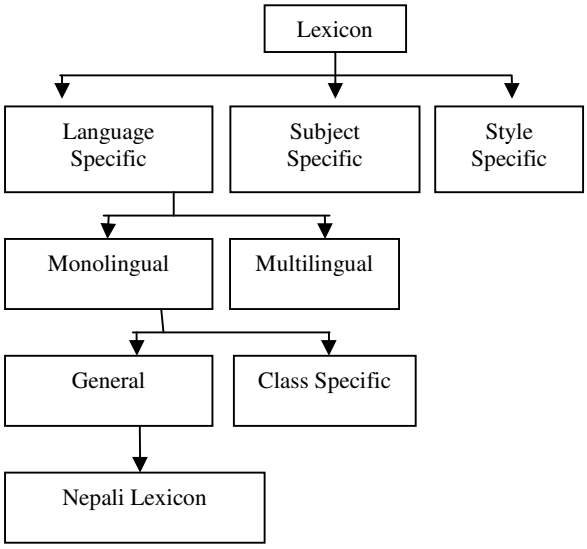

Figure 1: Nepali Lexicon in lexicon distribution

2. Objective and Scope

The main objective of the development of Nepali Lexicon is for its use in Nepali Spell Checking System. The lexicon is supposed to facilitate word level spell checking and is *monolingual* in nature. The lexicon targets to accommodate at least twenty five thousand words and is expected to contain all the words from dictionaries and some recently emerging words from technical fields like information and communication technology that are not enlisted by the existing Nepali dictionaries.

3. Development Process

The process chart given in Figure 2 is an overview of the overall process of Nepali Lexicon Development.

The following specific steps have been followed in the development of Nepali Lexicon:

3.1. Step I

The first step consists of planning, identifying the types, finalizing the source area and the volume to be managed by the author. This first step can further be divided into the following phases:

3.1.1. Planning.

This phase consist of the overall planning for the lexicon development process. Some of the specific activities are choice of using tools for development, frequency count utility, file format choice etc.

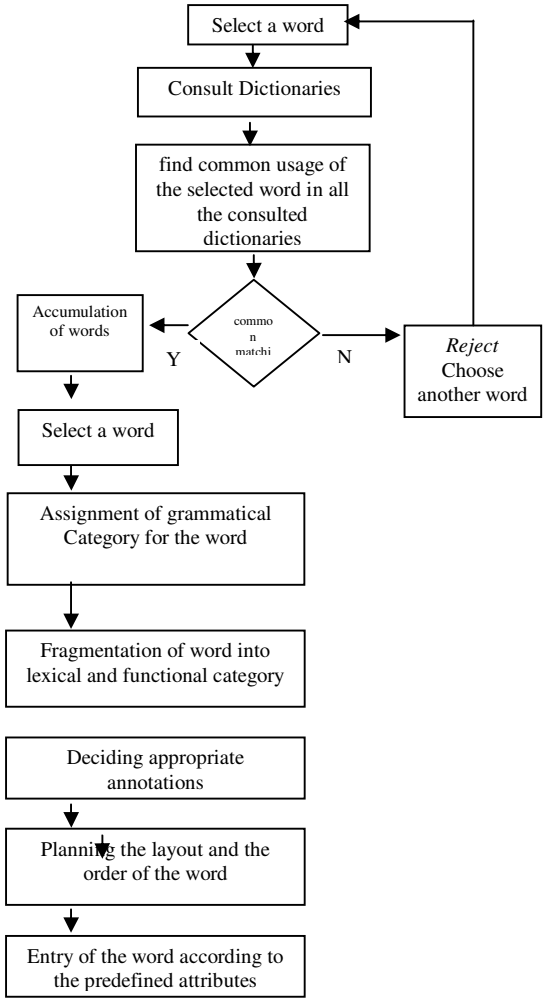

Figure 2: Process Chart for Nepali Lexicon Development

3.1.2. Material Collection and Indexing

This is an important phase of the lexicon development process. Material collection 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. The dictionaries referred for the word selection are: *Brihad Nepali Sabdakosh* [3], *Practical Nepali Sabdakosh* [4], *Sajha*

Sabdakosh [5] and *Technical Nepali Sabdakosh* [6]. Other sources such as fiction texts, newspapers, periodicals etc were also referred in this phase.

Earlier, in absence of a reliable tool for frequency counting of a word occurrence in texts, word selection was done on the basis of linguistic experience of a linguist. Later, a frequency count tool was available and hence was used in including words in lexicon according to the frequency measure of the word.

3.1.3. Entry Selection

This is the lexical phase of lexicon development. After material collection proper citation work is needed. The selection of entry words for citation is directed by the scope and the objective of this lexicon development. Relations of direct, inflection, derivation have been calculated properly to choose the proper head word.

Some words that have been considered are:

a. Technical words
b. Newly coined word
c. Obsolete words
d. Eco-empty words
e. Relational words like prepositions and post position.
f. Prefix suffix and infix
g. Compound Words
h. Idioms

3.2. Step II

In this part entry layout and order are addressed. Headword, lexical unit, grammar, meaning and other lexicon type sensitive features are described properly.

The following lexical attributes have been established for Nepali Lexicon:

a. Root word
b. Head word
c. Pronunciation
d. Syllable break
e. Meaning
f. Parts of Speech
g. Synonym
h. Idiom

3.3. Step III

This phase primarily focuses on lexicon formatting and enables the lexicon with proper shape and sequence. This is the final step of lexicon editing. Before the publication or bounding level (in electronic lexicon) every element should be checked.

In this part the following aspects have been addressed properly.

a. Entry sequence according to the text variety
b. Collation sequence
c. Subject division
d. Index
e. Bibliography

4. LexTool: A Tool developed and Used for Nepali Lexicon development

In order to ease the building and maintenance of Lexicon for Nepali language, LexTool has been developed and used throughout. Though the main aim of this tool is to assist the **development of Nepali Lexicon**, it can be used to develop the lexicon of any other languages because this system is Unicode aware which means this system supports multi-language.

The system stores the lexicon database in a XML file format (Unicode). The main reason for this is that XML is easily portable. The extensibility and the structured nature of XML allows it to be used for communication between different systems, which otherwise would be unable to communicate. The Lexicon in XML format can be exported to different formats according the specific need of the NLP systems.

The system requires manual entries of the stem words. After that it is feed with some morphological rules. The rules are stored as a separate database. The rules can be loaded and then the system can generate some other entries by manipulating the rules. Thus, it provides facility of semi-automatic expansion of the existing lexicon.

The following figure provides an overview of the architecture of LexTool.

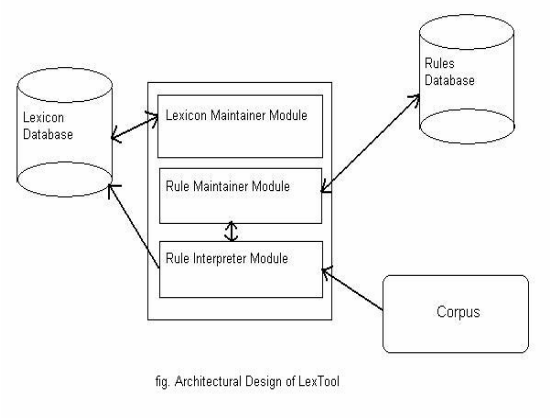

Figure 3: LexTool Architecture

LexTool consists of three major modules; **Lexicon Maintainer Module**, **Rule Maintainer Module** and **Rule Interpreter Module**. These modules interact with **Lexicon Database**, **Rules Database** and **Corpus**. The detailed description of each of these modules and the entity they interact with are given below:

a. Lexicon Database:

This is the lexicon database in XML format that contains several lexical entries. Each entry contains the language head word and all related attributes. There can be separate lexicon databases for different languages. Similarly, same language can have different lexicon database may be each for a specific category of words that share common attributes.

b. Lexicon Maintainer Module:

This module interacts with the **Lexicon Database.** It loads, modifies and saves the **Lexicon Database.**

c. Rules Database:

This database contains the morphological rules that can help in the generation of other entries in the **Lexicon Database.** It is also in XML format.

d. Rule Maintainer Module:

This module interacts with **Rules Database.** It can load the **Rule Database,** modify it and save it.

e. Corpus:

This is the corpus that is to be feed to the **Rule Interpreter Module** for semi-automatic expansion of the existing **Lexicon Database.** The corpuses are generally newspaper, journals and magazines in digital form.

f. Rule Interpreter Module:

This module interprets the rules that is loaded by **Rule Maintainer Module** and reads the **Corpus.** Depending upon the morphological rules in the **Rule Database,** this module generates some entries and stores them in the **Lexicon Database.**

4.1. Nepali Lexicon Development Using LexTool:

Nepali is a morphologically rich language. Both inflectional and derivational morphology are found in the words of this language. This information can be very nicely used in the semi-automatic generation of Nepali Lexicon using **LexTool.**

The morphological information found in this language can be put as rules database in **LexTool**. For example in Nepali language plural forms of nouns are generated by adding *"haruu"* at the end. Such inflectional morphology can be put in the **Rules Database.** While reading the corpus, if the **Rule Interpreter Module** finds a token (word) with ending, *"haruu "it* can detect that the root of the word is a noun and thus it can store the root as noun in **Lexicon Database.** Similarly, other rules can govern the correct determination of the attributes of stem words by morphological analysis. The exceptions found in the language can cause the errors in the lexical entries and they can be manual removed.

5. Results

The output that is currently available is a Nepali Lexicon of approximately five hundred root words. The result is categorized into different grammatical parts such as noun, pronoun, verb, adverb, preposition etc. The lexicon entries have been annotated according to the contemporary linguistics norms.

6. Conclusion

The initial phase of Nepali Lexicon development process suffered a lot due to the unavailability of technical tools to analyze several things like word frequency. Also unavailability of appropriate data of Nepali corpus still remains as a major in the lexicon development process. However, with some tools that have been developed on our own effort we are able to initiate the development process of Nepali Lexicon. The targeted lexicon is of 25,000 head words.

7. References

[1] N. Krishnaswamy, S.K Verma, and M.Nagarajan, *Modern Applied Linguistics*, Mac Millan, India, 2000

[2]] http://www.wikipedia.org

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

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

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

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