PAN Localization Project

Project No: 102042
Ref. No: PANL10n/Admn/PR/0801LK

PAN
Localization
www.PANL10n.net

Research Report on Sinhala WordNet

Country Component	SRI LANKA	**Report no.**	
Phase no.	2.1	**Report Ref no.**	

Prepared By: Dulip Herath
(Name)

Designation: Research Associate

Project Leader: Dr A R Weerasinghe
(Name)

Signature: ______________________

Date: ___________________________

TABLE OF CONTENTS

Sinhala WordNet

Introduction

Words can be treated as the basic building blocks of any natural language as they are the smallest meaningful categories that are ready to be used in sentences or utterances[1]. In general, meaning of a word is a broad concept and there have been a lot of theoretical discussions on this matter and several formalisms have been proposed in Lexical Semantics and Semantics in general to deal with the issues related to word meaning.

WordNet is one of the useful and important lexical resources for many crucial natural language processing and computational linguistic tasks such as Word Sense Disambiguation, Word Similarity, Information Retrieval and Extraction, Machine Translation, etc. It is based on the formalisms developed in Lexical Semantics and defines different senses associated with the meaning of a word (the idea of word senses are described in detail in Section 2) and other well-defined lexical relations such as synonyms, antonyms, hypernym, hyponyms, meronyms and holonyms (each of these relations are described with examples in Section 3).

Typically, words are organized in WordNet with respect to concepts which are directly related to word senses whereas in a traditional dictionary they are organized according to the alphabetical order. This facilitates the human users or computer applications to access not only the words that represent a particular concept or meaning but also words that represent other related concepts. In general, one word can stand for many senses (or concepts) and one sense (or concept) can be represented by many different words. In WordNet context a set of words that represent the same sense or concept is called a *Synonym Set* or *Synset*.

A detail analysis of word senses is given is Section 2 and a detail account of lexical relation are given in Section 3. Section 4 of this report provides a brief description of popular WordNet initiatives such as Princeton WordNet, Euro WordNet and Hindi WordNet. The issues to be dealt with Sinhala in WordNet building are discussed in

[1] In theoretical linguistics, *morpheme* is defined as the smallest meaningful unit. But in this context we refer to word as the smallest meaningful unit for practical purposes.

Section 5. Sections 6, 7 and 8 presents the resources used for, methodology adopted for and the design and implementation issues of Sinhala WordNet development.

Lexemes and Word Senses

In general linguistics, the idea of *lexeme* is defined as an entity that encapsulates an orthographic or phonological form and its meaning. Therefore lexeme is one that is needed to be listed in the *lexicon* as each lexeme represents a separate meaning. Usually, the lexicon consists of only the base forms of lexemes called *lemmas*. The process of obtaining lemmas is called *lemmatization*. Figure 2.1 depicts how meaning and form are related to each other and how lemmatization process maps all the possible word forms to a single base form called lemma.

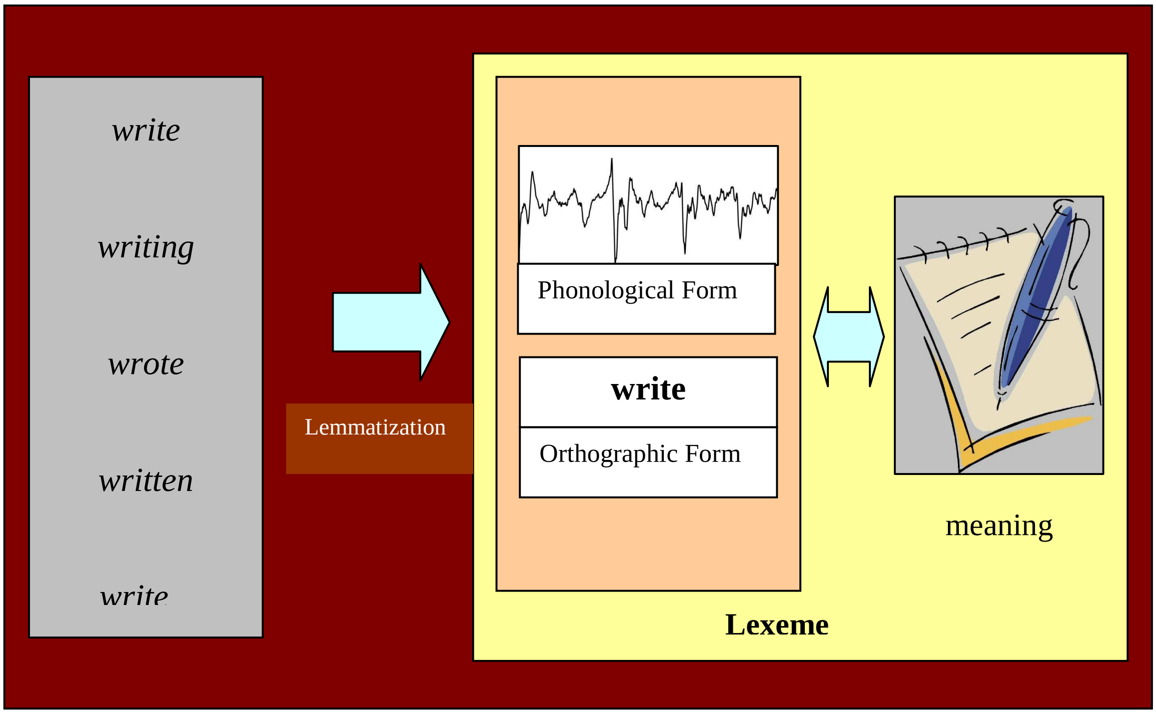

Figure 2.1: Lemmatization of Word Forms and Lexeme

Though the meaning is an integral part of a lexeme, it can vary in different contexts. These contextual variations are called *senses* (or word senses). Each sense represents a different aspect of the meaning of the word being considered.

The most popular English word with many different senses is **bank**. It can stand for either *sloping mound* or *financial institution*. The word **bank** has some other meanings as well, for example it can stand for any kind of repository such as *blood bank, egg bank, job bank*. Further the word **bank** can stand for a building that belongs to a banking organization. In semantics, these sense relationships are defined formally as follows:

Case 1:
bank (*financial institution*) and **bank** (*sloping mound*) are called **homonym** as they share the same orthographic form but not directly related to each other.

Case 2:
bank (*financial institution*) and **bank** (repository) are repositories for things that can be deposited and withdrawn. The semantic relationships of these two senses are structured and systematic and it is called **polysemy**.

Case 3:
bank (*building*) and **bank** (organization) is a subtype of **polysemy** and it is called **metonymy**. **Metonymy** is the use of one aspect of the meaning to refer to anoter aspect of the meaning. Similar relationship can be established between *Author* and *Works*, *Animal* and *Meat*, and *Tree* and *Fruit*.

It is extremely difficult to determine whether two senses are *homonymous* or *polysemous* and to determine how many senses word has and if so, to distinguish them. Some heuristics such as *independent truth conditions*, *different syntactic behavior, independent sense relations* or *exhibition of antagonistic meanings* can be utilized to determine polysemy of words.

Sense Relations

Current WordNet applications define ontological relations of word senses that are extremely useful natural language processing and computational linguistic tasks. Some of the main relations are described with example in this section.

Synonyms (Synsets)

If two senses of two different words are identical or nearly identical they are defined as synonyms. They are all supposed to refer to the same concept and therefore the set that consists of all the words that are synonymous with each other is called a *Synonym Set* or in short *Synset*. Practically, it is hard to find two words that are absolutely synonymous with each other and therefore several intuitive approaches are taken to find members of a synset. Example synsets are *car/automobile*, *water/*H_2O, *couch/ sofa* or *vomit/throw up*.

Antonyms

Antonyms are the words that have opposite meanings. It is also difficult to give a formal definition of antonyms. But roughly it can be said that two word senses are antonyms if they define *binary opposites* or are at *opposite ends of some scale*. For example *long/short*, *fast/slow* and *big/little* are extreme ends of length or size scale. *Reversives* is another category of antonyms. They describe movements and changes in opposite directions such as *rise/fall* or *up/down*. binary opposites:

Hyponyms

A word can be a hyponym of another if that word has a sense that denotes a more specific sense of a sense that is denoted by the other word. For example *car* is a hyponym of *vehicle*, *dog* is a hyponym of *animal* and *mango* is a hyponym of *fruit*. Therefore hyponym can be defined as a *kind-of* relation or *specialization*.

Hypernyms

Hypernym is the converse of hyponym which means it denotes a more general sense of another sense such as *vehicle* is a hypernym of *car*, *animal* is a hypernym of *dog* and *fruit* is a hypernym of *mango*. In other words, hypernym can be defined as a is-*a* relation or *generalization*.

Meronyms

A word sense is a meronym of another word sense if it denotes a part of the other word sense. Therefore it represents a *part-of* relation. One common example is *leg* is a part of *chair* hence *leg* is a meronym of *chair*.

Holonyms

Holonym is the inverse relation of meronym and it denotes a word sense that has another word sense as one of its parts. Therefore it represents a *has-a-part-called* relation. Therefore *chair* has a part called *leg* hence *chair* is a holonym of *leg*.

WordNet Initiatives

Several initiatives have been taken towards building WordNets for different languages such as English, European Languages, Hindi, and CJK languages. The research groups involved in developing consists of not only computer scientists but also linguists and psychologists. A brief overview of three prominent WordNet projects namely Princeton English WordNet (Fellbaum, 1998), Euro WordNet (Vossen, 2002), and Hindi WordNet (Narayan et al, 2002), (Chakrabarti and Bhattacharyya, 2004) were closely examined as a part of the Sinhala WordNet development project to understand the approachs taken, structure, language specific issues and the functionalities available in them are concisely described.

Available Resources

A survey of potential resources for the Sinhala WordNet project was carried out at the beginning of the project. As a result of this survey, it was found that the tradition of thesaurus building is not completely new to Sinhala language studies and in general it is a well established idea in traditional linguistic studies in ancient India.

Though there are some Sinhala language resources available in traditional Sinhala literature which are closer to the current work, it cannot be benefited directly from many those resources due to several reasons namely, *poor coverage of modern Sinhala* (mainly covers traditional ancient language) and *poverty of concept classification* (confined to religious and preliminary concepts). Having examined them thoroughly one main resource and a couple of supplementary resources were identified as preliminary sources of the project. They are as follows:

- ***Maha Sinhala Sabdakoshaya* by Dr Harishchandra Wijetunge** (Main Source)
- Sinhala-English Dictionary by Charles Carter
- Sinhala-English Dictionary by Benjamin Clough
- Sinhala-English Dictionary by Dr A. P. De Soysa
- Sanskrit- English Dictionary by Monier Williams
- Technical Glossaries published by the Official Language Department
- *Sinhala Namawaliya* (An ancient Sinhala thesaurus)
- *Ruwanmala* (An ancient Sinhala thesaurus)

The way the resources given above are used in the current project is explained in Section 6.2.

In addition to the above mentioned Sinhala language resources a set of English language resources are also used in the current project. These resources are mainly used to extract the semantic aspect of words and their relations and classifications. The resources are as follows:

- Princeton English WordNet (Fellbaum, 1998)
- Roget's Thesaurus (Roget, 1962)
- Suggested Upper Merged Ontology (SUMO) Ontology (www.ontologyportal.org)

The amount of work published on the semantic aspect of Sinhala language is relatively small due to the fact that it has not been studied formally by the linguists who are involved in Sinhala language research. This has led to a situation where it is difficult to express the semantics of words and their sense relations accurately. In order to address these issues, it was decided to complement the information given in above Sinhala language resources in an informal manner by working with a couple of senior linguists who have a strong theoretical background in both traditional grammar and modern linguistic analysis of Sinhala and English languages. The contribution of this human resource is equally important as the physical resources mention above.

Methodology

Having closely studied the approaches taken in other main WordNet initiatives, the strategy of the Sinhala WordNet development project was developed. As an enormous amount of work has been done in those projects, esp. Princeton English WordNet and Hindi WordNet, the strategy was developed in such a way that current project can be initiated with the existing resources developed under these projects and to avoid most of the issues that have been handled by the developers of those projects. The steps of the methodology of the Sinhala WordNet project can be listed as follows and Figure 6.1 depicts the work flow of the process:

1. *Word Selection Process*

   The words are chosen from the UCSC Sinhala Corpus according to their frequency. Initially the most frequently occurring 500 words excluding function words were chosen to build the prototype of the Sinhala WordNet. It is expected to select 5000 words based on the same principle to build the final system. As Sinhala is a morphologically rich language there are many different word forms for a given base form and only one single form called *lemma* is selected for the current system. If a word form different to the base form has a different semantic value that form is considered as a separate entry. Some words have alternative spellings and phonological variations that have led to semantic variations. These words are considered as separate entries in those cases.

2. *Sense Identification Process*

   As discussed in Section 2, one word can have more than one sense and it is extremely difficulty identify all the senses of a given word. In the current project two approaches are taken to identify the senses of words namely dictionary look up and look up the English translation of the word in the English WordNet. Finally, the list of senses for a given words is determined by a senior linguist after reviewing the potential senses given in the dictionary and the English WordNet. The main source for this exercise is *Maha Sinahla Sabdakoshaya* and the Princeton English WordNet.

3. *Sense Relation Extraction*

   Princeton English WordNet defines all the word sense relations such as synonyms, antonyms, hypernyms, hypronyms, meronyms and holonyms. As defining them from scratch is time consuming and requires a sophisticated expertise in lexical semantics it was decided to extract them from the WordNet databases and store them in a format which is readable for humans. The main motivation behind this decision is the fact that majority of the senses are language and culture independent. Therefore that helps incorporating Sinhala words with those relations to build the Sinhala WordNet with less effort.

4. *Sinhala to English Translation and English WordNet Query*

   The accurate English translation for a given sense of a Sinhala word is determined by a senior linguist who is conversant in both Sinhala and English language usages. Having translated the Sinhala word sense into English, it is in turn looked up in the English WordNet to obtain the synset identifier.

5. *Assign an Appropriate WordNet Sense Identifier and Insert Words to Sinhala WordNet*

   The Sinhala word with a particular word sense is then inserted to the Sinhala WordNet database with the sense identifier obtained according to the step described in 4. The final system will then consists of all the 5000 Sinhala words with their senses encoded with Princeton WordNet sense identifiers which in turn define the other sense relation automatically.

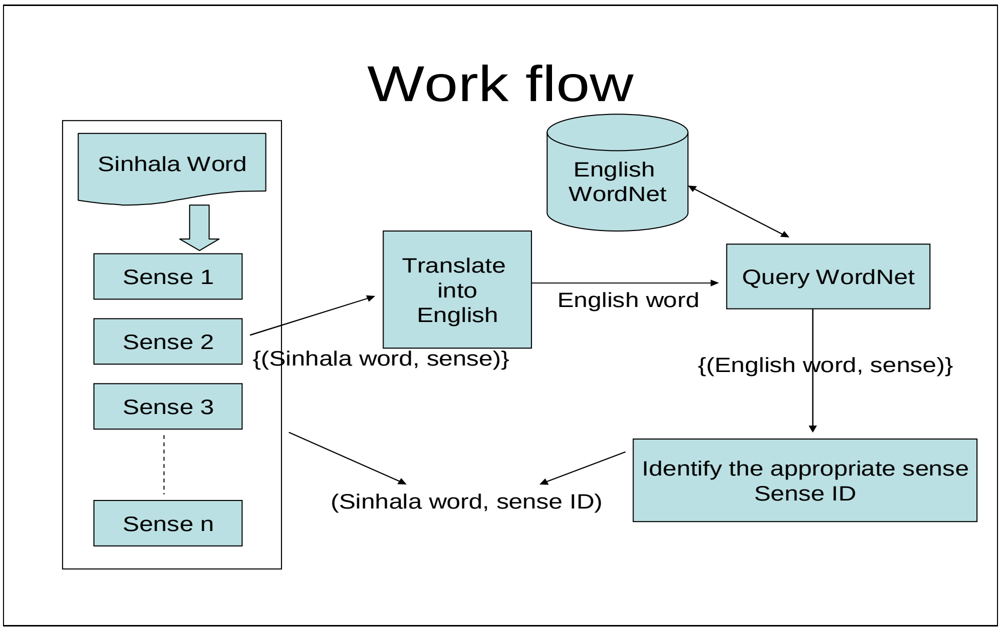

Figure 6.1: Work Flow of the Sinhala WordNet Development

Sinhala WordNet Issues

In the current project, several linguistic issues have to be addressed in order to capture the language specific features into the design of the system. Each of these issues are briefly described below:

1.1. Morphological Forms

As mentioned in many places of this report, Sinhala is a morphologically rich language which gives rise to 110 word forms for nouns and 282 word forms for verbs. Therefore it is extremely important to incorporate a morphological parser to map those word forms to their corresponding *lemma*. A complete morphological parser for Sinhala is being developed at the Language Technology Research Laboratory of the University of Colombo School of Computing and it is expected couple it with the Sinhala WordNet to improve its performance.

1.2. Compound Nouns and Verbs

Compounding is a very productive morphological process in Sinhala. Both Sinhala nouns and verbs formed by compounding nouns (nouns) and nouns with verbs (verbs, esp *do* and *be*). As a result of compounding the original sense of the constituents of the compound noun are changed and a new sense is derived.

Therefore it is extremely important to decide how to handle these issues when implementing the Sinhala WordNet.

1.3. Language and Culture Specific Senses

As the senses and sense relations of the Princeton WordNet are used in the current Sinhala WordNet project, it is important to decide the places in the ontology for the senses that are language and culture specific. There are two possible ways to deal with these senses namely, find the closest approximation in the existing ontology or extend the ontology appropriately to accommodate these concepts to represent them accurately.

1.4. Word Selection Criteria

The words for the Sinhala WordNet are chosen from the UCSC Sinhala Corpus as described in Section 6. Many of these words have senses in the dictionary which are not used in the modern language but in the ancient period. Taking these senses of words into account is not useful for the current project and therefore they should be ignored after carefully examining the period to which the usages of those words belong.

Reference:

- Fellbaum, C (ed) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Vossen, P (ed) 2002. *Euro WordNet General Document*. Vrije Universiteit, Amsterdam
- Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande and P. Bhattacharyya, *An Experience in Building the Indo WordNet - a WordNet for Hindi*, First International Conference on Global WordNet, Mysore, India, January 2002
- D. Chakrabarti and P. Bhattacharyya, *Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT,* Global WordNet Conference (GWC-2004), Czech Republic, January 2004.
- Roget, P.M , Roget, J. L. , Roget, S. R. 1962. *Thesaurus of English Words and Phrases*. Penguin Books.