Corpus-based Sinhala Lexicon

Ruvan Weerasinghe[1], Dulip Herath[2], Viraj Welgama[3]
Language Technology Research Laboratory,
University of Colombo School of Computing
35, Reid Avenue, Colombo 07,
Sri Lanka
{arw[1], dlh[2], wvw[3]}@ucsc.cmb.ac.lk

Abstract

Lexicon is in important resource in any kind of language processing application. Corpus-based lexica have several advantages over other traditional approaches. The lexicon developed for Sinhala was based on the text obtained from a corpus of 10 million words drawn from diverse genres. The words extracted from the corpus have been labeled with parts of speech categories defined according to a novel classification proposed for Sinhala. The lexicon reports 80% coverage over unrestricted text obtained from online sources. The lexicon has been implemented in Lexical Mark up Framework.

1 Introduction

The availability of lexical resources is central to many natural language processing tasks as words play a crucial role in defining higher level constructions such as phrases, clauses and sentences of any language. The most generic and basic lexical resource for such work is a lexicon, preferably with part of speech annotation and information about possible word forms. The latter is important especially for morphologically rich languages such as Sinhala. This kind of resource is extremely useful for part of speech tagging, grammar development and parsing, machine translations, speech processing applications, among others. As new knowledge is created, new concepts are introduced to the language in terms of words. Non-corpus-based lexicon development approaches are not capable of acquiring these new words into lexica due to their inherent limitations such as reliance on introspection and linguistic exposure of the human compiler(s). Therefore it is essential to adopt less expensive (less time consuming, labor intensive and robust) alternative strategies to develop wide-coverage lexica for less studied languages.

This paper presents a lexicon for Sinhala which has nearly 35,000 entries based on the text drawn from the UCSC Text Corpus of Contemporary Sinhala consisting of 10 million words from diverse genres. The corpus-based approach taken in this work can overcome the limitations that traditional approaches suffering from such as less reliance on less expert knowledge, the ability to capture modern usage based on recently introduced words and wide coverage.

The lexical entries defined in this approach are classified according to a novel classification in order to fulfill the requirements of language processing tasks. The broad classes defined are significantly different from those described in traditional Sinhala grammar. For declensional classes such as Nouns and Verbs, further subdivisions have been proposed based on their morpho-phonemic features. Each of the subdivision classes is associated with a set of rules that can be used to generate all possible morphological forms of that group. This has made a significant contribution to improve the coverage of the lexicon as for a given lexical entry it is hard to guarantee that all possible forms exist in the original corpus. However, the rules defined in each class guarantee recognize such unseen forms in the test data set.

In addition, a comprehensive set of function words has been defined based on some of the indeclinable classes such as Post-positions, Particles, Determiners, Conjunctions and Interjections. The lexicon also consists of the most commonly used named entities such as person and city names. Syllabified phonetic transcriptions of the lexical entries are also incorporated in order to make this resource useful in speech processing applications.

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 17–23,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

These characteristics are essential in building effective practical natural language processing applications. To the best of our knowledge, this is the first attempt to build a wide coverage lexicon for Sinhala from a computational linguistic perspective reported in the literature.

The rest of the paper describes the work carried out in detail. Section 2 gives a detailed description of the data acquisition stage, the part of speech categories and the subdivisions based on morphology and the phonetic transcription with syllabification. The implementation details of the lexicon and schemas defined for lexical entries using Lexical Mark up Framework (LMF) is given in Section 3. Section 4 comments on the results of the experiments conducted to measure the coverage of the lexicon. Finally, Section 5 discusses the issues and limitations of the current work with insights for future work.

2 Sinhala Lexicon

2.1 Data Acquisition

The data for the lexicon was obtained from the UCSC Sinhala Corpus which has text drawn from diverse genres namely, Creative Writing, Technical Writing and News Reportage. The corpus represents the modern usage of Sinhala in the above mentioned genres. This guarantees the robustness of the lexicon in practical language processing applications. The text distribution across genres in the corpus is given in Table 1.

Genre	Number of Words	% Number of Words
Creative Writing	2,340,999	23%
Technical Writing	4,357,680	43%
News Reportage	3,433,772	34%

Table 1. Distribution of Corpus Text across Genres

It is clear from the Table 1 that the corpus is fairly balanced across genres while Creative Writing and Technical Writing genres respectively make the lowest and the highest contributions to the total word count of the corpus.

In order to extract the candidate words for the lexicon, a distinct word list with frequencies was obtained by running a simple tokenizer on the corpus text. Misspelt and irrelevant tokens (numbers, foreign words, etc) were removed from the list after manual inspection. Further, the resultant words were manually classified into their respective parts of speech for subsequent processing based on a predefined classification. This was carried out by a team of five members including one senior linguist. At the initial stage of this phase, a substantial effort was made to train the manual classifiers to classify words according to the predefined set of classification criteria. In order to automate this process, the high frequency words were first classified into their respective parts of speech and then certain word ending patterns peculiar to each class were identified. These patterns were used to classify the rest of the list automatically by running regular expression matching followed by manual cleaning. This strategy significantly accelerated the data acquisition process.

In addition to the words taken from the corpus, a comprehensive list of named entities such as person, village/city, country, capital, product names was added to the lexicon after processing data sources obtained from the Departments of Census & Statistics and Inland Revenue. These entries were absorbed into the lexicon on the basis of complete enumeration.

2.2 Parts of Speech and Morphology

In traditional Sinhala grammar, several classifications have been proposed for parts of speech. This is mainly due to the existence of different grammatical schools in Sinhala language studies. They can be broadly classified into three main categories namely, notions based on Sanskrit grammar (Dharmarama, 1913), ideas inspired by language purism significantly different from those based on Sanskrit grammar (Kumaratunaga, 1963), and classifications proposed in the light of modern linguistics (Karunatilake, 2004). From a computational linguistic point of view, each of these classifications while having their own strengths is unable to capture phenomena which are useful for computational linguistics. They are mainly descriptive treatments of language which are used for pedagogical purposes whereas a computational model requires a formal analytical treatment in order to be of any use. Due to the limitations of the existing classifications of Sinhala words, a novel classification of part-of-speech categories was developed after studying the existing classifications closely; consulting linguists and reviewing part of speech tag set design initiatives for other Indic languages. Widely accepted and used tag sets for English were also taken into account when proposed classification was developed. As described in section 1, the current classification has improved the predictive power of each class and this has in turn improved the coverage that is essential for robustness of the lexicon in computational linguistic and natural language processing tasks.

Part of Speech	Frequency	
Noun	12264	35.89%
Verb	1018	3.00%
Adjective	2869	8.40%
Adverb	315	0.92%
Postposition	146	0.43%
Particle	145	0.42%
Conjunction	29	0.08%
Numeral	382	1.12%
Determiner	76	0.22%
Pronoun	150	0.44%
Proper Noun	16585	48.52%
Verb Particle	158	0.46%
Interjection	44	0.13%

Table 2. Part of Speech Categories and their Frequencies

Table 2 shows the thirteen broad classes of parts of speech used in the proposed lexicon. Names of these categories are self-explanatory except for *Verb Particle* which stands for a category of words that are used in Sinhala compound verbs, exemplified best by the terms ඉකුත් (Sinhala: *ikuth*, Sanskrit: *athikränthə*), පත් (Sinhala: *path*, Sanskrit: *präpthə*), and පළ (Sinhala: *palə*, Sanskrit: *prəkətə*). Most of these Verb Particles are localized forms of past participle forms of some Sanskrit verbs. For some historical reason, only past participle forms of these verbs are present in modern usage of Sinhala but not the other forms.

According to the frequency distribution of parts of speech categories given in Table 2, it is clear that nearly 50% of the lexical entries are Proper Nouns. Overall 85% of the total number of entries is nouns with only 3% being Verbs. This is mainly due to the fact that Sinhala has many compound verbs. Compound verbs are usually formed by using Nouns, Adjectives, Verb Particles and Participles of some verbs together with the helper verbs such as කරනවා (English: *do*), වෙනවා (English: *be*), දෙනවා (English: *give*), ගන්නවා (English: *take*), and දානවා (English: *put*). As this contextual information is absent it is hard to determine whether a particular noun or adjective or any other word has occurred as a constituent of a compound verb or not. Therefore they were classified as if they occurred in their primary category.

Even though the number of entries under Verb category is relatively small i.e. nearly 3%, it was found that the number of instances of those verbs is significantly high. In the distinct word list obtained from the original corpus, 4.64% of the entries were verbs (including inflected forms). The total number of instances of verbs (including inflected forms) in the corpus is 19.4% of the total number of words in the corpus. This implies that 3% of the lexicon has coverage of nearly 20% of the corpus. In addition, it was found that 27.7% of the verbs in the corpus are compound verbs since verbs that are essentially part of compound verbs (කරනවා, වෙනවා, දෙනවා, ගන්නවා, දානවා) have occurred 27.7% of the corpus.

It was also possible to identify a set of words which plays only functional roles in Sinhala sentences and have no lexical meaning. In the traditional treatments of grammar they are classified as *nipa:thə* which literally means "*things that fall in either initial or medial or final position of a sentence to express the relationships among elements of the sentence*". This definition does not take into account the different functional roles played by those words and therefore classifies them into one single class called *nipa:thə*. In the work described here, these words were classified into five classes namely, Postpositions, Particles, Conjunctions, Determiners and Interjections. A list of 440 words that belong to these five classes form the first function (stop) word list reported for Sinhala. Identifying the function words is important for applications such as information retrieval, prosody modeling in speech synthesis, semantic role labeling, and dependency parsing.

Nouns and Verbs are further classified into subclasses according to their inflectional/declension paradigms given in Table 3 and 4. These subclasses are mainly specified by the morphophonemic characteristics of stems/roots.

Gender	Subclass	Frequency
Masculine	Consonant-1	63
	Consonant-2	13
	Consonant Reduplication	973
	Front-Mid Vowel	1231
	Back Vowel	191
	Retroflex-1	81
	Retroflex-2	61
	Kinship	180
	Irregular	41
Feminine	Consonant	12

	Front-Mid Vowel	168
	Back Vowel	78
	Irregular	17
Neuter	Consonant	2303
	Consonant Reduplication	206
	Front-Mid Vowel	4379
	Mid Vowel	115
	Back Vowel	1097
	Retroflex-1	127
	Retroflex-2	523
	Uncountable	404
	Irregular	12

Table 3. Noun Subclasses

Nouns are primarily classified with respect to the phone type of the final position of the stem: *Consonant-1* and *Consonant-2* classes have stems that have a consonant ending. The difference between these two classes is defined by the phonological changes that take place when nominal and accusative suffixes are added to the stem. The noun stems belong Consonant-1 has the plural suffix (*-u*) and their final position consonant is reduplicated when the suffix is appended whereas noun stems belong to Consonant-2 has null suffix to mark plurality.

The noun stems that belong to *Consonant Reduplication* have either vowel /i/ or /u/ at the final position. When a nominative or accusative suffix (-a: / -O: / -an) is appended to the noun stem the final position vowel is deleted and the penultimate non-retroflex consonant is reduplicated. If the consonant is retroflex they are classified under *Retroflex-1*. If the noun stems that have vowel /ə/ at the final position and the penultimate consonant is retroflex then the vowel is deleted and the nominative or accusative suffix is appended to the remaining part of the stem. This class is named as *Retroflex-2*.

When a nominative or accusative suffix is appended to a noun stem that belongs to *Front-Mid Vowel* subclass, the semi-consonant /*y*/ is inserted between the noun stem and the suffix. Similarly, /*w*/ is inserted if the noun stem belongs to *Back Vowel* category.*Kinship* and *Uncountable* nouns[1] are inflected in a unique manner irrespective of the phonetic characteristics of stem endings. Each subcategory (*Masculine, Feminine*, and *Neuter*) has a set of stems that behaves irregularly.

Each category has a unique set of phonological rules and inflectional suffixes to generate 130 possible word forms.

Verbs have been classified into four main subclasses according to the phonetic characteristics of their roots.

Subclass	Frequency
ə-ending	488
e-ending	325
i-ending	90
irregular	115

Table 4. Verb Subclasses

As shown in Table 4 the most frequently occurring verbs belong to the ə-ending category. Each of these verb categories except for the irregular category has a unique set of phonological rules and suffixes to generate 240 possible word forms.

2.3 Phonetic Transcriptions

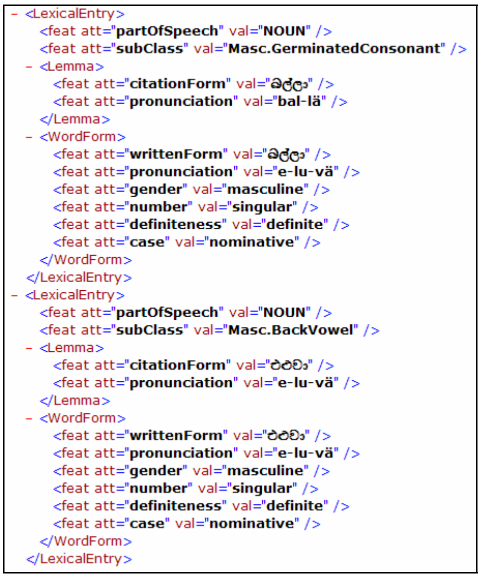
Sinhala orthography is relatively unambiguous as each character corresponds to a unique speech sound. However there are a few ambiguous cases that have to be resolved by taking the context into account. Though Sinhala orthography has different symbols to denote aspirated and unaspirated consonants, in present usage aspiration is not present. Similarly, the alveolar consonants such as ළ (/l/) and ණ (/n/) are now pronounced as their dental counterparts ල and න. Schwa epenthesis also plays a crucial role in Sinhala pronunciation as it leads to significant meaning changes of Sinhala words.

Having considered all these issues, it was decided to incorporate phonetic transcriptions of lexical entries in order to make the current lexicon general purpose. This piece of information is very useful for speech synthesis and recognition application development. Syllabified phonetic transcriptions were automatically derived by applying the grapheme to phoneme (G2P) rules and the syllabification algorithm described in (Wasala et al, 2006) and (Weerasinghe et al, 2005) respectively that report 98% on G2P and 99% accuracy on syllabification. All phonetic transcriptions are given in International Phonetic Alphabet (IPA) symbols.

[1] These two classes have been defined on a semantic basis whereas the other classes are based on phonetic and morphological characteristics of stems.

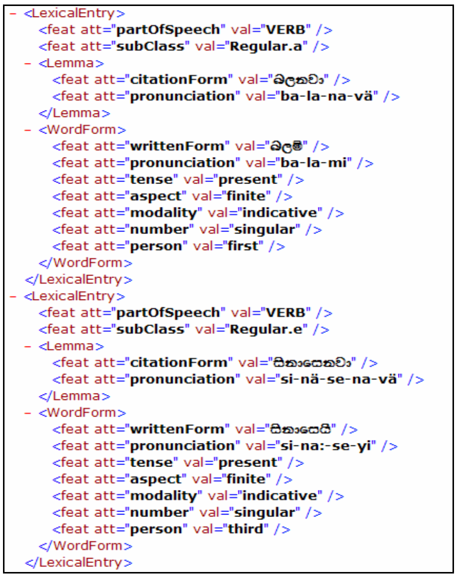
3 Implementation

The lexicon has been implemented in XML according to the specification given in Lexical Mark-up Framework (Francopoulo et al, 2006) which is now the ISO standard for lexicon development. The XML schema defined for Nouns and Verbs with some examples are shown in Figure 1 and 2 respectively.



```
- <LexicalEntry>
    <feat att="partOfSpeech" val="NOUN" />
    <feat att="subClass" val="Masc.GerminatedConsonant" />
  - <Lemma>
      <feat att="citationForm" val="බල්ලා" />
      <feat att="pronunciation" val="bal-lä" />
    </Lemma>
  - <WordForm>
      <feat att="writtenForm" val="බල්ලා" />
      <feat att="pronunciation" val="e-lu-vä" />
      <feat att="gender" val="masculine" />
      <feat att="number" val="singular" />
      <feat att="definiteness" val="definite" />
      <feat att="case" val="nominative" />
    </WordForm>
  </LexicalEntry>
- <LexicalEntry>
    <feat att="partOfSpeech" val="NOUN" />
    <feat att="subClass" val="Masc.BackVowel" />
  - <Lemma>
      <feat att="citationForm" val="එළුවා" />
      <feat att="pronunciation" val="e-lu-vä" />
    </Lemma>
  - <WordForm>
      <feat att="writtenForm" val="එළුවා" />
      <feat att="pronunciation" val="e-lu-vä" />
      <feat att="gender" val="masculine" />
      <feat att="number" val="singular" />
      <feat att="definiteness" val="definite" />
      <feat att="case" val="nominative" />
    </WordForm>
  </LexicalEntry>
```

Figure 1. Lexical Entries for
Nouns බල්ලා and එළුවා.



```
- <LexicalEntry>
    <feat att="partOfSpeech" val="VERB" />
    <feat att="subClass" val="Regular.a" />
  - <Lemma>
      <feat att="citationForm" val="බලනවා" />
      <feat att="pronunciation" val="ba-la-na-vä" />
    </Lemma>
  - <WordForm>
      <feat att="writtenForm" val="බලමි" />
      <feat att="pronunciation" val="ba-la-mi" />
      <feat att="tense" val="present" />
      <feat att="aspect" val="finite" />
      <feat att="modality" val="indicative" />
      <feat att="number" val="singular" />
      <feat att="person" val="first" />
    </WordForm>
  </LexicalEntry>
- <LexicalEntry>
    <feat att="partOfSpeech" val="VERB" />
    <feat att="subClass" val="Regular.e" />
  - <Lemma>
      <feat att="citationForm" val="සිනාසෙනවා" />
      <feat att="pronunciation" val="si-nä-se-na-vä" />
    </Lemma>
  - <WordForm>
      <feat att="writtenForm" val="සිනාසෙයි" />
      <feat att="pronunciation" val="si-na:-se-yi" />
      <feat att="tense" val="present" />
      <feat att="aspect" val="finite" />
      <feat att="modality" val="indicative" />
      <feat att="number" val="singular" />
      <feat att="person" val="third" />
    </WordForm>
  </LexicalEntry>
```

Figure 2. Lexical Entries for
Verbs බලනවා (see) and **සිනාසෙනවා** (smile)

As shown in Figure 1, a typical noun entry has main part of speech category (**partOfSpeech**), sub category (**subClass**). Each Lemma has two feature attributes namely citation form and pronunciation. **WordForm** has several feature attributes called *writtenForm* which is the *orthographic representation, pronunciation, number, gender, person, definiteness* and *case* of the particular word form.
In addition to the attributes available in Nouns, schema defined for Verbs have some attributes peculiar to verbs such as tense, aspect and modality. For Verb, only the present tense.

4 Evaluation

4.1 Test Data Sets

The coverage of the current lexicon was measured by conducting a set of experiments on test data prepared for each of the three genres: News Reportage, Technical Writing, and Creative Writing. This data was obtained from text available online: online newspapers, Sinhala Wikipedia, blogs and other websites. From this data two types of test data sets were prepared namely uncleaned and cleaned test data sets. The uncleaned test data contains all the text as they were whereas the cleaned test data contains words that have occurred more than once. Tables 5 and 6 respectively give the details of uncleaned and cleaned data sets.

Genre	Type	Size
Creative Writing	Full Text	108,018
	Distinct List	22,663
Technical Writing	Full Text	107,004
	Distinct List	25,786
News Reportage	Full Text	103,194
	Distinct List	20,225

Table 5. Un-Cleaned Test Data for
Three Main Genres

Genre	Type	Size
Creative Writing	Full Text	94,971
	Distinct List	9,616
Technical Writing	Full Text	91,323
	Distinct List	10,105
News Reportage	Full Text	91,838
	Distinct List	8,869

Table 6. Cleaned Test Data for
Three Main Genres

4.2 Lexicon Coverage

Initially, coverage of the lexicon was measured for each genre for both Full Text (FT) and Distinct Wordlist (DW) obtained from full text on each data sets: un-cleaned and cleaned. According to the results of this experiment shown in Table 7, the lexicon reports its highest coverage in Creative Writing genre and the lowest is reported in News Reportage.

Genre	Data Set			
	Un-cleaned		Cleaned	
	DW	FT	DW	FT
Creative Writing	60.11%	82.42%	72.21%	86.71%
Technical Writing	58.74%	80.32%	70.73%	84.15%
News Reportage	55.20%	79.82%	71.1%	85.81%

Table 7. Coverage Reported for each Genre on Un-cleaned and Cleaned Data Sets

There is a significant difference between the coverage reported on the Distinct Wordlists obtained from Un-cleaned and Cleaned datasets that is 60% to 72% in Creative Writing, 58% to 70% in Technical Writing and 55% to 71% in News Reportage. This consistent difference proves that a significant number of the words that could not be found in the lexicon were occurred only once in the test data set.

Relatively higher coverage can be achieved when the full text is used rather than a distinct list of words. As high frequency words occur in text more than once in practical situations the lexicon covers a large area of the text though it cannot recognize some low frequency words in the text. This is evident from the differences of coverage reported on Distinct Wordlists and Full Text for both un-cleaned and cleaned data sets (see Table 7). Around 20% coverage difference between Distinct Wordlist and Full Text was reported for each genre.

The average coverage of the lexicon was computed by averaging the coverage reported for three different genres on un-cleaned full-text (FT) data set, which is 80.9%.

In addition, a similar experiment was conducted to measure the significance of the classification proposed in the current work. In that experiment, the coverage of the lexicon was measured by taking only the word forms occurred in the original corpus but not all the forms of the words occurred in the original corpus. Then the rules defined in each subdivision of nouns and verbs were used to generate all possible forms of the words occurred in the original corpus. This experiment was carried out on the distinct word list obtained from the un-cleaned data set. The results show that there were 3.8%, 3.4% and 3.2% improvements in the coverage of creative writing, technical writing and news reportage genres respectively after introducing the generation rules for each subdivision of nouns and verbs.

4.3 Error Analysis

A comprehensive error analysis was done on the words that could not be found in the lexicon to identify the issues behind the errors reported. It was found that there were several types of errors that have contributed to the overall error. The identified error types are given in Table 8.

Error Type	Description
Word Division Errors (D)	Word does not follow standard word division policy
Spelling Error (E)	Word is incorrectly spelt
Foreign Word (F)	Foreign word written in Sinhala script
Non Standard Spelling (N)	Word does not follow standard spelling
Proper Nouns (P)	Word is a Proper Noun
Spoken Forms (S)	Word is a spoken form
Typographic Errors (T)	Word had typographic errors
Wrong Word Forms (W)	Word is an incorrect morphological form
Correct Forms (C)	Correct word not found in the lexicon

Table 8. Typical Errors Found in the Error Analysis

The distribution of these errors across three different genres is given in Table 9. These results were taken only for the cleaned data set. According to the reported results, it is clear that some errors are prominent in some genres are some are consistently present in all the genres. For example, word division errors (D), correct form errors (C), wrong word form errors (W), and non standard spelling errors (N) are consistently occurring in all three genres whereas spoken form errors (S) are prominent in Creative Writing genre (8.52%), Spelling Errors (E) are more prominent in Technical Writing genre, more foreign word errors (F) are found in Technical Writing genre, typographic errors (T) are prominent in found in News Reportage.

Error Type	Creative Writing		Technical Writing		News Reportage	
	DW	FT	DW	FT	DW	FT

C	38.88	39.59	25.81	25.20	17.16	10.57
D	31.84	32.28	23.16	25.36	33.82	33.15
E	5.01	4.49	6.31	5.72	2.31	2.60
F	3.01	2.27	5.27	3.37	2.12	2.19
N	2.30	2.77	6.34	5.94	5.40	6.44
P	2.86	1.89	11.37	11.59	11.03	7.88
S	8.52	8.64	3.20	2.14	5.40	4.34
T	6.22	6.84	14.96	17.88	18.97	29.10
W	1.36	1.22	3.58	2.79	3.78	3.73

Table 9. Different Error Types Distributed across Three Genres Reported on Distinct Wordlist (DW) and Full Text (FT)

It can be concluded from these observations that the errors that are genre independent occur more frequently than genre dependent error and they are the most general mistakes that writers make in their writings. The typographic errors that more frequent in Technical Writing and News Reportage genres are mainly due to the complications in Sinhala typing and Unicode representation. As Sinhala Unicode uses Zero Width Joiner character very often to represent combined characters typists make errors when typing by inserting this character incorrectly. It is hard for them to correct it by deleting that character as it is invisible to the typist on the computer screen. It is clear that from the results shown in Table 9 that there is no significant difference between the error distributions in distinct wordlist and full text test data.

5 Issues and Future Work

The current lexicon has 80% coverage over unrestricted text selected from online sources. In order to make this lexicon robust in practical language processing applications it is important to further improve its coverage in different domains.

It was observed that the number of verbs in the lexicon is relatively small due to the fact that fairly large numbers of Sinhala verbs are compound verbs. In the future it is expected to incorporate those compound verbs so that the coverage of verbs of the lexicon is relatively higher.

In the current implementation the word forms of nouns and verbs are generated by using third party commercial software. It is expected to incorporate a morphological analyzer and generator so that all the possible word forms can be generated by the lexicon itself.

6 Acknowledgements

Authors are very grateful to Prof Tissa Jayawardana at the University of Colombo for his generous support extended to linguistic aspect of this project. The contribution made by the members of the Development Team at the Language Technology Research Laboratory of the University of Colombo of School of Computing-Sri Lanka for manual data classification process is highly appreciated. Authors would also like to acknowledge the feedback given by three unknown reviewers as they have been extremely useful in improving the quality of the work reported.

References

Asanka Wasala, Ruvan Weerasinghe, Kumudu Gamage. 2006. *Sinhala Grapheme-toPhoneme Conversion and Rules for Schwa Epenthesis*. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia. pp. 890—897

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria. 2006. *Lexical Markup Framework*. LREC 2006

Kumaratunga Munidasa. 1963. *Vyakarana Vivaranaya*. M. D. Gunasena Publishers, Colombo

Rathmalane Dharmarama. 1913. *Sidath Sangarawa.* Author Publication.

Ruvan Weerasinghe Asanka Wasala, and Kumudu Gamage. 2005 *A Rule Based Syllabification Algtoithm for Sinhala*. Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05), Jeju Island, Korea. pp. 438-449

W S Karunatilaka. 2004. *Sinhala Bhasa Vyakaranaya*. 5th Edition, M. D Gunasena Publishers, Colombo