Defining the Gold Standard Definitions for the Morphology of Sinhala Words

Welgama Viraj[1], Weerasinghe Ruvan[1], and Mahesan Niranjan[2]

[1]University of Colombo School of Computing,
No:35, Reid Avenue, Colombo 00700
Sri Lanka.
[2]University of Southampton
Highfield, Southampton,
SO17 1BJ, UK.
[1]{wvw,arw}@ucsc.cmb.ac.lk
[2]mn@ec.soton.ac.uk

Abstract. In this work, we describe the steps and strategies we carried out on defining morpheme segmentation boundaries of Sinhala words (which we called *Gold Standard Definitions*). We measured the coverage of the defined resource against three different Sinhala corpora and obtained over 70% coverage for each corpora. Then we report some interesting facts and findings about the Sinhala language revealed due to this development and finally about some applications of this valuable linguistic resource.

Keywords: Sinhala Morphology, Gold Standard Definitions, POS categories for Sinhala

1 Introduction

Identifying the morpheme boundaries of a word is very essential for modern Natural Language Processing tasks. It is the fundamental goal of any automatic morpheme induction algorithm or any rule-based morphological analyzer. The accuracy of identifying morpheme boundaries effects to the permanence of its applications such as Speech Recognition, Machine Translation, Information Retrieval and Statistical Language Modeling, specially if those are performed with morphological reach languages.

There are two major approaches for identifying morpheme boundaries of a word namely; *knowledge-based* approaches and *data-driven* approaches. Though very successful, the knowledge-based approaches are very expensive with respect to the human resource they require. As a result, research on morphological segmentation is now moving towards more data-driven approaches, which require less expertise and heuristics, but rely on data [1]. However, in order to precisely evaluate such data-driven approaches it requires a pre-defined morpheme definitions, referred to as *Gold Standard definitions.* Some key competitions on developing data-driven approaches such as Morpho Challenge Competition [2]

pp. 163–171; rec. 2015-01-21; acc. 2015-02-25

have used gold standard definitions as one way of evaluating the algorithms and they have provided some sample Gold Standard definitions for English, German, Turkish and Finnish [3].

Our goal in this paper is to present the methodology and some findings on developing such resource for identifying morpheme segmentation boundaries of Sinhala words. Sinhala is an Indo-Aryan language spoken by more than 16 million people in Sri Lanka. Sinhala is a highly inflectional language as are many other Indic languages, and like many of them, can be considered as a low-resourced language with respect to the linguistic resources available for NLP. Therefore we assume that developing this kind of resource for Sinhala will provide a potential infrastructure for future research in Sinhala language. The rest of the paper describes the work carried out in detail.

2 POS Categories

Defining morpheme segmentation boundaries of words in a particular language is a highly challenging task, which needs lots of linguistic expertise and heuristic knowledge. Expert native speaker knowledge is required to classify words in to basic and sub POS categories . [4] have made some effort to define major POS categories of the Sinhala language and all the sub-structures of each category with a comprehensive list of words for each category. We used this work as the base for defining morpheme segmentation boundaries.

Having observing each POS category defined in [4], we decided to initially define morpheme segmentation boundaries only for five main POS categories namely; nouns, verbs, adjectives, adverbs and function words. [4] have introduced a novel sub classification for each of these categories according to their inflectional/declension paradigms and these subclasses are mainly specified by the morphophonemic characteristics of stems/roots.

2.1 Nouns

[4] have introduced 22 such sub categories for nouns based in their morphophonemic characteristics at the end of the word. We identified 26 sub categories based on their behavior in inflections and Table 1 shows all the sub categories defined for Sinhala nouns with number of words and number of inflected forms generate from each category with an example. [4] have identified 130 word forms for nouns in general, but we observed that non of these sub categories are inflected to all of these 130 forms.

As shown in the 4$^{\text{th}}$ column of the Table 1, masculine nouns generate the maximum number of inflected forms per sub category, which is 58. We classified 11,970 noun stems into these 26 sub categories and hence we were able to define morpheme segmentation boundaries for 529,781 distinct Sinhala nouns. The methodology we used to define these boundaries will describe later in this paper.

Table 1. Sub-categories for nouns

Group	Subclass	Words	Forms	Example
Masculine	FrontVowel. MidVowel	1,186	58	gɑwə(*cow*)
	Germinated Consonant	972	58	bɑlu (*dog*)
	BackVowel	190	58	elu (*goat*)
	Retroflex-1.1	48	58	kɑputu (*crow*)
	Retroflex-1.2	31	58	utumä (*lord*)
	Retroflex-2.1	19	58	kumərə(*prince*)
	Retroflex-2.2	37	30	sɑhɑkɑru (*partner*)
	Consonant-1	60	58	minis (*man*)
	Consonant-2	9	58	hɑrɑk (*bull*)
	Consonant-3	4	58	girä (*parrot*)
Feminine	FrontVowel. MidVowel	166	47	kuməri (*princess*)
	BackVowel	72	47	äryä (*lady*)
	Consonant	13	44	məw (*mother*)
Neuter	FrontVowel. MidVowel	4,234	42	mæsə(*table*)
	Germinated Consonant	207	42	kɑju (*nuts*)
	BackVowel	1,070	42	putu (*chair*)
	Retroflex-1	122	45	siruru (*body*)
	Retroflex-2	519	45	irə(*sun*)
	Consonant	2,272	42	gɑs (*tree*)
	MidVowel	116	33	kɑdə(*shops*)
kinship	kinship-1	31	42	ɑkkä (*sister*)
	kinship-2	32	46	gurutumä (*teacher)*
	kinship-3	102	27	mɑllë (*brother*)
Uncountable	Consonant Ending	187	12	käbən (*carbon*)
	Vowel Ending	214	12	sëni (*sugar*)
Irregular	Animate	57	16	nönä (*lady*)

2.2 Verbs

Even though verbs are playing the most significant role of the meaning of a sentence, number of verbs in a particular language is far below than the number of nouns of that language. Hence, the classification of verbs into sub categories is simpler than nouns. [4] have identified 4 sub categories for Sinhala verbs, but we further divided one of this category into two by considering their behavior when generating inflected forms. Table 2 shows all the sub categories defined for Sinhala verbs with number of words and number of inflected forms generate from each category with an example.

As shown in the table 2, number of inflected forms of Sinhala verbs are much higher than nouns. The reason behind of this higher number of inflected forms for Sinhala verbs is the gerund forms (*verbal nouns*). There are 3 main gerund forms for each category and each of those forms are inflected to around 40 different forms as in nouns. All together there are 117 gerund forms for each sub category. However, some of these gerund forms are high frequency nouns. for example the word “godənægillə” (*the building*) is a high frequency noun and a general person may not be aware that it is derived from the verb “godənɑgənəwä

Table 2. Sub-categories for verbs

Subclass	Words	Forms	Example
ə-ending	487	206	bɑlə (*to see*)
e-ending	323	198	sinäse (*smiling*)
i-ending-1	47	200	rɑki (*to protect*)
i-ending-2	44	200	ɑndi (*to dress*)
irregular	108	-	bo (*to drink*)

(*to build*). We decided to consider these gerund forms as derivatives of verbs, but we can still consider them as nouns whenever necessary since we have tagged them as *gerund*. We identified 1,009 Sinhala verb roots in all 5 sub categories and coverage of it will be described later in this paper.

2.3 Adjectives

There are two main categories for adjectives. One is playing the adjectival role in a sentence based on its position while the other category is pure adjectives such as "usə" (*tall*) or "hondə" (*good*). Most of the time the noun stems play the adjectival role as in "putu kɑkulə" (*chair's leg*) or "minis hɑndə" (*human voice*). We only consider pure adjectives under this category and we identified 2,576 pure adjectives for Sinhala. All the adjectives are inflected for 2 forms and we named them as "*conjunction form*" (for example "hondɑtə" (*good and*)) and "*final form*" (for example "hondɑyi" (*is* good)).

2.4 Adverbs

As adjectives, adverbs can also be divided into two categories as *derivative adverbs* and *pure adverbs*. We only considered pure adverbs under this category and 245 such adverbs were identified. All the adverbs are also inflected for 2 forms as in adjectives.

2.5 Function Words

We identified 6 types function words for Sinhala. 4 of them were further divided into two groups as "*vowel endings*" and "*consonant endings*" and it helps to programmatically generate the corresponding inflected forms of each category. We identified 619 function words for Sinhala in all of 6 sub categories and Table 3 shows its distribution over each sub category.

Table 3. Sub-categories for function words

Group	**Subclass**	**Words**	**Forms**
Conjunctions	vowel endings	17	3
	consonant endings	12	2
Determinants	vowel endings	52	3
	consonant endings	46	1
Interjections	-	44	1
Particles	vowel endings	110	3
	consonant endings	35	2
Postpositions	vowel endings	107	3
	consonant endings	39	2
Verbparticles	-	157	1

3 Methodology

As described in section 2, we grouped all Sinhala words into 43 sub categories based on their POS categories and word endings. The main objective of this classification is to programmatically generate the morpheme boundaries for rest of all the words of each category based on a given definition file from each category.

3.1 Creating the Definition File

To define the morpheme boundary definitions for each category, we selected a word from each category and manually define all the morpheme boundaries for each of its inflected forms with help of native language experts. We defined two types of definitions for each word namely; "*definitions with morphs*" and "*definitions with features*".

Definitions with Morphs

In these definitions, we tried to define morpheme boundaries of a word based on its orthography and we did not consider the orthographic changes happening into the word ending when adding a suffix. We split a word into its morphemes and defined its form (the morph realization at the particular word) and the definition of the morpheme separately. The fundamental rule we kept on splitting morphemes is that they should be able to produce the relevant word by simply concatenating all the morphs. The objective of following such rule is to use these gold standard definitions to evaluate machine generated morpheme boundaries, which always split a word into morpheme like units based on its spellings. We used colon (:) to separate the morphs from its name and Table 4 shows a sample of these definitions for the Sinhala word "ගවයා" (*the cow*), which is from the category *Nouns-Masculine.FrontVowel.MidVowel.*

˜ stands for the empty morph which we used to denote hidden morphs of a word which is highly utilized in Sinhala Nouns.

Table 4. Example for definitions with morphs

Word	Definition
ගව (*cow-Root*)	ගව:ගව-N+RT
ගවයා (*the cow*)	ගව:ගව-N+RT යා:+SG ˜:+DF ˜:+NOM
	ගව:ගව-N+RT යා:+SG ˜:+DF ˜:+ACC
ගවයාත් (*and cow*)	ගව:ගව-N+RT යා:+SG ˜:+DF ˜:+NOM ත්:+CJ
	ගව:ගව-N+RT යා:+SG ˜:+DF ˜:+ACC ත්:+CJ
ගවයායි (*is cow*)	ගව:ගව-N+RT යා:+SG ˜:+DF ˜:+NOM යි:+FN
	ගව:ගව-N+RT යා:+SG ˜:+DF ˜:+ACC යි:+FN
ගවයෙක් (*a cow*)	ගව:ගව-N+RT ෙ◌ය:+SG ක්:+ID ˜:+NOM

Definitions with Features

In these definitions, we defined morphological features of a particular word with its root or stem. Table 5 shows a sample of these definitions with features for the same example as above.

Table 5. Example for definitions with features

Word	Definition
ගව (*cow-Root*)	ගව:ගව-N+RT
ගවයා (*the cow*)	ගව:ගව-N +SG +DF +NOM ගව:ගව-N +SG +DF +ACC
ගවයාත් (*and cow*)	ගව:ගව-N +SG +DF +NOM +CJ ගව:ගව-N +SG +DF +ACC +CJ
ගවයායි (*is cow*)	ගව:ගව-N +SG +DF +NOM +FN ගව:ගව-N +SG +DF +ACC +FN
ගවයෙක් (*a cow*)	ගව:ගව-N+RT +SG +ID :+NOM

The objective behind this definition is to use this gold standard definitions as a resource for a *Sinhala Morphological Analyzer / Generator*. However, this is a derivative work of the *definitions with morphs*, but we kept defining this separately for the simplicity.

3.2 Creating the Roots File

The lexicon defined by [4] were used to create list of roots for a particular category. For some categories, the root form changes when adding some inflectional suffixes. For example, the verb root “bɑlə” (*to see*) become “bælæ” when adding inflectional suffixes for past tense. These changes can not be automatically predicted for some categories and hence we manually compiled the roots files with these alternative forms for each category.

After defining morpheme boundaries for all inflectional forms of a selected word in a category as described above, a computer program used to generate those definitions for all the other words of its category. The program requires the definition file and the list of roots of a particular category and it replaces the definition file’s root with other roots. This approach helped us to generate such morpheme boundary definitions for most Sinhala words with less effort.

4 Statistics

We managed to compile the first version of Sinhala Gold Standard Definitions with 736,084 Sinhala words using the above approach. Table 6 shows the number of root forms covered in each POS category with their percentage with respect to the total number of stems.

Table 6. Distribution of number of stems of each POS category

Category	**No. of Stems**	**%**
Nouns	11,971	72.90
Verbs	1,009	6.14
Adjectives	2,576	15.69
Adverbs	245	1.49
Function Words	619	3.77
Total	**16,420**	**100.00**

As shown in the Table 6, nouns cover nearly 73% of stems. That is expected because nouns are the most common POS category of a language. However, it is interesting to see that the number of adjectives in Sinhala is much higher than number of verb roots. This phenomenon is changed when we consider the total number of words of each category including their inflected forms. Table 7 shows number of total word forms of each category with their percentage with respect to the total number of words.

It is interesting to see that the coverage of nouns with respect to the total number of defined words are almost similar as it in stems. However, the percentage of verb forms of the language is significant with respect to the other categories other than nouns. As shown in the Table 7, number of adverbs in Sinhala is negligible with compare the total number of words of the language.

Table 7. Distribution of number of total words of each POS category

Category	No. of Words	%
Nouns	529,781	71.97
Verbs	196,873	26.75
Adjectives	7,503	1.02
Adverbs	671	0.09
Function Words	1,256	0.17
Total	**736,084**	**100.00**

5 Coverage

We measure the coverage of this defined resource against 3 different Sinhala corpora. The main resource we used to measure the coverage is the UCSC 10M words Sinhala corpus described in [5] and 70% of the 10 million words are covered by the defined resource. Interestingly, the coverage against the unique word list extracted from the above corpus is only 20.64% and that gives a clue on applicability of Zipf's law for the Sinhala language. We figured out that the most of uncovered words are proper nouns (which are not covered by the defined resource) and typos.

Second resource we used to measure the coverage of the defined resource is 2.4M words Sinhala news corpus extracted from online newspapers. The coverage of the defined resource is 72.65% against this news corpus and it is slightly better than 10M words open domain corpus. The reason behind this slight improvement may be due to less number of typos in online newspapers.

We used 0.95M words Sinhala news editorial corpus as the third resource for measuring the coverage and we obtained 78.27% coverage against the editorial corpus. It can be assumed that the low number of proper nouns and professional writing styles of newspaper editors are the reasons for this improvements.

6 Applications

One of the main objective of developing such resource is to use these definitions to evaluate machine learning approaches on automatic morpheme boundary detections. This resource can directly be used to check the accuracy of the output of such approaches and it will give a precise measure on the performance of morpheme induction algorithms other than any other evaluation methods. Such attempt has described in Kurimo et al. (2010).

Another direct application of such resource is a rule-based morphological analyzer for the particular language. We developed a rule-based morphological analyzer for Sinhala using this resource and that is the first such tool available for Sinhala. Currently, this morphological analyzer is using for research on Sinhala speech recognition and Sinhala-Tamil machine translation and the results of them are yet to be published. Other than serving as a language resource for NLP research in Sinhala, this resource is expected to be used as a learning material for Sinhala.

7 Conclusion

We presented the approach and the data sources we used to develop the Gold Standard Definitions on marking morpheme boundaries for Sinhala words. The defined resource covered over 70% open domain Sinhala words. This is the first attempt on define such resource for Sinhala language and we hope that this resource will be useful for many NLP applications for Sinhala in the future.

Acknowledgments

The authors are very grateful to past and current members at the Language Technology Research Laboratory of the University of Colombo of School Computing, Sri Lanka for their significant contribution on developing basic linguistic resources for Sinhala language such as corpora and lexica, which make it possible for us to carry out the work described above.

References

1. Welgama V., Weerasinghe R., and Mahesan M.: Evaluating a Machine Learning Approach to Sinhala Morphological Analysis. In: Proceedings of the 10th International Conference on Natural Language Processing, Noida, India (2013)
2. Morpho Challenge 2010 - Semi-supervised and Unsupervised Analysis, `http://research.ics.aalto.fi/events/morphochallenge2010/`
3. Kurimo M., Virpioja S., Turunen V., and Lagus K.:Morpho Challenge 2005-2010: Evaluations and Results. In: Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology, pp. 87–95. Association for Computational Linguistics, Uppsala, Sweden. (2010)
4. Weerasinghe R., Herath D., and Welgama V.: Corpus-based Sinhala Lexicon. In: Proceedings of the 7th Workshop on Asian Language Resources, pp. 17–23. Association for Computational Linguistics, Singapore (2009)
5. Weerasinghe R., Herath D., Welgama V., Medagoda N., Wasala A., and Jayalatharachchi E.: UCSC Sinhala Corpus - PAN Localization Project-Phase I. Language Technology Research Laboratory, University of Colombo School of Computing (2007)