Survey on Publicly Available Sinhala Natural Language Processing Tools and Research

Nisansa de Silva

Abstract—

Sinhala is the native language of the Sinhalese people who make up the largest ethnic group of Sri Lanka. The language belongs to the globe-spanning language tree, Indo-European. However, due to poverty in both linguistic and economic capital, Sinhala, in the perspective of Natural Language Processing tools and research, remains a resource-poor language which has neither the economic drive its cousin English has nor the sheer push of the law of numbers a language such as Chinese has. A number of research groups from Sri Lanka have noticed this dearth and the resultant dire need for proper tools and research for Sinhala natural language processing. However, due to various reasons, these attempts seem to lack coordination and awareness of each other. The objective of this paper is to fill that gap of a comprehensive literature survey of the publicly available Sinhala natural language tools and research so that the researchers working in this field can better utilize contributions of their peers. As such, we shall be uploading this paper to arXiv and perpetually update it periodically to reflect the advances made in the field.

Index Terms—Sinhala, Natural Language Processing, Resource Poor Language

1 Introduction

Sinhala[1] language, being the native language of the Sinhalese people [2–4], who make up the largest ethnic group of the island country of Sri Lanka, enjoys being reported as the mother tongue of approximately 16 million people [5, 6]. To give a brief linguistic background for the purpose of aligning the Sinhala language with the baseline of English, primarily it should be noted that Sinhala language belongs to the same the Indo-European language tree [7, 8]. However, unlike English, which is part of the Germanic branch, Sinhala belongs to the Indo-Aryan branch. Further, Sinhala, unlike English, which borrowed the Latin alphabet, has its own writing system, which is a descendant of the Indian Brahmi script [9–15]. By extension, this makes Sinhala Script a member of the Aramaic family of scripts [16, 17]. Thus by inheritance, Sinhala writing system is *abugida* (*alphasyllabary*), which to say that consonant-vowel sequences are written as single units [18]. It should be noted that the modern Sinhala language have loanwords from languages such as Tamil, English, Portuguese, and Dutch due to various historical reasons [19]. Regardless of the rich historical array of literature spanning several millennia (starting between 3^{rd} to 2^{nd} century BCE [20, 21]), modern natural language processing tools for the Sinhala language are scarce [22].

Natural Language Processing (NLP) is a broad area covering all computational processing and analysis of human languages. To achieve this end, NLP systems operate at different levels [23–25]. A graphical representation of NLP layers and application domains are shown in Figure 1. On one hand, according to Liddy [24], these systems can be categorized into the following layers; *phonological*, *morphological*, *lexical*, *syntactic*, *semantic*, *discourse*, and *pragmatic*. The *phonological* layer deals with the interpretation of language sounds. As such, it consists of mainly speech-to-text and text-to-speech systems. In cases where one is working with written text of the language rather than speech, it is possible to replace this layer with tools which handle Optical Character Recognition (OCR) and language rendering standards (such as Unicode [26]). The *morphological* layer analyses words at their smallest units of meaning. As such, analysis on word lemmas and prefix-suffix-based inflection are handled in this layer. *Lexical* layer handles individual words. Therefore tasks such as Part of Speech (PoS) tagging happens here. The next layer, *syntactic*, takes place at the phrase and sentence level where grammatical structures are utilized to obtain meaning. *Semantic* layer attempts to derive the meanings from the word level to the sentence level. Starting with Named Entity Recognition (NER) at the word level and working its way up by identifying the contexts they are set in until arriving at overall meaning. The *discourse* layer handles meaning in textual units larger than a sentence. In this, the function of a particular sentence maybe contextualized within the document it is set in. Finally, the *pragmatic* layer handles contexts read into contents without having to be explicitly mentioned [23, 24]. Some forms of anaphora (coreference) resolution [27–31] fall into this application.

On the other hand, Wimalasuriya and Dou [25] categorize NLP tools and research by utility. They introduce three categories with increasing complexity; *Information Retrieval* (IR), *Information Extraction* (IE), and *Natural Language Understanding* (NLU). *Information Retrieval* covers applications, which search and retrieve information which are relevant to a given query. For pure IR, tools and methods up-to and including the *syntactic* layer in the above analysis are

- *Nisansa de Silva is with the Department of Computer Science & Engineering, University of Moratuwa.*
  E-mail: nisansa@cse.mrt.ac.lk

Manuscript revised: May 29, 2022.

1. Englebretson and Genetti [1] observe that in some contexts the *Sinhala* language is also referred as *Sinhalese*, *Singhala*, and *Singhalese*

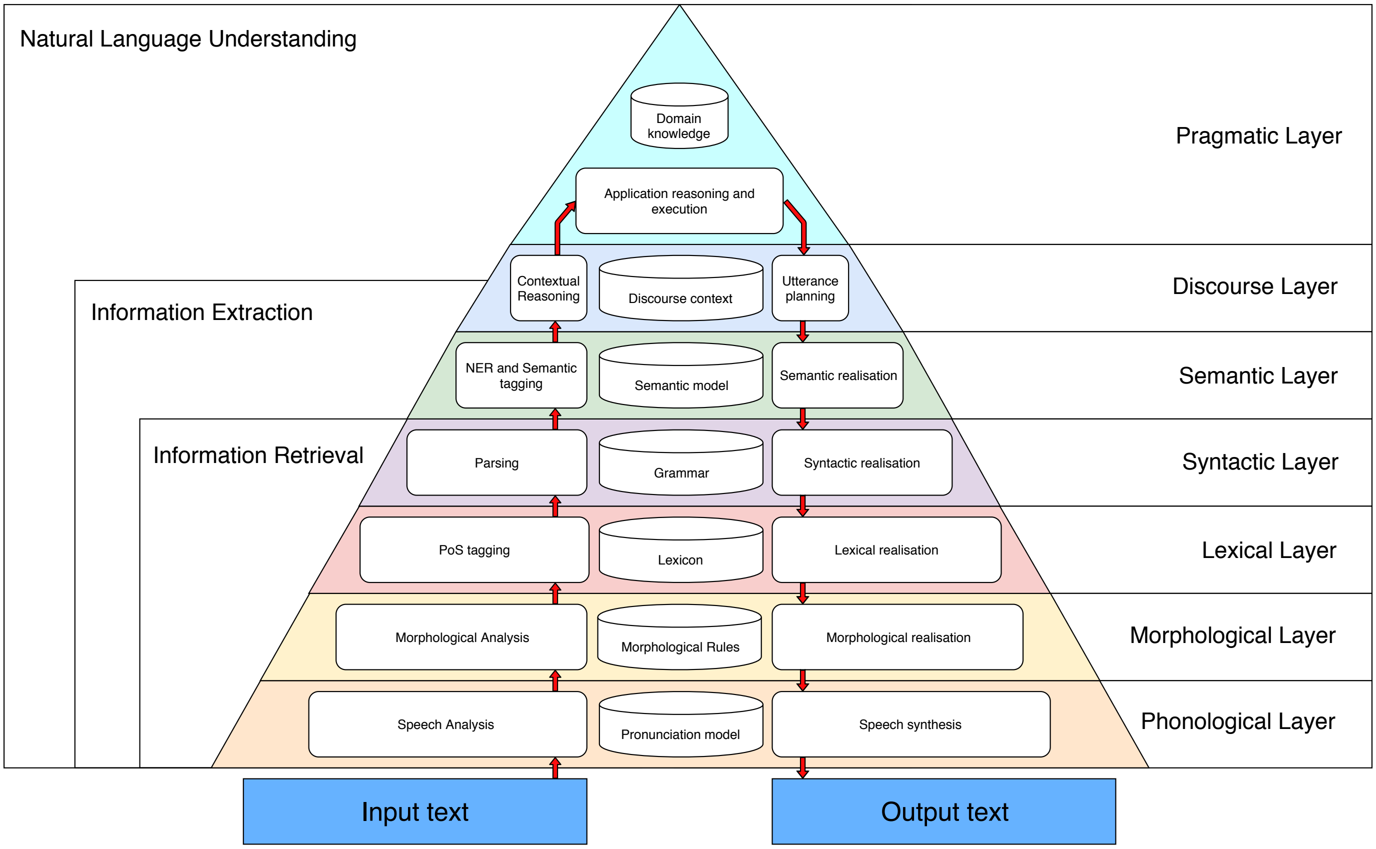


Fig. 1: NLP layers and tasks [23]

used. *Information Extraction*, on the other hand, extracts structured information. The difference between IR and IE is the fact that IR does not change the structure of the documents in question. Be them structured, semi-structured, or unstructured, all IR does is fetching them as they are. In comparison, IE, takes semi-structured or unstructured text and puts them in a machine readable structure. For this, IE utilizes all the layers used by IR and the *semantic* layer. *Natural Language Understanding* is purely the idea of cognition. Most NLU tasks fall under AI-hard category and remain unsolved [23]. However, with varying accuracy, some NLU tasks such as machine translation[2] are being attempted. The *pragmatic* layer of the above analysis belongs to the NLU tasks while the *discourse layer* straddles information extraction and natural language understanding [23].

The objective of this paper is to serve as a comprehensive survey on the state of natural language processing resources for the Sinhala language. The initial structure and content of this survey are heavily influenced by the preliminary surveys carried out by de Silva [22] and Wijeratne et al. [23]. However, our hope is to host this survey at arXiv as a perpetually evolving work which continuously gets updated as new research and tools for Sinhala language are created and made publicly available. Hence, it is our hope that this work will help future researchers who are engaged in Sinhala NLP research to conduct their literature surveys efficiently and comprehensively. For the success of this survey, we shall also consider the Sri Lankan NLP tools repository, *lknlp*[3]. This manuscript is at version 4.1.0. The latest version of the manuscript can be obtained from arXiv[4] or ResearchGate[5].

Figure 3 in Appendix A shows the most prolific researchers in the domain of Sinhala NLP. The nodes contain the name of the researcher along with the total number of Sinhala NLP papers that researcher has authored. The edges between two researchers are labeled with the number of Sinhala NLP papers the relevant pair of researchers have co-authored. Given that the objective of the visualization is to portray the co-operation between researchers, the threshold used here is to have at least 3 papers on Sinhala NLP with on another researcher. We have also added labels to clusters in cases where all or the majority of researchers in those clusters have the same affiliation. It is observable that the cluster from the *Department of Computer Science & Engineering*, *University of Moratuwa* is the most prolific in Sinhala NLP research.

The remainder of this survey is organized as follows; Section 2 introduces some important properties and conventions of the Sinhala language which are important for the development and understanding of Sinhala NLP. Section 3 discusses the various tools and research available for Sinhala NLP. In this section we would discuss both pure Sinhala NLP tools and research as well as hybrid Sinhala-

2. This is, however, not without the criticism of being nothing more than a *Chinese room* [32] rather than true NLU.

3. https://github.com/lknlp/lknlp.github.io
4. https://arxiv.org/abs/1906.02358
5. http://bit.ly/31AhvvR

English work. We will also discuss research and tools which contributes to Sinhala NLP either along with or by the help of Tamil, the other official language of Sri Lanka. Section 4 gives a brief introduction to the primary language sources used by the studies discussed in this work. Finally, Section 5, concludes the survey.

2 Properties of the Sinhala Language

Before moving on to discussing Sinhala NLP resources, we shall give a brief introduction to some of the important properties of Sinhala language, which impact the development of Sinhala NLP resources. Sinhala grammar has two forms: written (literary) and spoken. These forms differ from each-other in their core grammatical structures [1, 35, 36]. The written form strictly adheres to the *SOV* (Subject, Object, and Verb) configuration [37, 38]. Further, in the written form, *subject-verb agreement* is enforced [39] such that, in order to be grammatically correct, the subject and the verb must agree in terms of: gender (male/female), number (singular/plural) and person (1st/2nd/3rd). However, in spoken Sinhala, the *SOV* order can be neglected [40] and *male singular 3rd person* verb can be used for all nouns [39]. Sinhala is also a head-final language, where the complements and modifiers would appear before their heads [41] this is similar to that of English and dissimilar to that of French. In total, according to Abhayasinghe [42], there are 25 types of simple sentence structures in Sinhala. Similar to many Indo Aryan languages, *animacy* plays a major role in Sinhala grammar in syntactic and semantic roles [43–45]. Comparative studies done by Noguchi [46] and by Miyagishi [36, 47] have found that *animacy* extends its influence from phrase level to sentence level in Sinhala (e.g., Usage of post-positions [37, 48]). On this matter, Table 1 explains grammatical cases and inflections of animate common nouns while Table 2 explains grammatical cases and inflections of inanimate common nouns. We provide a comparative analysis of parsing the very simple English sentence "*I eat a red apple*" and its Sinhala, Hindi, and French translations in Fig 2. English and French parsing was done using the *Stanford Parser*[6]. Hindi parsing was done using the *IIIT-Hyderabad Parser*[7] and the study by Singh et al. [49].

Herath et al. [50, 51] argue that pure Sinhala words did not have suffixes and that adding suffixes was incorporated to Sinhala after 12th century BC with the influx of Sanskrit words. With this, they declare Sinhala to have to following types of words:

1) Suffixes
2) Nouns
3) Cases
4) Verbs
5) Conjunctions and articles
6) Adjectives
7) Demonstratives, Interrogatives, and negatives
8) Particles and prefixes

They further divide nouns into five groups: material, agentive, common, abstract, and proper. In addition to these, they also introduce compound nouns.We show the noun categorization proposed by Herath et al. [50] in Table 3.

6. http://nlp.stanford.edu:8080/parser/
7. http://ltrc.iiit.ac.in/analyzer/

Herath et al. [51] categorize Sinhala suffixes along the attributes of: gender, number, definiteness, case, and conjunctive. They further claim that there are 3 types of suffixes: *Suf1* adds gender, number, and definiteness; *Suf2* adds case; and *Suf3* adds conjunctive. Conjunctive is claimed to be equivalent to *too* and *and* in English. We show an extension of the suffix structure proposed by Herath et al. [51] in Table 4.

3 Sinhala NLP resources

In this section we generally follow the structure shown in Figure 1 for sectioning. However, in addition to that, we also discuss topics such as available corpora, other data sets, dictionaries, and WordNets. We focus on NLP tools and research rather than the mechanics of language script handling [52–57]. One of the earliest attempts on Sinhala NLP was done by Herath et al. [58]. However, progress on that project has been minimal due to the limitations of their time. The later work by Nandasara [59] has not caught much of the advances done up to the time of its publication. Given that it was a decade old by the time the first edition of this survey was compiled, we observe the existence of many new discoveries in Sinhala NLP which have not been taken into account by it. A review on some challenges and opportunities of using Sinhala in computer science was done by Nandasara and Mikami [60]. At this point, it is worthy to note that the largest number of studies in Sinhala NLP has been on optical character recognition (OCR) rather than on higher levels of the hierarchy shown in Figure 1. On the other hand, the most prolific single project of Sinhala NLP we have observed so far is an attempt to create an end-to-end Sinhala-to-English translator [18, 61–78]. Tamil, the other official language of Sri Lanka is also a resource poor language. However, due to the existence of larger populations of Tamil speakers worldwide, including but not limited to economic powerhouses such as India, there are more research and tools available for Tamil NLP tasks [23]. Therefore, it is rational to notice that Sinhala and Tamil NLP endeavours can help each other. Especially, given the above fact, that these are official languages of Sri Lanka, results in the generation of parallel data sets in the form of official government documents and local news items. A number of researchers make use of this opportunity. We shall be discussing those applications in this paper as well. Further, there have been some fringe implementations, which bridge Sinhala with other languages such as Japanese [8, 21, 79–81].

3.1 Corpora

For any language, the key for NLP applications and implementations is the existence of adequate corpora. On this matter, a relatively substantial Sinhala text corpus[8] was created by Upeksha et al. [82, 83] by web crawling. It was later extended by adding *Jathaka Stories*[9] and more web crawled news articles[10]. Later a smaller Sinhala news corpus[11] was created by de Silva [22]. Both of the above

8. https://osf.io/a5quv/
9. https://bit.ly/JathakaTxt
10. https://bit.ly/3osodBj
11. https://osf.io/tdb84/

TABLE 1: **Examples for grammatical cases and inflection of animate common nouns**

<table>
<tr><th rowspan="3">Form</th><th rowspan="3">Case</th><th colspan="4">Singular</th><th colspan="2">Plural</th></tr>
<tr><th colspan="2">Masculine</th><th colspan="2">Feminine</th><th rowspan="2">Masculine</th><th rowspan="2">Feminine</th></tr>
<tr><th>Definite</th><th>Indefinite</th><th>Definite</th><th>Indefinite</th></tr>
<tr><td>1</td><td>Nominative</td><td>මිනිසා</td><td>මිනිසෙක්</td><td>ගැහැණිය</td><td>ගැහැණියක්
ගැහැණියෙක්</td><td>මිනිස්සු</td><td>ගැහැණු</td></tr>
<tr><td>2</td><td>Accusative</td><td rowspan="2">මිනිසා</td><td rowspan="2">මිනිසෙකු</td><td rowspan="2">ගැහැණිය</td><td rowspan="2">ගැහැණියක</td><td rowspan="2">මිනිසුන්</td><td rowspan="2">ගැහැණුන්</td></tr>
<tr><td>3</td><td>Auxiliary</td></tr>
<tr><td>4</td><td>Dative</td><td>මිනිසාට</td><td>මිනිසෙකුට</td><td>ගැහැණියට</td><td>ගැහැණියකුට
ගැහැණියකට</td><td>මිනිසුන්ට</td><td>ගැහැණුන්ට</td></tr>
<tr><td>5</td><td>Genitive</td><td rowspan="2">මිනිසාගේ</td><td rowspan="2">මිනිසෙකුගේ</td><td rowspan="2">ගැහැණියගේ</td><td rowspan="2">ගැහැණියකගේ</td><td rowspan="2">මිනිසුන්ගේ</td><td rowspan="2">ගැහැණුන්ගේ</td></tr>
<tr><td>6</td><td>Locative</td></tr>
<tr><td>7</td><td>Instrumental</td><td rowspan="2">මිනිසාගෙන්</td><td rowspan="2">මිනිසෙකුගෙන්</td><td rowspan="2">ගැහැණියගෙන්</td><td rowspan="2">ගැහැණියකගෙන්</td><td rowspan="2">මිනිසුන්ගෙන්</td><td rowspan="2">ගැහැණුන්ගෙන්</td></tr>
<tr><td>8</td><td>Ablative</td></tr>
<tr><td>9</td><td>Vocative</td><td>මිනිස</td><td>මිනිස</td><td>ගැහැණිය</td><td>ගැහැණිය</td><td>මිනිසුනි</td><td>ගැහැණුනි</td></tr>
</table>

TABLE 2: **Examples for grammatical cases and inflection of inanimate common nouns**
Note that the grammatical cases of *Auxiliary* and *Vocative* do not exist for inanimate nouns in Sinhala.

<table>
<tr><th rowspan="2">Form</th><th rowspan="2">Case</th><th colspan="2">Singular</th><th rowspan="2">Plural</th></tr>
<tr><th>Definite</th><th>Indefinite</th></tr>
<tr><td>1</td><td>Nominative</td><td rowspan="2">පොත</td><td rowspan="2">පොතක්</td><td rowspan="2">පොත්</td></tr>
<tr><td>2</td><td>Accusative</td></tr>
<tr><td>4</td><td>Dative</td><td>පොතට</td><td>පොතකට</td><td>පොත්වලට</td></tr>
<tr><td>5</td><td>Genitive</td><td rowspan="2">පොතේ
පොතෙහි</td><td rowspan="2">පොතක</td><td rowspan="2">පොත්වල</td></tr>
<tr><td>6</td><td>Locative</td></tr>
<tr><td>7</td><td>Instrumental</td><td rowspan="2">පොතෙන්
පොතින්</td><td rowspan="2">පොතකින්</td><td rowspan="2">පොත්වලින්</td></tr>
<tr><td>8</td><td>Ablative</td></tr>
</table>

TABLE 3: **Noun categorization by Herath et al. [50]**

Type	**Examples**
Material	පුටු, ගෙය
Agentive	දුවන්නා, බීම
Common	ගොවියා, මිනිසා
Abstract	සුදු, උස
Proper	කොළඹ, අමර
Compound	කිරිබත් (කිරි+බත්), සුදුමල් (සුදු+මල්)

corpora are publicly available. However, none of these come close to the massive capacity and range of the existing English corpora. A word corpus of approximately 35,000 entries was developed by Weerasinghe et al. [84]. But it does not seem to be online anymore. A number of Sinhala-English parallel corpora were introduced by Guzmán et al. [85]. This includes a 600k+ Sinhala-English subtitle pairs[12] initially collected by [86], 45k+ Sinhala-English sentence pairs from GNOME[13], KDE[14], and Ubuntu[15]. Guzmán et al. [85] further provided two monolingual corpora for Sinhala.

12. http://bit.ly/2KsFQxm
13. http://bit.ly/2Z8q0fo
14. http://bit.ly/2WLY6bI
15. http://bit.ly/2wLVZGt

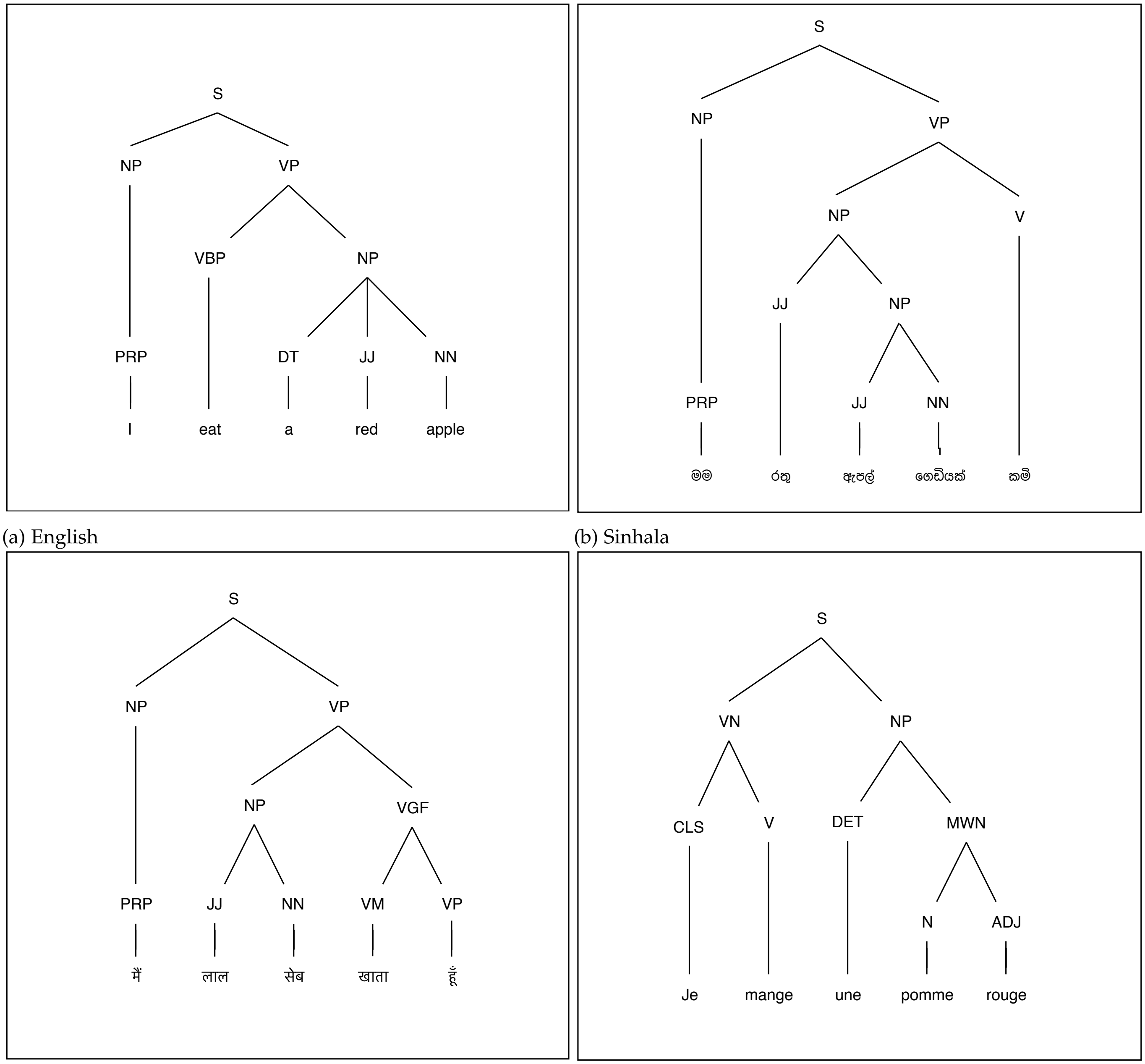
Those were a 155k+ sentences of filtered Sinhala Wikipedia[16] and 5178k+ sentences of Sinhala common crawl[17]. Wijeratne and de Silva [87] have publicly released[18] a massive corpus of text and stop words taken from a decade of Sinhala Facebook posts. A parallel corpus of Sinhala and English was collected by Banón et al. [88] containing 217,407 sentences and available to download from their website[19]. However the later audit by Caswell et al. [89] raised issues on the quality of that data set. A parallel corpus[20] of aligned Sinhala-English documents and sentences obtained from crawling the web was released by Sachintha et al. [90].

As for Sinhala-Tamil corpora, Hameed et al. [91] claim to have built a sentence aligned Sinhala-Tamil parallel corpus and Mohamed et al. [92] claim to have built a word aligned Sinhala-Tamil parallel corpus. However, at the time of writing this paper, neither of them was publicly available. A very small Sinhala-Tamil aligned parallel corpus created by Farhath et al. [93] using order papers of government of Sri Lanka is available to download[21]. A Sinhala and Tamil annotated corpus for emotion analysis was created by Jenarthanan et al. [94]. Vasantharajan and Thayasivam [95] used Printed Character Recognition (PCR) to create a large scale Tamil-Sinhala-English parallel corpus which they caimed to be available on Github [22].

3.2 Data Sets

Specific data sets for Sinhala, as expected, is scarce. However, a Sinhala PoS tagged data set [96–98] is available to download from github[23]. Further, a Sinhala NER data set created by Manamini et al. [99] is also available to download from github[24]. Liyanage et al. [100] analyzed Sinhala fastText and Word2Vec in the context of cross lingual embedding spaces.

16. http://bit.ly/2EQZ7oM
17. http://bit.ly/2ZaQFZo
18. https://bit.ly/2GEI4d6
19. https://www.paracrawl.eu/
20. https://github.com/kdissa/comparable-corpus
21. http://bit.ly/2HTMEme
22. https://github.com/Chaarangan/tamizhinet-corpus
23. http://bit.ly/2Krhrbv
24. http://bit.ly/2XrwCoK



(a) English (b) Sinhala

(c) Hindi (d) French

Fig. 2: **Parse trees for the sentence "*I eat a red apple*" in four languages.**

Facebook has released FastText [101–103] models for the Sinhala language trained using the Wikipedia corpus. They are available as both text models[25] and binary files[26]. Using the above models by Facebook, Lakmal et al. [104] have created an extended FastText model trained on Wikipedia, News, and official government documents. The binary file[27] of the trained model is available to be downloaded. Herath et al. [105] has compiled a report on the Sinhala lexicon for the purpose of establishing a basis for NLP applications. A comparative analysis of Sinhala word embedding has been conducted by Lakmal et al. [104]. A dataset[28] consisting of 3576 Sinhala documents drawn from Sri Lankan news websites and tagged (*CREDIBLE*, *FALSE*, *PARTIAL* or *UNCERTAIN*) was published by Jayawickrama et al. [106] A benchmark data set for Sinhala spell correction was created by Sonnadara et al. [107] which they put online on Github[29]. De Saa and Ranathunga [108] has released a data set[30] for Sinhala hate speech detection which consists of comments from youtube. Another data set consisting of Sinhala hate

25. http://bit.ly/2JXAyL8
26. http://bit.ly/2JY5J9c
27. http://bit.ly/2WowH0h
28. https://github.com/LIRNEasia/MisinformationCorpusSinhala
29. https://github.com/chason94/SinNeuSpellCorrector
30. https://bit.ly/3FTNaMZ

TABLE 4: **Extension of the suffix structure proposed by Herath et al. [51]**
Number = Singular (S) / Plural (P)
Definite = Definite (D) / Indefinite (I) / Undecided (U)
Case = Nominative (N) / Accusative (A) / Dative (Da) / Genitive (G) / Instrumentive (In) / Auxiliary (Au) / Locative (L) / Ablative (Ab)
Conjun = with suffix *th* (Y) / without suffix *th* (N)

Suf1	Suf2	Suf3	Noun	*-tree-*	Attributes			
					Number	Definite	Case	Conjun
-	-	-	ගස්	(-s)	P	U	N, A, V	N
අ	-	-	ගස	(the-)	S	D	N, A, V	N
අක්	-	-	ගසක්	(a-)	S	I	N, A	N
අක	-	-	ගසක	(of a-)	S	I	G	N
අ	ට	-	ගසට	(to the-)	S	D	Da	N
-	ඒ	-	ගසේ	(in the-)	S	D	Au	N
-	වල	-	ගස්වල	(on -s)	P	U	L	N
-	වලට	-	ගස්වලට	(to -s)	P	U	Da	N
-	එහි	-	ගසෙහි	(of the-)	S	D	G	N
අක්	ඉන්	-	ගසකින්	(from a-)	S	I	Au	N
-	-	උත්	ගසුත්	(-s and)	P	U	N, A	Y
අ	-	ත්	ගසත්	(-and)	S	D	N, A	Y
අක්	-	උත්	ගසකුත්	(a- too)	S	I	N, A	Y
අක	-	ත්	ගසකත්	(of a -, too)	S	U	N, A	Y
අ	ට	ත්	ගසටත්	(to the- too)	S	D	Da	Y
-	ඒ	ත්	ගසේත්	(in -s too)	S	D	Au	Y
-	වල	ත්	ගස්වලත්	(on -s too)	P	U	L	Y
-	වලට	ත්	ගස්වලටත්	(to -s too)	P	U	Da	Y
-	එහි	ත්	ගසෙහිත්	(of the- too)	S	D	G	Y
අක්	ඉන්	උත්	ගසකිනුත්	(from a- too)	S	I	Au	Y

speech comments pulled from Facebook is available[31] on Kaggle but it does not have an accompanying paper. A text-to-speech data set[32] with 3300 Sinhala sentences with 7.5 hours of recordings was released by the *Path Nirvana Foundation*.

3.3 Dictionaries

A necessary component for the purpose of bridging Sinhala and English resources are English-Sinhala dictionaries. The earliest and most extensive Sinhala-English dictionary available for consumption was by Malalasekera [109]. However, this dictionary is locked behind copyright laws and is not available for public research and development. This copyright issue is shared with other printed dictionaries [110–115] as well. The dictionary by Kulatunga [116] is publicly available for usage through an online web interface but does not provide API access or means to directly access the data set. The largest publicly available English-Sinhala dictionary data set is from a discontinued FireFox plug-in *EnSiTip* [117] which bears a more than passing resemblance to the above dictionary by Kulatunga [116]. Hettige and Karunananda [64] claim to to have created a lexicon to help in their attempt to create a system capable of English-to-Sinhala machine translation. A review on the requirements for English-Sinhala smart bilingual dictionary was conducted by Samarawickrama and Hettige [118].

There exists the government sponsored trilingual dictionary [119], which matches Sinhala, English, and Tamil. However, other than a crude web interface on the ministry website, there is no efficient API or any other way for a researcher to access the data on this dictionary. Weerasinghe and Dias [120] have created a multilingual place name database for Sri Lanka which may function both as a dictionary and a resource for certain NER tasks.

3.4 WordNets

WordNets [121] are extremely powerful and act as a versatile component of many NLP applications. They encompass a number of linguistic properties which exist between the words in the lexicon of the language including but not limited to: *hyponymy*, *hypernymy*, *synonymy*, and *meronymy*. Their uses range from simple gazetteer listing applications [25] to information extraction based on semantic similarity [122, 123] or semantic oppositeness [124]. An attempt has been made to build a Sinhala Wordnet by Wijesiri et al. [125]. For a time it was hosted on [126] but it too is now

31. https://bit.ly/3j9r3Za
32. https://github.com/pnfo/sinhala-tts-dataset

defunct and all the data and applications are lost other than what Arukgoda et al. [127] have cloned to use in their application uploaded on github[33]. However, even at its peak, due to the lack of volunteers for the crowd-soured methodology of populating the WordNet, it was at best an incomplete product. Another effort to build a Sinhala Wordnet was initiated by Welgama et al. [128] independently from above; but it too have stopped progression even before achieving the completion level of Wijesiri et al. [125].

3.5 Morphological Analyzers

As shown in Fig 1, morphological analysis is a ground level necessary component for natural language processing. Given that Sinhala is a highly inflected language [22, 39, 40], a proper morphological analysis process is vital. The earliest attempt on Sinhala morphological analysis we have observed are the studies by Herath et al. [50, 51]. They are more of an analysis of Sinhala morphology rather than a working tool. As such we discussed the observations and conclusions of these works at Section 2. It is also worth to note that these works predates the introduction of Sinhala unicode and thus use a transliteration of Sinhala in the Latin alphabet.

The next attempt by Herath et al. [129] creates a modular unit structure for morphological analysis of Sinhala. Much later, as a step on their efforts to create a system with the ability to do English-to-Sinhala machine translation, Hettige and Karunananda [61] claim to have created a morphological analyzer (void of any public data or code), which links to their studies of a Sinhala parser [62] and computational grammar [18]. Hettige et al. [73] further propose a multi-agent System for morphological analysis. Welgama et al. [130] attempted to evaluate machine learning approaches for Sinhala morphological analysis. Yet another independent attempt to create a morphological parser for Sinhala verbs was carried out by Fernando and Weerasinghe [131]. Later, another study, which was restricted to morphological analysis of Sinhala verbs was conducted by Dilshani and Dias [132]. There was no indication on whether this work was continued to cover other types of words. Further, other than this singular publication, no data or tools were made publicly accessible. Nandathilaka et al. [133] proposed a rule based approach for Sinhala lemmatizing. The work by Welgama et al. [134] claim to have set a set of gold standard definitions for the morphology of Sinhala Words; but given that their results are not publicly available, further usage or confirmation of these claims cannot not be done. The table 5 provides a comparative summery of the discussion above. The combined study introduced a rule based stemmer [135] and a tokenizer [136] for Sinhala. A later work by Kumarasinghe et al. [137] named *SinMorphy* used a comprehensive vocabulary of Sinhala words to conduct rule-based morphological analysis on Sinhala.

3.6 Part of Speech Taggers

The next step after morphological analysis is Part of Speech (PoS) tagging. The PoS tags differ in number and functionality from language to language. Therefore, the first step in creating an effective PoS tagger is to identifying the PoS tag set for the language. This work has been accomplished by Fernando et al. [96] and Dilshani et al. [97]. Expanding on that, Fernando et al. [96] has introduced an SVM Based PoS Tagger for Sinhala and then Fernando and Ranathunga [98] give an evaluation of different classifiers for the task of Sinhala PoS tagging. While here it is obvious that there has been some follow up work after the initial foundation, it seems, all of that has been internal to one research group at one institution as neither the data nor the tools of any of these findings have been made available for the use of external researchers. Several attempts to create a stochastic PoS tagger for Sinhala has been done with the studies by Herath and Weerasinghe [138], Jayaweera and Dias [139], and Jayasuriya and Weerasinghe [140] being the most notable. Within a single group which did one of the above stochastic studies [139], yet another set of studies was carried out to create a Sinhala PoS tagger starting with the foundation of Jayaweera and Dias [141] which then extended to a Hidden Markov Model (HMM) based approach [142] and an analysis of unknown words [143, 144]. Further, this group presented a comparison of few Sinhala PoS taggers that are available to them [145]. A RESTFul PoS tagging web service created by Jayaweera and Dias [146] using the above research can still be accessed[34] via POST and GET. A hybrid PoS tagger for Sinhala language was proposed by Gunasekara et al. [147]. The study by Kothalawala et al. [148] discussed the data availability problem in NLP with a Sinhala POS tagging experiment among others. Withanage and Silva [149] proposed a stochastic POS tagger based on a small 10,000 word corpus drawn from Facebook and Twitter.

3.7 Parsers

The PoS tagged data then needs to be handed over to a parser. This is an area which is not completely solved even in English due to various inherent ambiguities in natural languages. However, in the case of English, there are systems which provide adequate results [150] even if not perfect yet. The Sinhala state of affairs, is that, the first parser for the Sinhala language was proposed by Hettige and Karunananda [62] with a model for grammar [18]. The study by Liyanage et al. [40] is concentrated on the same given that they have worked on formalizing a computational grammar for Sinhala. While they do report reasonable results, yet again, do not provide any means for the public to access the data or the tools that they have developed. Kanduboda [39] have worked on Sinhala differential object markers relevant for parsing.

The first attempt at a Sinhala parser, as mentioned above, was by Hettige and Karunananda [62] where they created prototype Sinhala morphological analyzer and a parser as part of their larger project to build an end-to-end translator system. The function of the parser is based on three dictionaries: *Base Dictionary*, *Rule Dictionary*, and *Concept Dictionary*. They are built as follows:

- **The Base Dictionary:** *prakurthi* (base words), *nipatha* (prepositions), *upasarga* (prefixes), and *vibakthi* (Irregular Verbs).

33. https://github.com/jseanm1/aruthSWSD

34. http://bit.ly/2F0jKid

TABLE 5: **Morphological Analyzers comparison**
Base: Rule-based (RB) / Machine Learning (ML)
Able to Handle Part of Speech (Handles): Yes (Y) / No (N)
Outputs: Yes (Y) / No (N) / No Information (O)
Abbreviations: Nouns (Nu), Verbs (Ve), Adjectives (Aj), Adverbs (Av), Function Words (Fn), Root (R), Person (P), Number (Nb), Gender (G), Article (A), Case (C)

	Base	Modus Operandi	Handles					Output					
			Nu	Ve	Aj	Av	Fn	R	P	Nb	G	A	C
Hettige and Karunananda [61]	RB	Finite State Automata	Y	Y	N	N	N	Y	Y	Y	Y	Y	Y
Hettige et al. [73]	RB	Agent-based	Y	Y	Y	Y	Y	N	N	N	N	N	N
Nandathilaka et al. [133]	RB	N/A	Y	N	N	N	N	Y	N	N	N	N	N
Welgama et al. [130]	ML	Morfessor algorithm	Y	Y	Y	Y	Y	Y	N	N	N	N	N
Fernando and Weerasinghe [131]	RB	Finite State Transducer	N	Y	N	N	N	Y	Y	Y	Y	N	Y
Dilshani and Dias [132]	RB	N/A	N	Y	N	N	N	O	O	O	O	O	O

- **The Rule Dictionary:** inflection rules used to generate various forms of verbs and nouns from the base words.
- **The Concept Dictionary:** synonyms and antonyms for the words found in the base dictionary.

Parsers are, in essence, a computational representation of the grammar of a natural language. As such, in building Sinhala parsers, it is crucial to create a computational model for Sinhala grammar. The first such attempt was taken by Hettige and Karunananda [18] with special consideration given to Morphology and the Syntax of the Sinhala language as an extension to their earlier work [62]. Here, it is worthy to note that, unlike in their earlier attempt [62], where they explicitly mentioned that they are building a parser, in this study [18], they use the much conservative claim of building a computational grammar. Under Morphology, they again handled Sinhala inflection. Their system is based on a Finite State Transducer (FST) and Context-Free Grammar (CFG) where they they modeled 85 rules for nouns and 18 rules for verbs. The specific implementation is more partial to a rule-based composer rather than parser. It is also worthy to note that this system could only handle simple sentences which only contained the following 8 constituents: *Attributive Adjunct of Subject*, *Subject*, *Attributive Adjunct of Object*, *Object*, *Attributive Adjunct of Predicate*, *Attributive Adjunct of the Complement of Predicate*, *Complement of Predicate*, and *Predicate*. With these, they propose the following grammar rules for Sinhala:

```
S = Subject Akkyanaya
Subject = SimpleSubject | ComplexSubject
ComplexSubject = SimpleSubject ConSub
SimpleSubject = Noun | Adjective Noun
ConSub = Conjunction SimpleSubject
Akkyanaya = VerbP | Object VerbP
Object = SimpleObject | ComplexObject
ComplexObject = Conjunction SimpleObject
SimpleObject = Noun | Adjective Noun
VerbP = Verb | Adverb Verb
```

The later work by Liyanage et al. [40] also involves formalizing a computational grammar for Sinhala. They claim that Sinhala can have any order of words in practice. However, they do not note that this is happening because practices of the spoken language, which does not share the strong SOV conventions of the written language, are slowly seeping into written text. However, they do make note of how Sinhala grammar is modeled as a head-final language [41]. They propose the Sinhala Noun Phrase (NP) to be defined as shown in equation 1 where NN is a noun which can be of types: common noun (N), pronoun (PrN) or proper noun ($PropN$). The adjectival phrase ($ADJP$) is then defined as as shown in equation 2 where: Det is a Determiner, Adj is the adjective, and Deg is an optional operator *Degrees* which can be used to intensify the meaning of the adjective in cases where the adjective is qualitative. While they note that according to Gunasekara [151], there has to three classes of adjectives (*qualitative*, *quantitative*, and *demonstrative*), they do not implement this distinction in their system. Similarly, they propose Sinhala Verb Phrase (VP) to be defined as shown in equation 3 where V is a *single verb*. They here note that they are ignoring *compound verbs* and *auxiliary verbs* in their grammar. The adverbial phrases ($ADVP$) are then recursively defined as as shown in equation 4.

$$NP = [ADJP][NN] \tag{1}$$

$$ADJP = \Big[[Det]\Big[[Deg][Adj]\Big]\Big] \tag{2}$$

$$VP = [ADVP][V] \tag{3}$$

$$ADVP = \Bigg[[NP]\Bigg[[ADVP]\Big[[Deg][ADV]\Big]\Bigg]\Bigg] \tag{4}$$

Similar to Hettige and Karunananda [18], the work by Liyanage et al. [40] also builds a CFG for Sinhala covering 10 out of the 25 types of simple sentence structures in Sinhala reported by Abhayasinghe [42]. This parser is unable to parse sentences where inanimate subjects do not consider the number. Further, sentences which contain, *compound verbs*, *auxiliary verbs*, *present participles*, or *past participles* cannot be handled by this parser. If the verbs have *imperative mood* or *negation* those too cannot be handled by this. Non-verbal sentences which end with *adjectives*, *oblique nominals*, *locative predicates*, *adverbials*, or any other language entity which is not a verb cannot be handled by this parser.

The study by Kanduboda [39] covers not the whole of Sinhala parsing but analyzes a very specific property of Sinhala observed by Aissen [152] which states that it is possible to notice Differential Object Marking (DOM) in Sinhala active sentences. Kanduboda [39] define this as the choice of */wa/* and */ta/* object markers. They further observe three unique aspects of DOM in Sinhala: (a) it is only observed in active sentences which contain transitive verbs, (b) it can occur with accusative marked nouns but not with any other cases, (c) it exists only if the sentence has placed an animate noun in the accusative position. They do a statistical analysis and provide a number of short gazetteer lists as appendixes. However, they observe that further work has to be done for this particular language rule in Sinhala given that they found some examples which proved to be exceptions to the general model which they proposed.

3.8 Named Entity Recognition Tools

As shown in Fig 1, once the text is properly parsed, it has to be processed using a Named Entity Recognition (NER) system. The first attempt of Sinahla NER was done by Dahanayaka and Weerasinghe [153]. Given that they were conducting the first study for Sinhala NER, they based their approach on NER research done for other languages. In this, they gave prominent notice to that of Indic languages. On that matter, they were the first to make the interesting observation that NER for Indic languages (including, but not limited to Sinhala) is more difficult than that of English by the virtue of the absence of a capitalization mechanic. Following prior work done on other languages, they used Conditional Random Fields (CRF) as their main model and compared it against a baseline of a Maximum Entropy (ME) model. However, they only use the *candidate word*, *Context Words* around the candidate word, and a simple analysis of Sinhala *suffixes* as their features.

The follow up work by Senevirathne et al. [154] kept the CRF model with all the previous features but did not report comparative analysis with an ME model. The innovation introduced by this work is a richer set of features. In addition to the features used by Dahanayaka and Weerasinghe [153], they introduced, *Length of the Word* as a threshold feature. They also introduced *First Word* feature after observing certain rigid grammatical rules of Sinhala. A feature of *clue Words* in the form of a subset of *Context Words* feature was first proposed by this work. Finally, they introduced a feature for *Previous Map* which is essentially the NE value of the preceding word. Some of these feature extractions are done with the help of a rule-based post-processor which utilizes context-based word lists.

The third attempt at Sinhala NER was by Manamini et al. [99] who dubbed their system *Ananya*. They inherit the CRF model and ME baseline from the work of Dahanayaka and Weerasinghe [153]. In addition to that, they take the enhanced feature list of Senevirathne et al. [154] and enrich it further more. They introduce a *Frequency of the Word* feature based on the assumption that most commonly occurring words are not NEs. Thus, they model this as a Boolean value with a threshold applied on the word frequency. They extend the *First Word* feature proposed by Senevirathne et al. [154] to a *First Word/ Last Word of a Sentence* feature noting that Sinhala grammar is of *SOV* configuration. They introduce a *(PoS) Tag* feature and a *gazetteer lists* based feature keeping in line with research done on NER in other languages. They formally introduce *clue Words*, which was initially proposed as a sub-feature by Dahanayaka and Weerasinghe [153], as an independent feature. Utilizing the fact that they have the ME model unlike Dahanayaka and Weerasinghe [153], they introduce a complementary feature to *Previous Map* named *Outcome Prior*, which uses the underlying distribution of the outcomes of the ME model. Finally, they introduce a *Cutoff Value* feature to handle the over-fitting problem.

The table 6 provides comparative summery of the discussion above. It should be noted that all three of these models only tag NEs of types: *person names*, *location names* and *organization names*. The *Ananya* system by Manamini et al. [99] is available to download at GitHub [35]. The data and code for the approaches by Dahanayaka and Weerasinghe [153] and by Senevirathne et al. [154] are not accessible to the public. Azeez and Ranathunga [155] proposed a fine-grained NER model for Sinhala building on their earlier work on NER [99] and PoS tagging [98]. Anuruddha [156] proposed a method based on reinforcement learning for Sinhala NER.

3.9 Semantic Similarity

A Sinhala semantic similarity measure has been developed for short sentences by Kadupitiya et al. [157]. This work has been then extended by Kadupitiya et al. [158] for the application use case of *short answer grading*. Data and tools for these projects are not publicly available. A sentence similarity measurement which uses Siamese neural networks was developed by Nilaxan and Ranathunga [159] where they demonstrate their results for Sinhala and Tamil. A cross-lingual document similarity measurement using the use-case of Sinhala and English was developed by Isuranga et al. [160].

3.10 Text Classification

Text classification is a popular application on the semantic layer of the NLP stack. A very basic Sinhala text classification using Naïve Bayes Classifier, Zipf's Law Behavior, and SVMs was attempted by Gallege [161]. A smaller implementation of Sinhala news classification has been attempted by de Silva [22]. As mentioned in Section 3.2, their news corpus is publicly available[36]. Another attempt on Sinhala text classification using six popular rule based algorithms was done by Lakmali and Haddela [162]. Eventhough they talk about building a corpus named *SinNG5*, they do not indicate of means for others to obtain the said corpus. Another study by Kumari and Haddela [163] utilizes the *SinNG5* corpus as the data set for their attempt to use LIME [164] for human interpretability of Sinhala document classification. However, they too do not provide access corpus. Nanayakkara and Ranathunga [165] have implemented a system which uses corpus-based similarity measures for Sinhala text classification. Gunasekara and Haddela [166] claim to have created a context aware stop word extraction method for Sinhala text classification based on simple TF-IDF. An LSTM based textual entailment system for Sinhala was proposed by Jayasinghe and Sirts

35. http://bit.ly/2XrwCoK
36. https://osf.io/tdb84/

TABLE 6: **NER system comparison**
* Denotes a baseline.
F_1 to F_{11} denotes *Context Words*, *Word Prefixes and Suffixes*, *Length of the Word*, *Frequency of the Word*, *First Word/ Last Word of a Sentence*, *(POS) Tags*, *Gazetteer Lists*, *Clue Words*, *Outcome Prior*, *Previous Map*, and *Cutoff Value*

	CRF	ME	Features										
			F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}
Dahanayaka and Weerasinghe [153]	Yes	Yes*	Yes	Yes	No	No	No	No	No	No	No	No	No
Senevirathne et al. [154]	Yes	No	Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	No
Manamini et al. [99]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

[167]. Demotte and Ranathunga [168] proposed a dual-state capsule network architecture for text classification where they demonstrated their methodology on the Sinhala data set established by Senevirathne et al. [169]. The attempt by Sameemdeen and Selvanthan [170] considers three classical machine learning algorithms (Naïve Bayes,SVM,and KNN) and then goes on to briefly discusses the pros and cons of previous attempts by: Nanayakkara and Ranathunga [165], Gunasekara and Haddela [166], Lakmali and Haddela [162], and Buddhika et al. [171]. Then they propose active learning [172, 173] as an alternative. However, no experimental results of how active learning would improve Sinhala text classification is given. The study by Bandara et al. [174] proposed and ontology-based approach for Sinhala fake news detection. However their literature survey did not cover seminal papers in OBIE such as the work by Wimalasuriya and Dou [25]. This has impacted the overall methodology that was presented. A novel, domain-idependant, and domain-adaptive text classification framework named *AdaptText* for Sinhala text classification was proposed by Kodithuwakku and Hettiarachchi [175]. A simple TF-IDF based Sinhala text classification system was proposed by Koralage [176].

3.11 Sentiment Analysis

A simple MLP based method to classify sentiments in Sinhala text was initially proposed by Medagoda [177] based on their prior work [178]. A word2vec based tool[37] for sentiment analysis of Sinhala news comments is available. A methodology for constructing a sentiment lexicon for Sinhala Language in a semi-automated manner based on a given corpus was proposed by Chathuranga et al. [179]. Demotte et al. [180] proposed a sentiment analysis system based on sentence-state LSTM Networks for Sinhala news comments. In the subsequent work [181], they used word similarity to generate a Sinhala semantic lexicon. They followed this up with a further study [169] which discussed a number of other deep learning models such as RNN and Bi-LSTM in the domain of Sinhala sentiment analysis. Jayasuriya et al. [182] proposed a method to classify Sinhala posts in the domain of sports into positive and negative class sentiments. Ranathunga and Liyanage [183] claimed that using word embedding models as semantic features can compensate for the lack of well developed language-specific linguistic or language resources in the case of analysing sentiment of Sinhala news comments. Jayasuriya et al. [184] conducted a comparative study between word N-grams and character N-grams in the task of semantic classification of Sinhala content in social media. Which they soon followed up with an ensemble approach [185]. The work by Karunarathne [186] used word embedding for to analyse the sentiment of manually annotated Sinhala Tweets. Abeyratne and Jayaratne [187] conducted a multi model analysis on classifying Sinhala songs by emotion.

The work by Jayawickrama et al. [188] used the data set released by Wijeratne and de Silva [87] to predict the reactions induced by Sinhala Facebook posts. They then extended the work [189] and compared the results obtained with their data set against that of Senevirathne et al. [169].

3.12 Hate Speech Detection

A machine learning approach to detect hate speech in Sinhala was proposed by De Silva [190]. A feature model and a data set[38] for Sinhala hate speech detection for youtube was proposed by De Saa and Ranathunga [108]. Sandaruwan et al. [191] have attempted to identify abusive Sinhala comments in social media using text mining and machine learning techniques. A cyberbullying comment classification study for Sinhala was conducted by [192] where they used classical machine learning algorithms. The study by Hettiarachchi et al. [193] used classical machine learning methods to detect hate speech in Romanized Sinhala social media posts. While the basic idea is the same, they have avoided mentioning *transliteration* in their paper. The study by Samarasinghe et al. [194] proposed using CNNs for detecting hate speech in Sinhala text. Kariyawasam [195] proposed a machine learning approach for identifying toxic Sinhala language on social media.

3.13 Other Semantic Tools

Applications of the semantic layer are more advanced than the ones below it in Figure 1. But even with the obvious lack of resources and tools, a number of attempts have been made on semantic level applications for the Sinhala Language. The earliest attempt on semantic analysis was done by Herath et al. [196] using their earlier work which dealt with Sinhala morphological analysis [50].

A deterministic process flow for automatic Sinhala text summarizing was proposed by Welgama [197]. The study by Wimalasuriya [198], which has the same name as the above work by Welgama [197], uses graph based TextRank algorithm for automatic Sinhala text summarizing.

There have been multiple attempts to do word sense disambiguation (WSD) [199–203] for Sinhala. For this, Arukgoda et al. [127] have proposed a system named *Aruth*

37. http://bit.ly/2QKI9Np

38. https://bit.ly/3FTNaMZ

based on the Lesk Algorithm[204]. An online tool[39], an API[40] of the algorithm, and code along with data on github[41] are available. For the same task, Marasinghe et al. [205] have proposed a system based on probabilistic modeling. A dialogue act recognition system which utilizes simple classification algorithms has been proposed by Palihakkara et al. [206].

3.14 Phonological Tools

On the case of phonological layer, a report on Sinhala phonetics and phonology was published by Wasala and Gamage [207]. Wickramasinghe et al. [208] discussed the practical issues in developing Sinhala Text-to-Speech and Speech Recognition systems. Based on the earlier work by Weerasinghe et al. [209], Wasala et al. [210] have developed methods for Sinhala grapheme-to-phoneme conversion along with a set of rules for schwa epenthesis. This work was then extended by Nadungodage et al. [211]. Weerasinghe et al. [212] developed a Sinhala text-to-speech system. However, it is not publicly accessible. They internally extended it to create a system capable of helping a mute person achieve synthesized real-time interactive voice communication in Sinhala [213]. A rule based approach for automatic segmentation of a small set of Sinhala text into syllables was proposed by Kumara et al. [214]. An *ew prosodic phrasing* method to help with Sinhala Text-to-Speech process was proposed by Bandara et al. [215, 216, 217]. Sodimana et al. [218] proposed a text normalization methodology for Sinhala text-to-speech systems. Further, Sodimana et al. [219] formalized a step-by-step process for building text-to-speech voices for Sinhala. Both Jayamanna [220] and Mishangi [221] have created Sinhala document readers for visually impaired persons to be used on Android devices. An OCR and Text-to-Speech system for Sinhala named Bhashitha was proposed by De Zoysa et al. [222]. The work by Lakmal et al. [223] adapted MaryTTS [224] to synthesize Sinhala speech. The study by Jayawardhana et al. [225] used *Deep Voice* [226] for Sinhala and English TTS. Anuradha and Thelijjagoda [227] proposed a machine translation system to convert Sinhala and English Braille documents into voice. A separate group has done work on Sinhala text-to-speech systems independent to above [228].

On the converse, Nadungodage et al. [229] have done a series of work on Sinhala speech recognition with special notice given to Sinhala being a resource poor language. This project divides its focus on: continuity [230], active learning [231], and speaker adaptation [232]. A Sinhala speech recognition for voice dialing which is speaker independent was proposed by Amarasingha and Gamini [233] and on the other end, a Sinhala speech recognition methodology for interactive voice response systems, which are accessed through mobile phones was proposed by Manamperi et al. [234]. Priyadarshani [235] proposes a method for speaker dependant speech recognition based on their previous work on: dynamic time warping for recognizing isolated Sinhala words [236], genetic algorithms [237], and syllable segmentation method utilizing acoustic envelopes [238].

39. http://aruth.herokuapp.com/
40. https://bit.ly/3sJEYbS
41. https://github.com/jseanm1/aruthSWSD

The method proposed by Gunasekara and Meegama [239] utilizes an HMM model for Sinhala speech-to-text. A Sinhala speech recognizer supporting bi-directional conversion between Unicode Sinhala and phonetic English was proposed by Punchimudiyanse and Meegama [240]. The work by Karunanayake et al. [241] transfer learns CNNs for transcribing free-form Sinhala and Tamil speech data sets for the purpose of classification. Dilshan [242] conducted a study for the specific use case of transcribing number sequences in continuous Sinhala speech. Gamage et al. [243] explored the use of combinational acoustic models such as Deep Neural Network - Hidden Markov Model (DNN-HMM) [244] and Subspace Gaussian Mixture Model (SGMM) [244] in Sinhala speech recognition. Karunathilaka et al. [245] explore Sinhala speech recognition using deep learning models such as: pre-trained DNN, DNN, TDNN, TDNN+LSTM.

The Sinhala speech classification system proposed by Buddhika et al. [171] does so without converting the speech-to-text. However, they report that this approach only works for specific domains with well-defined limited vocabularies.The work by Dinushika et al. [246] uses automatic speech recognition of Sinhala for speech command classification. Extending that, Kavmini et al. [247] presented a Sinhala speech command classification system which can be used for downstream applications. The work by Karunanayake et al. [248] used English phoneme-Based Automatic Speech Recognition (ASR) for intent identification in Sinhala and Tamil. Ignatius and Thayasivam [249] proposed a speaker-invariant speech-to-intent classification model with i-vector based speaker normalization, which was then evaluated on Sinhala, and Tamil speech intent data sets. The later work by Yadav et al. [250] used pre-trained embeddings for Sinhala speech intent classification.

3.15 Optical Character Recognition Applications

While it is not necessarily a component of the NLP stack shown in Fig 1, which follows the definition by Liddy [24], it is possible to swap out the bottom most phonological layer of the stack in favour of an Optical Character Recognition (OCR) and text rendering layer.

An attempt for Sinhala OCR system has been taken by Rajapakse et al. [251] before any other work has been done on the topic. Much later, a linear symmetry based approach was proposed by Premaratne and Bigun [252, 253]. They then used hidden Markov model-based optimization on the recognized Sinhala script [254]. Similarly, Hewavitharana et al. [255] used hidden Markov models for off-line Sinhala character recognition. Statistical approaches with histogram projections for Sinhala character recognition is proposed by Hewavitharana and Kodikara [256], by Ajward et al. [257], and by Madushanka et al. [258]. Karunanayaka et al. [259] also did off-line Sinhala character recognition with a use case for postal city name recognition. A separate group had attempted Sinhala OCR [260] mainly involving the nearest-neighbor method [261]. A study by Ediriweera [262] uses dictionaries to correct errors in Sinhala OCR. An early attempt for Sinhala OCR by Dias et al. [263] has been extended to be online and made available to use via desktops [264] and hand-held devices [265] with the ability to recognize handwriting. A simple neural network based approach for Sinhala OCR was utilized by Rimas et al. [266]. A fuzzy-based model for identifying printed Sinhala characters was proposed by Gunarathna et al. [267]. Premachandra et al. [268] proposes a simple back-propagation artificial neural network with hand crafted features for Sinhala character recognition. Another neural network with specialized feature extraction for Sinhala character recognition was proposed by Jayamaha and Naleer [269]. On the matter of neural networks and feature extraction, a feature selection process for Sinhala OCR was proposed by Kumara and Ragel [270]. Jayawickrama et al. [271] worked on Sinhala printed characters with special focus on handling diacritic vowels. However, they opted to refer to diacritic vowels as *modifiers* in their work. Gunawardhana and Ranathunga [272] proposed a limited approach to recognize Sinhala letters on Facebook images. A CNN-based methodology to improve printed Sinhala character OCR was proposed by Liyanage [273]. Printed Character Recognition (PCR) was used by Vasantharajan and Thayasivam [95] to create a large scale Tamil-Sinhala-English parallel corpus. A meta-study on the effects of text genre, image resolution, and algorithmic complexity needed for Sinhala OCR from books and newspapers was conducted by Anuradha et al. [274]. Anuradha et al. [275] used *Tesseract 3*[42] [276] for Sinhala OCR. An OCR and Text-to-Speech system for Sinhala named Bhashitha was proposed by De Zoysa et al. [222]. Anuradha and Thelijjagoda [227] uses OCR on Sinhala and English Braille documents on which they then run a machine translation system in order to convert them into voice. The work by de Silva and Liyanage [277] uses KNN, SVM, and a simple ANN system to recognize Sinhala Braille text.

Fernando et al. [278] claim to have created a database for handwriting recognition research in Sinhala language and further claims that the data set is available at National Science Foundation (NSF) of Sri Lanka. However, the paper provides no URLs and we were not able to find the data set on the NSF website either. The work by Karunanayaka et al. [279] is focused on noise reduction and skew correction of Sinhala handwritten words. A genetic algorithm-based approach for non-cursive Sinhala handwritten script recognition was proposed by Jayasekara and Udawatta [280]. Nilaweera et al. [281] compare projection and wavelet-based techniques for recognizing handwritten Sinhala script. Silva and Kariyawasam [282] worked on segmenting Sinhala handwritten characters with special focus on handling diacritic vowels. A comparative study of few available Sinhala handwriting recognition methods was done by Silva et al. [283]. Silva et al. [284] uses contour tracing for isolated characters in handwritten Sinhala text. A Sinhala handwriting OCR system which utilizes zone-based feature extraction has been proposed by Dharmapala et al. [285]. The study by Walawage and Ranathunga [286] specifically focuses on segmentation of overlapping and touching Sinhala handwritten characters. Silva and Jayasundere [287] focused on recognizing character modifiers in Sinhala handwriting. The similarly named studies by Mariyathas et al. [288] and Wasalthilake and Kartheeswaran [289], both utilize CNN to recognize Sinhala handwriting.

42. https://tesseract-ocr.github.io/

Summarizing on optically recognized old Sinhala text for the purpose of archival search and preservation was explored by Rathnasena et al. [290]. The work of Peiris [291] also focused on OCR for ancient Sinhala inscriptions. A neural network based method for recognizing ancient Sinhala inscriptions was proposed by Karunarathne et al. [292]. Chanda et al. [293] proposed a Gaussian kernel SVM based method for word-wise Sinhala, Tamil, and English script identification.

3.16 Translators

A series of work has been done by a group towards English to Sinhala translation as mentioned in some of the above subsections. This work includes; building a morphological analyzer [61], lexicon databases [64], a transliteration system [65], an evaluation model [70], a computational model of grammar [18], and a multi-agent solution [76]. After working on human-assisted machine translation [66], Hettige and Karunananda [69, 71] have attempted to establish a theoretical basics for English to Sinhala machine translation. A very simplistic web based translator was proposed [67, 68]. The same group have worked on a Sinhala ontology generator for the purpose of machine translation [75] and a phrase level translator [77] based on the previous work on a multi-agent system for translation [74]. Further, an application of the English to Sinhala translator on the use case of selected text for reading was implemented [72]. They later continued their work on multi agent English to Sinhala translation with the AGR organizational model [78].

Another group independently attempted English-to-Sinhala machine translation [294] with a statistical approach [295]. Wijerathna et al. [296] and De Silva et al. [297] have proposed simple rule based translators. An example-based method applied on the English-Sinhala sentence aligned government domain corpus was proposed by Silva and Weerasinghe [298]. A translator based on a look-up system was proposed by Vidanaralage et al. [299]. In a preprint, Joseph et al. [300] proposes an evolutionary algorithm for Sinhala to English translation with a basis of Pointwise Mutual Information (PMI) and claims that the code will be shared once the paper is accepted. However, they do not report any quantitative results to be compared and the reported qualitative results are also superficial. Pushpananda et al. [301] utilized statistical machine translation to translate between Sinhala and Tamil. Fernando et al. [302] tries to solve the Out of vocabulary (OOV) problem for Sinhala in the context of Sinhala-English-Tamil statistical machine translation.

Fonseka et al. [303] used Byte Pair Encoding (BPE) for English to Sinhala neural machine translation. As an another solution to the OOV problem, an analysis of subword techniques to improve English to Sinhala Neural Machine Translation (NMT) was conducted by Naranpanawa et al. [304]. A data augmentation method to expand bilingual lexicon terms based on case-markers for the purpose of solving the OOV problem in the domain of NMT was proposed by Fernando et al. [305]. Epaliyana et al. [306] proposed iterative filtering and data selection be used to improve Sinhala-English NMT. Perera et al. [307] used English Part-of-Speech (PoS) tags to improve English to Sinhala NMT.

The early work by Goonetilleke et al. [308] attempted Sinhala transliteration through the Latin alphabet. However, their work does not use the word *transliteration* and instead focuses on the predictive aspect. Priyadarshani et al. [309] used statistical machine learning for transliteration of names between Sinhala, Tamil, and English. A rule-based system on trigrams was proposed by Liwera and Ranathunga [310] for Singlish to Sinhala transliteration of social media text. A Singlish to Sinhala converter which uses an LSTM was proposed by De Silva [311]. A rule based approach for the same was proposed by de Silva and Ahangama [312].

Most of the cross Sinhala and Tamil work has been done in the domain of machine translation. A neural machine translation for Sinhala and Tamil languages was initiated by Tennage et al. [313, 314]. Then they further enhanced it with transliteration and byte pair encoding [315] and used synthetic training data to handle the rare word problem [316]. This project produced *Si-Ta* [317] a machine translation system of Sinhala and Tamil official documents. In the statistical machine translation front, Farhath et al. [318] worked on integrating bilingual lists. The attempts by Weerasinghe [319] and Sripirakas et al. [320] were also focused on statistical machine translation while Jeyakaran and Weerasinghe [321] attempted a kernel regression method. A yet another attempt was made by Pushpananda et al. [322] which they later extended with some quality improvements [323]. An attempt on real-time direct translation between Sinhala and Tamil was done by Rajpirathap et al. [324]. Dilshani et al. [325] have done a study on the linguistic divergence of Sinhala and Tamil languages in respect to machine translation. Mokanarangan [326] claims to have built a named entity translator between Sinhala and Tamil for official government documents. But this work is locked behind an institutional repository wall and thus is not accessible by other researchers. Arukgoda et al. [327] studied the possibility of using deep learning techniques to improve Sinhala-Tamil translation which they further improved later [328]. Pramodya et al. [329] compared Transformers, Recurrent Neural Networks, and Statistical Machine Translation (SMT) in the context of Tamil to Sinhala machine translation. The work by Nissanka et al. [330] used monolingual word embedding to improve NMT between Sinhala and Tamil. Thillainathan et al. [331] uses pre-trained mBART models for six directional translations between Sinhala, Tamil, and English.

There have been attempts to link Sinhala NLP with Japanese by Herath et al. [21, 79, 80], Thelijjagoda et al. [81], Thelijjagoda et al. [332], and Kanduboda [8]. There has also been an attempt to use dictionary-based machine translation [333] between Sinhala and the liturgical language of Buddhism, Pali [334–336]. The study by Anuradha and Thelijjagoda [227] uses machine translation on the unique application of converting Sinhala and English Braille documents, which they have run OCR on, into voice.

3.17 Spelling and Grammar

The open-source data driven approach proposed by Wasala et al. [337, 338] claims to be able to check and correct spelling errors in Sinhala. The approach by Jayalatharachchi et al. [339] attempts to obtain synergy between two algorithms for the same purpose. These efforts [337, 339] were then extended by Subhagya et al. [340]. A rule-based Sinhala spell checker named SinSpell based on *Hunspell*[43] was introduced by Liyanapathirana et al. [341]. They have also made the tool available[44] online for use. The study by Sithamparanathan and Uthayasanker [342] extended the *Generic Environment for context-aware spell correction* to handle Sinhala and Tamil. Sonnadara et al. [107] created a benchmark data set for Sinhala spell correction along with a neural model. A model for detecting grammatical mistakes in Sinhala was developed by Pabasara and Jayalal [343]. They followed this up with an grammatical error detection and correction model [344]. Gunasekara et al. [345] used annotation projection for semantic role labeling for Sinhala.

3.18 Chat Bots

A simple Sinhala chat bot which utilizes a small knowledge base has been proposed by Hettige and Karunananda [63]. A study on the effect of word embeddings on a Sinhala chat bot was conducted by Gamage et al. [346] where they used, the fasttext model trained by Facebook [101–103], on a RASA[45] chat bot. A Sinhala chat bot for train information was proposed by Harshani [347]. Similarly, the tool proposed by Chandrasena et al. [348] serves as a chat bot-based recommendation system for Sri Lankan traditional dancers. The chat bot discussed by Kumanayake [349] has the very specific purpose of answering user inquiries about the degree programs at University of Ruhuna.

3.19 Miscellaneous Applications

In this section, we discuss NLP tools and research which are either hard to categorize under above sections or are reasonably involving multiples of them. The first miscellaneous application of Sinhala NLP is sign language, strides have been made in the domains of computer interpreting for written Sinhala [350] and animation of finger-spelled words and number signs [351]. Dissanayake and Hettige [352] implemented a question and answer generator for Sinhala with the limited PoS: pronouns, adjectives, verbs, and adverbs.

Fernando [353] proposed a method for inexact matching of Sinhala proper names. A study on determining canonical word order of colloquial Sinhala sentences using priority information was conducted by Kanduboda and Tamaoka [354] which they later extended [355–357]. Jayakody et al. [358] uses simple KNN and SVM methods on a PoS tagged Sinhala corpus to create a question-answering system which they name *Mahoshadha*. An extremely simple plagiarism detection tool which only uses n-grams of simply tokenized text was proposed by Basnayake et al. [359]. Another simple plagiarism detection tool that uses synonymy and Hyponymy-Hypernymy (which they call *Generalization* in the paper) was attempted by Rajamanthri and Thelijjagoda [360]. They later extended this work [361] to propose a more advanced plagiarism detection tool which uses Internet resources.

43. http://hunspell.github.io/
44. http://nlp-tools.uom.lk/sinspell/
45. https://rasa.com/

A dataset consisting of Sinhala documents drawn from Sri Lankan news websites was published by Jayawickrama et al. [106] along with the benchmark misinformation classification models. The work by Liyanage and Ranathunga [362, 363] attempt to use LSTMs for mathematical word problem generation in Sinhala and other languages. The problem of recognizing Sinhala and English code-mixed data where the Sinhala text is written in Singlish was explored by Smith and Thayasivam [364] using an XGB classifier and a CRF model building on their previous work [365], which analysed such data. Sandathara et al. [366] proposed a system which they named *Arunalu* that they claimed to use Voice recognition, Natural Language Processing, Machine Learning, and Deep Learning concepts to help individuals with dyslexia overcome problems of reading Sinhala. Rajitha et al. [367] used the data set that they introduced in their previous work [90] to prove that task specific supervised distance learning metrics outperform their unsupervised counterparts, for document alignment. Kumari and Hettiarachchi [368] proposed an algorithm for Sinhala topic modelling based on LDA [369] and RAKE [370]. Arambewela et al. [371] proposed a Sinhala writing assistant tool utilising CNNS. A news aggregator with news categorization, comment filtering, and two types of recommendation systems was proposed by Malsha et al. [372]. Jayaweera et al. [373] used classical machine learning methods to propose dynamic stop word removal. They claim to have released a corpus of 100,000+ Sinhala documents in their paper. But they provide no information on where to obtain this corpus. A trending topic detection model for Sinhala tweets using simple clustering and ranking algorithms was proposed by Jayasekara and Ahangama [374]. Sandamini et al. [375] proposed a post recommendation system, which supports Singlish, for social media.

4 Primary Sources

Even though the main objective of this survey is to cover NLP tools and research, we noticed that much of these NLP tools and research depend on primary sources of Sinhala language such as printed books in the role of knowledge sources and ground truth. Therefore, for the benefit of other researchers who venture into Sinhala NLP, we decided to add a short introduction to the available primary sources of Sinhala language used by their peers. We note that the body of work by a single scholar, Disanayake [2, 37, 376–386], is quite prominent in the case of being used for NLP applications. For formal introduction of the language, the books by Disanayake [2] and Perera [3] are commonly used. In cases which deal with the Sinhala alphabet, the introduction by Indrasena [387] and by Disanayake [376, 377] have been used. An analysis of modern Sinhala linguistics has been done by Jayathilake [388] and by Pallatthara and Weihene [38]. The early study by Henadeerage [389] covers a number of topics on the Sinhala language such as grammatical relations, argument structure, phrase structure and focus constructions.

As we discussed in Section 3.3, a number of printed Sinhala dictionaries exist, Malalasekera [109] being the most prominent English-Sinahala dictionary among them. In addition to that seminal work, previous researchers of Sinhala NLP have utilized a number of other dictionaries of various configurations such as: English-Sinhala [111, 114, 115], Sinhala-Sinhala [110, 113], and English-Sinhala-Tamil [112].

A number of NLP applications have utilized first sources intended to teach children [390–394] or foreigners [151, 395, 396][46]. This makes sense given that an introduction written for children would start with basic principles and thus be ideal for crafting rule based NLP systems and an introduction written for foreigners would have Sinhala language described in terms of English, making easy the process of rule based translation of English NLP tools to Sinhala.

For applications where a rule based approach for Sinhala spelling correction is utilized, the books by Disanayake [383, 384], by Koparahewa [397], and by Gair and Karunatillake [398] are used to provide a basis. A number of NLP applications which handle spoken Sinhala in the capacity of phonological layer (Section 3.14) applications or otherwise, make note of the fact that spoken Sinhala is considerably distinguishable from written Sinhala, as such, they refer primary sources which explicitly deal with spoken Sinhala [37, 382, 386, 392, 395, 399, 400].

Primary sources used in NLP application for Sinhala grammar are varied. A number of them provide overviews of the entirety of Sinhala grammar [19, 38, 41, 385, 401–410]. There are specific primary sources focusing on verbs [381, 393, 411], nouns [380, 394], prepositions [392], compounds [379], derivation [378], case system [412], and sentence structure [42] of the Sinhala language. The book by Rajapaksha [413] is commonly used in NLP applications as a guide for word tagging and punctuation mark handling. NLP studies that tackle the hard problem of handling questions expressed in Sinhala often refer to the book by Kariyakarawana [414]. Kekulawala [415] has aptly discussed the much controversial topic of the situation of *future tense* of Sinhala.

5 Conclusion

At this point, a reader might think, there seems to be a significant number of implementations of NLP for Sinhala. Therefore, how can one justify listing Sinhala as a resource poor language? The important point which is missing in that assumption is that in the cases of almost all of the above listed implementations and findings, the only thing that is publicly available for a researcher is a set of research papers. The corpora, tools, algorithm, and anything else that were discovered through these research are either locked away as properties of individual research groups or worse lost to the time with crashed ancient servers, lost hard drives, and expired web hosts. This reason and probably academic/research rivalry have caused these separate research groups not to cite or build upon the works of each-other. In many cases where similar work is done, it is a re-hashing on the same ideas adopted from resource rich languages because of, the unavailability of (or the reluctance to), referring and building on work done by another group. This has resulted in multiple groups building multiple foundations behind closed doors but no one ending up with a completed end-to-end NLP work-flow. In conclusion,

46. Note that [396] is an extension of [151].

what can be said is that, even though there are islands of implementations done for Sinhala NLP, they are of very small scale and/or are usually not readily accessible for further use and research by other researchers. Thus, so far, sadly, Sinhala stays a resource poor language.

Acknowledgments

The authors would like to thank Romain Egele for checking the examples we have provided in French for their accuracy. Similarly, the authors would also like to thank Shravan Kale for checking the examples we have provided in Hindi for their accuracy.

References

[1] R. Englebretson and C. Genetti, "Santa barbara papers in linguistics: Proceeding from the workshop on sinhala linguistics," *Santa Barbara, CA: Department of Linguistics at the University of California, Santa Barbara*, 2005.
[2] J. B. Disanayake, *National Languages of Sri-Lanka: Sinhala*. Department of cultural affairs, 1976.
[3] T. G. Perera, *Sinhala Bhashawa*, 1985.
[4] L. Bauer, *Linguistics Student's Handbook*. Edinburgh University Press, 2007.
[5] Department of Census and Statistics Sri Lanka. Percentage of population aged 10 years and over in major ethnic groups by district and ability to speak sinhala, tamil and english languages. [Online]. Available: https://goo.gl/nnVZSd
[6] J. W. Gair and W. Karunatilaka, *Literary Sinhala.* ERIC, Cornell University. New York, 1974.
[7] H. Young. A language family tree - in pictures — education — the guardian. [Online]. Available: https://www.theguardian.com/education/gallery/2015/jan/23/a-language-family-tree-in-pictures
[8] A. B. Kanduboda, "The role of animacy in determining noun phrase cases in the sinhalese and japanese languages," *Science of words*, vol. 24, pp. 5–20, 2011.
[9] P. Fernando, "Palaeographical development of the brahmi script in ceylon from 3rd century bc to 7th century ad," 1949.
[10] D. Bandara, N. Warnajith, A. Minato, and S. Ozawa, "Creation of precise alphabet fonts of early brahmi script from photographic data of ancient sri lankan inscriptions," *Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition*, vol. 3, no. 3, pp. 33–39, 2012.
[11] P. T. Daniels and W. Bright, *The world's writing systems*. Oxford University Press on Demand, 1996.
[12] M. H. Sirisoma, "Brahmi inscriptions of sri lanka from 3rd century bc to 65 ad," pp. 3–54, 1990.
[13] M. Dias, "Lakdiwa sellipiwalin heliwana sinhala bhashawe prathyartha namayange vikashanaya," *Department of Archaeology, Colombo Sri Lanka*, p. 1, 1996.
[14] A. S. Hettiarachchi, "Investigation of 2nd, 3rd and 4th century inscriptions," *Inscriptions: Volume Two, Archaeological Department Centenary (1890–1990), Commemorative Series. Colombo: Department of Archaeology*, pp. 57–104, 1990.
[15] S. Paranavitana and S. L. P. Depārtamēntuva, *Inscriptions of Ceylon*. Dept. of Archaeology, 1970.
[16] R. Salomon, *Indian epigraphy: a guide to the study of inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan languages*. Oxford University Press, 1998.
[17] H. Falk, *Schrift im alten Indien: ein Forschungsbericht mit Anmerkungen*. Gunter Narr Verlag, 1993, vol. 56.
[18] B. Hettige and A. S. Karunananda, "Computational model of grammar for english to sinhala machine translation," in *Advances in ICT for Emerging Regions (ICTer), 2011 International Conference on*. IEEE, 2011, pp. 26–31.
[19] A. M. Gunasekara, *A Comprehensive Grammar of the Sinhalese Language*. Asian Educational Services, New Delhi, Madras, India, 1986.
[20] H. P. Ray, *The archaeology of seafaring in ancient South Asia*. Cambridge University Press, 2003.
[21] A. Herath, Y. Hyodo, Y. Kawada, T. Ikeda, and S. Herath, "A practical machine translation system from japanese to modern sinhalese," *Gifu University*, pp. 153–162, 1994.
[22] N. de Silva, "Sinhala Text Classification: Observations from the Perspective of a Resource Poor Language," 2015.
[23] Y. Wijeratne, N. de Silva, and Y. Shanmugarajah, "Natural Language Processing for Government: Problems and Potential," *LIRNEasia*, 2019.
[24] E. D. Liddy, "Natural language processing," 2001.
[25] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *Journal of Information Science*, vol. 36, no. 3, pp. 306–323, 2010.
[26] U. Consortium *et al.*, "The unicode standard: A technical introduction," *online document, http://www. unicode. org/unicode/standards/principles. html*, 1996.
[27] R. A. Van der Sandt, "Presupposition projection as anaphora resolution," *Journal of semantics*, vol. 9, no. 4, pp. 333–377, 1992.
[28] S. Lappin and H. J. Leass, "An algorithm for pronominal anaphora resolution," *Computational linguistics*, vol. 20, no. 4, pp. 535–561, 1994.
[29] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational linguistics*, vol. 27, no. 4, pp. 521–544, 2001.
[30] V. Ng and C. Cardie, "Improving machine learning approaches to coreference resolution," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 104–111.
[31] R. Mitkov, *Anaphora resolution*. Routledge, 2014.
[32] J. Preston and M. J. M. Bishop, *Views into the Chinese room: New essays on Searle and artificial intelligence*. OUP, 2002.
[33] D. R. Shanahan, "A living document: reincarnating the research article," *Trials*, vol. 16, no. 1, pp. 1–5, 2015.
[34] N. M. Sopinka, L. E. Coristine, M. C. DeRosa, C. M. Rochman, B. L. Owens, and S. J. Cooke, "Envisioning the scientific paper of the future," *Facets*, vol. 5, no. 1, pp. 1–16, 2020.
[35] G. H. Fairbanks, J. W. Gair, and M. W. S. D. Silva, *Colloquial Sinhalese.* ERIC, Cornell University. New York, 1968.
[36] T. Miyagishi, "Accusative subject of subordinate clause in literary sinhala," *Journal of Yasuda Women's University*, vol. 33, 2005.
[37] J. B. Disanayake, *Say it in Sinhala*. Lake House Investments Limited, 1985.
[38] S. Pallatthara and P. Weihene, *Sinhala Grammar in Linguistic Perspective*, 1966.
[39] A. P. B. Kanduboda, "On the usage of sinhalese differential object markers object marker /wa/ vs. object marker /ta/," *Theory and Practice in Language Studies*, vol. 3, no. 7, p. 1081, 2013.
[40] C. Liyanage, R. Pushpananda, D. L. Herath, and R. Weerasinghe, "A computational grammar of Sinhala," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2012, pp. 188–200.
[41] W. S. Karunatilaka, *Sinhala Bhasa Vyakaranaya*. M D Gunasena Publishers, Colombo, 1997.
[42] A. A. Abhayasinghe, *Sinhala bhashave sarala vakya vibagaya*, 1998.
[43] C. Jany, "The relationship between case marking and s, a, and o in spoken sinhala," *Santa Barbara Papers in Linguistics*, no. 17, pp. 68–84, 2006.
[44] J. Garland, "Morphological typology and the complexity of nominal morphology in sinhala," *Santa Barbara Papers in Linguistics*, no. 17, pp. 1–19, 2005.
[45] M. Henderson, "Between lexical and lexico-grammatical classification: nominal classification in sinhala," *Santa Barbara Papers in Linguistics*, p. 29, 2005.
[46] T. Noguchi, "Shinharago nyuumon [introductory to the sinhalese language]," *Tokyo: Daigaku Shorin*, 1984.
[47] T. Miyagishi, "A comparison of word order between japanese and sinhalese," *Bulletin of Japanese Language and Literature*, pp. 101–107, 2003.
[48] D. Chandralal, *Sinhala*. John Benjamins Publishing, 2010, vol. 15.
[49] S. P. Singh, A. Kumar, P. Sahu, and P. Verma, "Syntax based machine translation using blended methodology," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. IEEE, 2016, pp. 242–247.
[50] S. Herath, T. Ikeda, S. Yokoyama, H. Isahara, and S. Ishizaki, "Sinhalese morphological analysis: a step towards machine processing of sinhalese," in *[Proceedings 1989] IEEE International Workshop on Tools for Artificial Intelligence*. IEEE, 1989, pp. 100–107.
[51] S. Herath, T. Ikeda, S. Ishizaki, and Y. Anzai, "Formalization of sinhalese morphology," in *Proc. of the 40th National Congress of IPSJ. Jyouhou Syori Gakkai*, vol. 1, 1990, pp. 327–328.

[52] V. K. Samaranayake, J. B. Disanayaka, and S. T. Nandasara, "A standard code for sinhala characters," *Proceedings, 9th Annual Sessions of the Computer Society of Sri Lanka, Colombo*, 1989.

[53] V. K. Samaranayake, S. T. Nandasara, J. B. Disanayaka, A. R. Weerasinghe, and H. Wijayawardhana, "An introduction to unicode for sinhala characters," *University Of Colombo School of Computing*, 2003.

[54] G. Dias and A. Goonetilleke, "Development of standards for Sinhala computing," in *1st Regional Conference on ICT and E-Paradigms*, 2004.

[55] G. V. Dias, "Challenges of enabling it in the sinhala language," in *27th Internationalization and Unicode Conference*, 2005.

[56] A. R. Weerasinghe, D. L. Herath, and K. Gamage, "The sinhala collation sequence and its representation in unicode," *Localization Focus*, 2006.

[57] G. Sandeva, "Design and evaluation of user-friendly yet efficient sinhala input methods," 2009.

[58] S. Herath, S. Ishizaki, T. Ikeda, Y. Anzai, and H. Aiso, "Machine processing of sinhala natural language: a step toward intelligent systems," *Cybernetics and systems*, vol. 22, no. 3, pp. 331–348, 1991.

[59] S. T. Nandasara, "From the past to the present: Evolution of computing in the sinhala language," *IEEE Annals of the History of Computing*, vol. 31, no. 1, pp. 32–45, 2009.

[60] S. T. Nandasara and Y. Mikami, "Bridging the digital divide in sri lanka: some challenges and opportunities in using sinhala in ict," *International Journal on Advances in ICT for Emerging Regions (ICTer)*, vol. 8, no. 1, 2016.

[61] B. Hettige and A. S. Karunananda, "A morphological analyzer to enable english to sinhala machine translation," in *Information and Automation, 2006. ICIA 2006. International Conference on*. IEEE, 2006, pp. 21–26.

[62] ——, "A parser for sinhala language-first step towards english to sinhala machine translation," in *Industrial and Information Systems, First International Conference on*. IEEE, 2006, pp. 583–587.

[63] ——, "First sinhala chatbot in action," *Proceedings of the 3rd Annual Sessions of Sri Lanka Association for Artificial Intelligence (SLAAI), University of Moratuwa*, 2006.

[64] ——, "Developing lexicon databases for english to sinhala machine translation," in *Industrial and Information Systems, 2007. ICIIS 2007. International Conference on*. IEEE, 2007, pp. 215–220.

[65] ——, "Transliteration system for english to sinhala machine translation," in *Industrial and Information Systems, 2007. ICIIS 2007. International Conference on*. IEEE, 2007, pp. 209–214.

[66] ——, "Using human-assisted machine translation to overcome language barrier in sri lanka," *Proceedings of 4th Annual session of Sri Lanka Association for Artificial Intelligence*, p. 10, 2007.

[67] ——, "Web-based english-sinhala translator in action," in *2008 4th International Conference on Information and Automation for Sustainability*. IEEE, 2008, pp. 80–85.

[68] ——, "Web-based english to sinhala selected texts translation system," *Sri Lanka Association for Artificial Intelligence*, p. 26, 2008.

[69] ——, "Theoretical based approach to english to sinhala machine translation," in *2009 International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2009, pp. 380–385.

[70] ——, "An evaluation methodology for english to sinhala machine translation," in *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on*. IEEE, 2010, pp. 31–36.

[71] ——, "Varanageema: A theoretical basics for english to sinhala machine translation," in *Sri Lanka Association for Artificial Intelligence (SLAAI)*, 2010.

[72] B. Hettige, G. Rzevski, and A. S. Karunananda, "Selected text machine translator for english to sinhala," 2013.

[73] B. Hettige, A. S. Karunananda, and G. Rzevski, "Multi-agent system technology for morphological analysis," *Proceedings of the 9th Annual Sessions of Sri Lanka Association for Artificial Intelligence (SLAAI), Colombo*, 2012.

[74] ——, "Masmt: A multi-agent system development framework for english-sinhala machine translation," *International Journal of Computational Linguistics and Natural Language Processing (IJCLNLP)*, vol. 2, no. 7, pp. 411–416, 2013.

[75] ——, "Sinhala ontology generator for english to sinhala machine translation," in *Proc. of KDU International Research Conference*, 2014.

[76] ——, "A multi-agent solution for managing complexity in english to sinhala machine translation," *Complex Systems: Fundamentals & Applications*, vol. 90, p. 251, 2016.

[77] ——, "Phrase-level english to sinhala machine translation with multi-agent approach," in *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2017, pp. 1–6.

[78] ——, "Masmt4: The agr organizational model-based multi-agent system development framework for machine translation," in *Inventive Computation and Information Technologies*. Springer, 2021, pp. 691–702.

[79] A. Herath, Y. Hyodo, T. Ikeda, and S. Herath, "Generation of sinhalese units from japanese bunsetsu structure," 1993.

[80] A. Herath, Y. Hyodo, Y. Kunieda, T. Ikeda, and S. Herath, "Bunsetsu-based japanese-sinhalese translation system," *Information sciences*, vol. 90, no. 1-4, pp. 303–319, 1996.

[81] S. Thelijjagoda, Y. Imai, and T. Ikeda, "Japanese-sinhalese mt system (jaw/sinhalese)," in *Proceedings of Asian Symposium on Natural Language Processing to Overcome Language Barriers, IJCNLP-04 Satellite Symposium*, 2004.

[82] D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. H. N. D. de Silva, and G. Dias, "Implementing a Corpus for Sinhala Language," in *Symposium on Language Technology for South Asia 2015*, 2015.

[83] ——, "Comparison between performance of various database systems for implementing a language corpus," in *International Conference: Beyond Databases, Architectures and Structures*. Springer, May 2015, pp. 82–91.

[84] R. Weerasinghe, D. Herath, and V. Welgama, "Corpus-based sinhala lexicon," in *Proceedings of the 7th Workshop on Asian Language Resources*. Association for Computational Linguistics, 2009, pp. 17–23.

[85] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english," *arXiv preprint arXiv:1902.01382*, 2019.

[86] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[87] Y. Wijeratne and N. de Silva, "Sinhala language corpora and stopwords from a decade of sri lankan facebook," *arXiv preprint arXiv:2007.07884*, 2020.

[88] M. Banón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Espla-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn *et al.*, "Paracrawl: Web-scale acquisition of parallel corpora," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4555–4567.

[89] I. Caswell, J. Kreutzer, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. J. O. Suárez, I. Orife, K. Ogueji, R. A. Niyongabo, T. Q. Nguyen, M. Müller, A. Müller, S. H. Muhammad, N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. F. P. Dossou, S. Dlamini, N. de Silva, S. Çabuk Ballı, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. A. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi, "Quality at a glance: An audit of web-crawled multilingual datasets," *arXiv preprint arXiv:2103.12028*, 2021.

[90] D. Sachintha, L. Piyarathna, C. Rajitha, and S. Ranathunga, "Exploiting parallel corpora to improve multilingual embedding based document and sentence alignment," *arXiv preprint arXiv:2106.06766*, 2021.

[91] R. A. Hameed, N. Pathirennehelage, A. Ihalapathirana, M. Z. Mohamed, S. Ranathunga, S. Jayasena, G. Dias, and S. Fernando, "Automatic creation of a sentence aligned sinhala-tamil parallel corpus," in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, 2016, pp. 124–132.

[92] M. Z. Mohamed, A. Ihalapathirana, R. A. Hameed, N. Pathirennehelage, S. Ranathunga, S. Jayasena, and G. Dias, "Automatic creation of a word aligned sinhala-tamil parallel corpus," in *Engineering Research Conference (MERCon), 2017 Moratuwa*. IEEE, 2017, pp. 425–430.

[93] F. Farhath, P. Theivendiram, S. Ranathunga, S. Jayasena, and G. Dias, "Improving domain-specific smt for low-resourced languages using data from different domains," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[94] R. Jenarthanan, Y. Senarath, and U. Thayasivam, "Actsea: annotated corpus for tamil & sinhala emotion analysis," in *2019 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2019,

pp. 49–53.
[95] C. Vasantharajan and U. Thayasivam, “Tamizhi-Net OCR: Creating A Quality Large Scale Tamil-Sinhala-English Parallel Corpus Using Deep Learning Based Printed Character Recognition (PCR),” *arXiv preprint arXiv:2109.05952*, 2021.
[96] S. Fernando, S. Ranathunga, S. Jayasena, and G. Dias, “Comprehensive part-of-speech tag set and svm based pos tagger for sinhala,” in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, 2016, pp. 173–182.
[97] N. Dilshani, S. Fernando, S. Ranathunga, S. Jayasena, and G. Dias, “A comprehensive part of speech (pos) tag set for sinhala language.” The Third International Conference on Linguistics in Sri Lanka, ICLSL 2017 . . . , 2017.
[98] S. Fernando and S. Ranathunga, “Evaluation of different classifiers for sinhala pos tagging,” in *2018 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2018, pp. 96–101.
[99] S. A. P. M. Manamini, A. F. Ahamed, R. A. E. C. Rajapakshe, G. H. A. Reemal, S. Jayasena, G. V. Dias, and S. Ranathunga, “Ananya-a named-entity-recognition (ner) system for sinhala language,” in *Moratuwa Engineering Research Conference (MERCon), 2016*. IEEE, 2016, pp. 30–35.
[100] A. Liyanage, S. Ranathunga, and S. Jayasena, “Bilingual lexical induction for sinhala-english using cross lingual embedding spaces,” in *2021 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2021, pp. 579–584.
[101] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, “Fasttext. zip: Compressing text classification models,” *arXiv preprint arXiv:1612.03651*, 2016.
[102] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
[103] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 427–431.
[104] D. Lakmal, S. Ranathunga, S. Peramuna, and I. Herath, “Word embedding evaluation for sinhala,” in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 1874–1881.
[105] D. Herath, K. Gamage, and A. Malalasekara, “Research report on sinhala lexicon,” *Langugae Technology Research Laboratory, UCSC*.
[106] V. Jayawickrama, A. Ranasinghe, D. C. Attanayake, and Y. Wijeratne, “A corpus and machine learning models for fake news classification in sinhala,” 2021.
[107] C. Sonnadara, S. Ranathunga, and S. Jayasena, “Sinhala Spell Correction A Novel Benchmark with Neural Spell Correction.”
[108] E. De Saa and L. Ranathunga, “Self-reflective and introspective feature model for hate content detection in sinhala youtube videos,” in *2020 From Innovation to Impact (FITI)*, vol. 1. IEEE, 2020, pp. 1–6.
[109] G. P. Malalasekera, “English-sinhalese dictionary.” 1967.
[110] D. B. Jayathilake, *Sinhala Shabdakoshaya (Sinhala Dictionary), Prathama Bhagaya (Vol 1)*. Sri Lankan branch of Royal Asian Society, 1937.
[111] S. Maitipe, *Gunasena English-Sinhalese Concise Dictionary*. MD Gunasena & Co., 1988.
[112] A. Weerasinghe and C. P. Weerasinghe, “Godage english-sinhala-tamil dictionary,” *Sri Lanka: S. Godage and brothers, Godage book shop*, vol. 661, 1999.
[113] H. Wijayathunga, *Maha Sinhala Sabdakoshaya*. M. D. Gunasena & Co. Ltd., Colombo, 2003.
[114] S. Ranaweera, *Wasana English-Sinhala Dictionary*. Wasana Prakashakayo, Dankotuwa, Sri Lanka, 2004.
[115] K. Gunaratne, *Ratna English-Sinhalese Dictionary*. Ratna Poth Prakasakayo, 513, Maradana road, Colombo 10, Sri Lanka, 2006.
[116] M. Kulatunga. Madura english-sinhala dictionary - online language translator. [Online]. Available: https://maduraonline.com/
[117] A. Wasala and R. Weerasinghe, “Ensitip: a tool to unlock the english web,” in *11th international conference on humans and computers, Nagaoka University of Technology, Japan*, 2008, pp. 20–23.
[118] L. Samarawickrama and B. Hettige, “Requirements for an english-sinhala smart bilingual dictionary: A review.”
[119] Department of Official Languages, Sri Lanka. Tri-lingual dictionary. [Online]. Available: https://www.trilingualdictionary.lk/
[120] A. Weerasinghe and G. Dias, “Construction of a multilingual place name database for sri lanka,” 2013.
[121] G. A. Miller, “Wordnet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[122] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.
[123] J. J. Jiang and D. W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy,” in *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING’97*. Citeseer, 1997.
[124] N. de Silva, D. Dou, and J. Huang, “Discovering inconsistencies in pubmed abstracts through ontology-based information extraction,” in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 362–371.
[125] I. Wijesiri, M. Gallage, B. Gunathilaka, M. Lakjeewa, D. Wimalasuriya, G. Dias, R. Paranavithana, and N. de Silva, “Building a wordnet for Sinhala,” in *Proceedings of the Seventh Global WordNet Conference*, 2014, pp. 100–108.
[126] Sinhala wordnet. [Online]. Available: http://www.wordnet.lk/
[127] J. Arukgoda, V. Bandara, S. Bashani, V. Gamage, and D. Wimalasuriya, “A word sense disambiguation technique for sinhala,” in *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*. IEEE, 2014, pp. 207–211.
[128] V. Welgama, D. L. Herath, C. Liyanage, N. Udalamatta, R. Weerasinghe, and T. Jayawardana, “Towards a sinhala wordnet,” in *Proceedings of the Conference on Human Language Technology for Development*, 2011.
[129] S. Herath, T. Ikeda, S. Ishizaki, Y. Anzai, and H. Aiso, “Analysis system for sinhalese unit structure,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 4, no. 1, pp. 29–48, 1992.
[130] V. Welgama, R. Weerasinghe, and M. Niranjan, “Evaluating a machine learning approach to sinhala morphological analysis,” in *Proceedings of the 10th International Conference on Natural Language Processing, Noida, India*, 2013.
[131] N. Fernando and R. Weerasinghe, “A morphological parser for sinhala verbs,” in *Proceedings of the International Conference on Advances in ICT for Emerging Regions*, 2013.
[132] W. S. N. Dilshani and G. Dias, “A corpus-based morphological analysis of sinhala verbs.” The Third International Conference on Linguistics in Sri Lanka, ICLSL 2017 . . . , 2017.
[133] M. Nandathilaka, S. Ahangama, and G. T. Weerasuriya, “A rule-based lemmatizing approach for sinhala language,” in *2018 3rd International Conference on Information Technology Research (ICITR)*. IEEE, 2018, pp. 1–5.
[134] V. Welgama, R. Weerasinghe, and N. Mahesan, “Defining the gold standard definitions for the morphology of sinhala words.”
[135] K. Kariyawasam, S. Senanayake, and P. S. Haddela, “A rule based stemmer for sinhala language,” in *2019 14th Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2019, pp. 326–331.
[136] S. Y. Senanayake, K. Kariyawasam, and P. S. Haddela, “Enhanced tokenizer for sinhala language,” in *2019 National Information Technology Conference (NITC)*. IEEE, 2019, pp. 84–89.
[137] K. Kumarasinghe, G. Dias, and I. Herath, “SinMorphy: A Morphological Analyzer for the Sinhala Language,” in *2021 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2021, pp. 681–686.
[138] D. L. Herath and A. R. Weerasinghe, “A stochastic part of speech tagger for sinhala,” in *Proceedings of the 06th International Information Technology Conference*, 2004, pp. 27–28.
[139] A. J. P. M. P. Jayaweera and N. G. J. Dias, “Evaluation of stochastic based tagging approach for sinhala language,” 2012.
[140] M. Jayasuriya and A. R. Weerasinghe, “Learning a stochastic part of speech tagger for sinhala,” in *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*. IEEE, 2013, pp. 137–143.
[141] A. J. P. M. P. Jayaweera and N. G. J. Dias, “Part of speech (pos) tagger for sinhala language,” 2011.
[142] ——, “Hidden markov model based part of speech tagger for sinhala language,” *arXiv preprint arXiv:1407.2989*, 2014.
[143] ——, “Unknown words analysis in pos tagging of sinhala language,” in *Advances in ICT for Emerging Regions (ICTer), 2014 International Conference on*. IEEE, 2014, pp. 270–270.
[144] ——, “Handling issues with unknown words in pos tagging.” Book of Abstracts, Annual Research Symposium 2014, 2014.

[145] M. Jayaweera and N. G. J. Dias, "Comparison of part of speech taggers for sinhala language," 2016.

[146] A. J. P. M. P. Jayaweera and N. G. J. Dias, "Restful pos tagging web service for sinhala language," in *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2015, pp. 50–57.

[147] D. Gunasekara, W. V. Welgama, and A. R. Weerasinghe, "Hybrid part of speech tagger for sinhala language," in *Advances in ICT for Emerging Regions (ICTer), 2016 Sixteenth International Conference on*. IEEE, 2016, pp. 41–48.

[148] B. Kothalawala, R. Weerasinghe, and P. Kumarasinghe, "Online learning for solving data availability problem in natural language processing." in *NL4AI@ AI* IA*, 2019.

[149] S. G. Withanage and T. Silva, "A stochastic part of speech tagger for the sinhala language based on social media data mining," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 137–142.

[150] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[151] A. M. Gunasekara, *A comprehensive grammar of the Sinhalese language: adapted for the use of English readers and prescribed for the Civil Service examinations*. GJA Skeen, 1891.

[152] J. Aissen, "Differential object marking: Iconicity vs. economy," *Natural Language & Linguistic Theory*, vol. 21, no. 3, pp. 435–483, 2003.

[153] J. K. Dahanayaka and A. R. Weerasinghe, "Named entity recognition for sinhala language," in *Advances in ICT for Emerging Regions (ICTer), 2014 International Conference on*. IEEE, 2014, pp. 215–220.

[154] K. U. Senevirathne, N. S. Attanayake, A. W. M. H. Dhananjanie, W. A. S. U. Weragoda, A. Nugaliyadde, and S. Thelijjagoda, "Conditional random fields based named entity recognition for sinhala," in *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2015, pp. 302–307.

[155] R. Azeez and S. Ranathunga, "Fine-grained named entity recognition for sinhala," in *2020 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2020, pp. 295–300.

[156] H. Anuruddha, "Reinforcement learning for sinhala named entity recognition," Ph.D. dissertation, 2021.

[157] J. C. S. Kadupitiya, S. Ranathunga, and G. Dias, "Sinhala short sentence similarity calculation using corpus-based and knowledge-based similarity measures," in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, 2016, pp. 44–53.

[158] ——, "Sinhala short sentence similarity measures using corpus-based simi-larity for short answer grading," in *6th Workshop on South and Southeast Asian Natural Language Processing*, 2017, pp. 44–53.

[159] S. Nilaxan and S. Ranathunga, "Monolingual sentence similarity measurement using siamese neural networks for sinhala and tamil languages," in *2021 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2021, pp. 567–572.

[160] U. Isuranga, J. Sandaruwan, U. Athukorala, and G. Dias, "Improved cross-lingual document similarity measurement," 2020.

[161] S. Gallege, "Analysis of sinhala using natural language processing techniques," 2010.

[162] K. B. N. Lakmali and P. S. Haddela, "Effectiveness of rule-based classifiers in sinhala text categorization," in *2017 National Information Technology Conference (NITC)*. IEEE, 2017, pp. 153–158.

[163] P. K. S. Kumari and P. S. Haddela, "Use of lime for human interpretability in sinhala document classification," in *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*. IEEE, 2019, pp. 97–102.

[164] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[165] P. Nanayakkara and S. Ranathunga, "Clustering sinhala news articles using corpus-based similarity measures," in *2018 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2018, pp. 437–442.

[166] S. V. S. Gunasekara and P. S. Haddela, "Context aware stopwords for sinhala text classification," in *2018 National Information Technology Conference (NITC)*. IEEE, 2018, pp. 1–6.

[167] S. H. Jayasinghe and K. Sirts, "Deep learning textual entailment system for sinhala language," 2019.

[168] P. Demotte and S. Ranathunga, "Dual-state capsule networks for text classification," *arXiv preprint arXiv:2109.04762*, 2021.

[169] L. Senevirathne, P. Demotte, B. Karunanayake, U. Munasinghe, and S. Ranathunga, "Sentiment Analysis for Sinhala Language using Deep Learning Techniques," *arXiv preprint arXiv:2011.07280*, 2020.

[170] A. Sameemdeen and N. Selvanthan, "Topic classification using active learning for sinhala language documents," in *2021 Asian Conference on Innovation in Technology (ASIANCON)*. IEEE, 2021, pp. 1–5.

[171] D. Buddhika, R. Liyadipita, S. Nadeeshan, H. Witharana, S. Javasena, and U. Thayasivam, "Domain specific intent classification of sinhala speech data," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 197–202.

[172] B. Novak, D. Mladenič, and M. Grobelnik, "Text classification with active learning," in *From Data and Information Analysis to Knowledge Engineering*. Springer, 2006, pp. 398–405.

[173] B. Yang, J.-T. Sun, T. Wang, and Z. Chen, "Effective multi-label active learning for text classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 917–926.

[174] O. Bandara, D. Jayarathne, D. Shashinika, and L. Ranathunga, "Ontology based fake news detection for sinhala language," in *2021 6th International Conference on Information Technology Research (ICITR)*. IEEE, 2021, pp. 1–6.

[175] Y. Kodithuwakku and S. Hettiarachchi, "Adapttext: A novel framework for domain-independent automated sinhala text classification," in *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*. IEEE, 2021, pp. 240–245.

[176] A. D. Koralage, "Sinclassify-sinhala text classification system," Ph.D. dissertation, 2019.

[177] N. Medagoda, "Sentiment analysis on morphologically rich languages: An artificial neural network (ann) approach," in *Artificial Neural Network Modelling*. Springer, 2016, pp. 377–393.

[178] N. Medagoda, S. Shanmuganathan, and J. Whalley, "Sentiment lexicon construction using sentiwordnet 3.0," in *2015 11th International Conference on Natural Computation (ICNC)*. IEEE, 2015, pp. 802–807.

[179] P. D. T. Chathuranga, S. A. S. Lorensuhewa, and M. A. L. Kalyani, "Sinhala sentiment analysis using corpus based sentiment lexicon," in *International Conference on Advances in ICT for Emerging Regions (ICTer)*, vol. 1, 2019, p. 7.

[180] P. Demotte, L. Senevirathne, B. Karunanayake, U. Munasinghe, and S. Ranathunga, "Sentiment Analysis of Sinhala News Comments using Sentence-State LSTM Networks," in *2020 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2020, pp. 283–288.

[181] B. Karunanayake, U. Munasinghe, P. Demotte, L. Senevirathne, and S. Ranathunga, "Sinhala Sentiment Lexicon Generation using Word Similarity," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 77–82.

[182] P. Jayasuriya, R. Munasinghe, and S. Thelijjagoda, "Sentiment classification of sinhala content in social media: A comparison between word n-grams and character n-grams."

[183] S. Ranathunga and I. U. Liyanage, "Sentiment analysis of sinhala news comments," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 4, pp. 1–23, 2021.

[184] P. Jayasuriya, R. Munasinghe, and S. Thelijjagoda, "Sentiment classification of sinhala content in social media: A comparison between stemmers and n-gram features," in *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, pp. 134–139.

[185] ——, "Sentiment classification of sinhala content in social media: An ensemble approach," in *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, pp. 140–145.

[186] W. I. Karunarathne, "Sentiment analysis of sinhala tweets," Ph.D. dissertation, 2020.

[187] K. Abeyratne and K. Jayaratne, "Classification of sinhala songs based on emotions," in *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, vol. 250. IEEE, 2019, pp. 1–10.

[188] V. Jayawickrama, G. Weeraprameshwara, N. de Silva, and Y. Wijeratne, "Seeking sinhala sentiment: Predicting facebook reactions of sinhala posts," *arXiv preprint arXiv:2112.00468*, 2021.

[189] G. Weeraprameshwara, V. Jayawickrama, N. de Silva, and Y. Wijeratne, "Sentiment analysis with deep learning models: A comparative study on a decade of sinhala language facebook data," *arXiv preprint arXiv:2201.03941*, 2022.

[190] D. H. A. De Silva, "An approach to hate speech detection," 2019.

[191] S. T. Sandaruwan, S. A. S. Lorensuhewa, and K. Munasinghe, "Identification of abusive sinhala comments in social media using text mining and machine learning techniques," *ICTer*, vol. 13, no. 1, 2020.

[192] H. I. Amali and S. Jayalal, "Classification of cyberbullying sinhala language comments on social media," in *2020 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2020, pp. 266–271.

[193] N. Hettiarachchi, R. Weerasinghe, and R. Pushpanda, "Detecting hate speech in social media articles in romanized sinhala," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 250–255.

[194] S. Samarasinghe, R. Meegama, and M. Punchimudiyanse, "Machine learning approach for the detection of hate speech in sinhala unicode text," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 65–70.

[195] S. Kariyawasam, "A machine learning approach in the identification of sinhala toxic language on social media," Ph.D. dissertation, 2019.

[196] S. Herath, S. Ishizaki, T. Ikeda, Y. Anzai, and H. Aiso, "Syntactic and semantic analysis of sinhala: a step towards intelligence computing systems," in *Proceedings. 5th IEEE International Symposium on Intelligent Control 1990*. IEEE, 1990, pp. 316–324.

[197] W. V. Welgama, "Automatic text summarization for sinhala," 2012.

[198] O. S. Wimalasuriya, "Automatic text summarization for sinhala," Ph.D. dissertation, 2019.

[199] D. Yarowsky, "Word-sense disambiguation using statistical models of roget's categories trained on large corpora," in *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1992, pp. 454–460.

[200] N. Ide and J. Véronis, "Introduction to the special issue on word sense disambiguation: the state of the art," *Computational linguistics*, vol. 24, no. 1, pp. 2–40, 1998.

[201] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *33rd annual meeting of the association for computational linguistics*, 1995, pp. 189–196.

[202] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in *International conference on intelligent text processing and computational linguistics*. Springer, 2002, pp. 136–145.

[203] R. Navigli, "Word sense disambiguation: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 2, p. 10, 2009.

[204] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in *Proceedings of the 5th annual international conference on Systems documentation*. Citeseer, 1986, pp. 24–26.

[205] C. Marasinghe, S. Herath, and A. Herath, "Word sense disambiguation of sinhala language with unsupervised learning," in *Proc. International Conference on Information Technology and Applications*, 2002, pp. 25–29.

[206] S. Palihakkara, D. Sahabandu, A. Shamsudeen, C. Bandara, and S. Ranathunga, "Dialogue act recognition for text-based sinhala," in *Proceedings of the 12th International Conference on Natural Language Processing*, 2015, pp. 367–375.

[207] A. Wasala and K. Gamage, "Research report on phonetics and phonology of sinhala," *Language Technology Research Laboratory, University of Colombo School of Computing*, vol. 35, 2005.

[208] R. I. P. Wickramasinghe, K. H. Kumara, and N. G. J. Dias, "Practical issues in the development of tts and sr for the sinhala language," 2007.

[209] R. Weerasinghe, A. Wasala, and K. Gamage, "A rule based syllabification algorithm for sinhala," in *International Conference on Natural Language Processing*. Springer, 2005, pp. 438–449.

[210] A. Wasala, R. Weerasinghe, and K. Gamage, "Sinhala grapheme-to-phoneme conversion and rules for schwa epenthesis," in *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006, pp. 890–897.

[211] T. Nadungodage, C. Liyanage, A. Prerera, R. Pushpananda, and R. Weerasinghe, "Sinhala g2p conversion for speech processing," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pp. 112–116.

[212] R. Weerasinghe, A. Wasala, V. Welgama, and K. Gamage, "Festival-si: A sinhala text-to-speech system," in *International Conference on Text, Speech and Dialogue*. Springer, 2007, pp. 472–479.

[213] M. S. Amarasekara, K. M. N. S. Bandara, B. V. A. I. Vithana, D. H. De Silva, and A. Jayakody, "Real-time interactive voice communication-for a mute person in sinhala (rtivc)," in *2013 8th International Conference on Computer Science & Education*. IEEE, 2013, pp. 671–675.

[214] K. H. Kumara, N. G. J. Dias, and H. Sirisena, "Automatic segmentation of given set of sinhala text into syllables for speech synthesis," pp. 53–62, 2007.

[215] W. M. C. Bandara, W. M. S. Lakmal, T. D. Liyanagama, S. V. Bulathsinghala, G. Dias, and S. Jayasena, "A ew prosodic phrasing method for sinhala language," 2017.

[216] W. M. C. Bandara, S. V. Bulathsinghala, W. M. S.Lakmal, T. D. Liyanagama, G. Dias, and S. Jayasena, "Sinhala text to speech system," 2009.

[217] W. M. C. Bandara, V. M. S. Lakmal, T. D. Liyanagama, S. V. Bulathsinghala, G. Dias, and S. Jayasena, "A new prosodic phrasing model for sinhala language," 2013.

[218] K. Sodimana, P. De Silva, R. Sproat, A. Theeraphol, C. F. Li, A. Gutkin, S. Sarin, and K. Pipatsrisawat, "Text normalization for bangla, khmer, nepali, javanese, sinhala, and sundanese text-to-speech systems," 2018.

[219] K. Sodimana, K. Pipatsrisawat, L. Ha, M. Jansche, O. Kjartansson, P. De Silva, and S. Sarin, "A step-by-step process for building tts voices using open source data and framework for bangla, javanese, khmer, nepali, sinhala, and sundanese," 2018.

[220] D. S. Jayamanna, "Android based sinhala document reader for visually impaired persons," 2014.

[221] A. K. P. D. Mishangi, "Android based sinhala document reader for visually impaired people," 2018.

[222] D. S. S. De Zoysa, J. M. Sampath, E. M. P. De Seram, D. M. I. D. Dissanayake, L. Wijerathna, and S. Thelijjagoda, "Project bhashitha-mobile based optical character recognition and text-to-speech system," in *2018 13th International Conference on Computer Science & Education (ICCSE)*. IEEE, 2018, pp. 1–5.

[223] M. Lakmal, K. Methmini, D. Rupasinghe, D. Hettiarachchi, V. Piyawardana, M. Senarathna, S. Reyal, and K. Pulasinghe, "Adapting MaryTTS for Synthesizing Sinhalese Speech to Communicate with Children," in *2021 6th International Conference on Information Technology Research (ICITR)*. IEEE, 2021, pp. 1–6.

[224] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[225] P. Jayawardhana, A. Aponso, N. Krishnarajah, and A. Rathnayake, "An intelligent approach of text-to-speech synthesizers for english and sinhala languages," in *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2019, pp. 229–234.

[226] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[227] K. S. Anuradha and S. Thelijjagoda, "Machine translation system to convert sinhala and english braille documents into voice," in *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*. IEEE, 2020, pp. 7–16.

[228] L. Nanayakkara, C. Liyanage, P.-T. Viswakula, T. Nagungodage, R. Pushpananda, and R. Weerasinghe, "A human quality text to speech system for sinhala," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pp. 157–161.

[229] T. Nadungodage, R. Weerasinghe, and M. Niranjan, "Speech recognition for low resourced languages: Efficient use of training data for sinhala speech recognition by active learning."

[230] T. Nadungodage and R. Weerasinghe, "Continuous sinhala speech recognizer," in *Conference on Human Language Technology for Development, Alexandria, Egypt*, 2011, pp. 2–5.

[231] T. Nadungodage, R. Weerasinghe, and M. Niranjan, "Efficient use of training data for Sinhala speech recognition using active learning," in *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*. IEEE, 2013, pp. 149–153.

[232] ——, "Speaker Adaptation Applied to Sinhala Speech Recognition." *Int. J. Comput. Linguistics Appl.*, vol. 6, no. 1, pp. 117–129,

2015.

[233] W. G. T. N. Amarasingha and D. D. A. Gamini, "Speaker independent sinhala speech recognition for voice dialling," in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*. IEEE, 2012, pp. 3–6.

[234] W. Manamperi, D. Karunathilake, T. Madhushani, N. Galagedara, and D. Dias, "Sinhala speech recognition for interactive voice response systems accessed through mobile phones," in *2018 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2018, pp. 241–246.

[235] P. G. N. Priyadarshani, "Speaker dependent speech recognition on a selected set of sinhala words," 2012.

[236] P. G. N. Priyadarshani, N. G. J. Dias, and A. Punchihewa, "Dynamic time warping based speech recognition for isolated sinhala words," in *2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 2012, pp. 892–895.

[237] ——, "Genetic algorithm approach for sinhala speech recognition," in *2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 2012, pp. 896–899.

[238] P. G. N. Priyadarshani and N. G. J. Dias, "Automatic segmentation of separately pronounced sinhala words into syllables," 2011.

[239] M. K. H. Gunasekara and R. G. N. Meegama, "Real-time translation of discrete sinhala speech to unicode text," in *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2015, pp. 140–145.

[240] M. Punchimudiyanse and R. G. N. Meegama, "Unicode sinhala and phonetic english bi-directional conversion for sinhala speech recognizer," in *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2015, pp. 296–301.

[241] Y. Karunanayake, U. Thayasivam, and S. Ranathunga, "Transfer learning based free-form speech command classification for low-resource languages," in *Proceedings of the 57th Conference of the Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 288–294.

[242] K. A. D. C. Dilshan, "Transcribing number sequences in continuous sinhala speech," 2018.

[243] B. Gamage, R. Pushpananda, R. Weerasinghe, and T. Nadungodage, "Usage of combinational acoustic models (dnn-hmm and sgmm) and identifying the impact of language models in sinhala speech recognition," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 17–22.

[244] D. Giuliani and B. BabaAli, "Large vocabulary children's speech recognition with dnn-hmm and sgmm acoustic modeling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[245] H. Karunathilaka, V. Welgama, T. Nadungodage, and R. Weerasinghe, "Low-resource sinhala speech recognition using deep learning," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 196–201.

[246] T. Dinushika, L. Kavmini, P. Abeyawardhana, U. Thayasivam, and S. Jayasena, "Speech command classification system for sinhala language based on automatic speech recognition," in *2019 International Conference on Asian Language Processing (IALP)*. IEEE, 2019, pp. 205–210.

[247] L. Kavmini, T. Dinushika, U. Thayasivam, and S. Jayasena, "Improved speech command classification system for sinhala language based on automatic speech recognition," *International Journal of Asian Language Processing*, p. 2050009, 2020.

[248] Y. Karunanayake, U. Thayasivam, and S. Ranathunga, "Sinhala and tamil speech intent identification from english phoneme based asr," in *2019 International Conference on Asian Language Processing (IALP)*. IEEE, 2019, pp. 234–239.

[249] A. Ignatius and U. Thayasivam, "Speaker-invariant speech-to-intent classification for low-resource languages," in *International Conference on Speech and Computer*. Springer, 2021, pp. 279–290.

[250] H. Yadav, A. Gupta, S. K. Rallabandi, A. W. Black, and R. R. Shah, "Intent classification using pre-trained embeddings for low resource languages," *arXiv preprint arXiv:2110.09264*, 2021.

[251] R. K. Rajapakse, A. R. Weerasinghe, and E. K. Seneviratne, "A neural network based character recognition system for sinhala script," *Department of Statistics and Computer Science, University of Colombo*, 1995.

[252] H. L. Premaratne and J. Bigun, "Recognition of printed sinhala characters using linear symmetry," in *The 5th Asian Conference on Computer Vision*, 2002, pp. 23–25.

[253] ——, "A segmentation-free approach to recognise printed sinhala script using linear symmetry," *Pattern recognition*, vol. 37, no. 10, pp. 2081–2089, 2004.

[254] H. L. Premaratne, E. Järpe, and J. Bigun, "Lexicon and hidden markov model-based optimisation of the recognised sinhala script," *Pattern recognition letters*, vol. 27, no. 6, pp. 696–705, 2006.

[255] S. Hewavitharana, H. C. Fernando, and N. D. Kodikara, "Off-line sinhala handwriting recognition using hidden markov models." in *ICVGIP*, 2002.

[256] S. Hewavitharana and N. D. Kodikara, "A statistical approach to sinhala handwriting recognition," in *Proc. of the International Information Technology Conference (IITC), Colombo, Sri Lanka*, 2002.

[257] S. Ajward, N. Jayasundara, S. Madushika, and R. Ragel, "Converting printed sinhala documents to formatted editable text," in *2010 Fifth International Conference on Information and Automation for Sustainability*. IEEE, 2010, pp. 138–143.

[258] P. T. C. Madushanka, R. Bandara, and L. Ranathunga, "Sinhala handwritten character recognition by using enhanced thinning and curvature histogram based method," in *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2017, pp. 46–50.

[259] M. L. M. Karunanayaka, N. D. Kodikara, and G. D. S. P. Wimalaratne, "Off line sinhala handwriting recognition with an application for postal city name recognition," *Il'I'C 2004*, 2004.

[260] R. Weerasinghe, A. Wasala, D. Herath, and V. Welgama, "Nlp applications of sinhala: Tts & ocr," in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

[261] A. R. Weerasinghe, D. L. Herath, and N. P. K. Medagoda, "A nearest-neighbor based algorithm for printed sinhala character recognition," *Innovations for a Knowledge Economy*, p. 11, 2006.

[262] D. N. Ediriweera, "Improviing the accuracy of the output of sinhala ocr by using a dictionary," Ph.D. dissertation, University of Moratuwa Sri Lanka, 2012.

[263] G. Dias, T. N. P. Patikirikorala, C. I. Arambewela, R. P. M. Darshana, and N. D. Alahendra, "Sinhala optical character recognition for desktops," 2013.

[264] G. Dias, T. N. P. Patikirikorala, C. I. Arambewela, R. P. M. Darshani, and N. D. Alahendra, "Online sinhala handwritten character recognition for desktops," 2013.

[265] M. H. P. Ranmuthugala, G. D. N. C. Pathiragoda, S. H. C. Jayasundara, G. Dias, and A. S. Karunananda, "Online sinhala handwritten character recognition on handheld devices," *Innovations for a Knowledge Economy*, p. 1, 2006.

[266] M. Rimas, R. P. Thilakumara, and P. Koswatta, "Optical character recognition for sinhala language," in *2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS)*. IEEE, 2013, pp. 149–153.

[267] G. I. Gunarathna, M. A. P. Chamikara, and R. G. Ragel, "A fuzzy based model to identify printed sinhala characters," in *7th International Conference on Information and Automation for Sustainability*. IEEE, 2014, pp. 1–6.

[268] H. W. H. Premachandra, C. Premachandra, T. Kimura, and H. Kawanaka, "Artificial neural network based sinhala character recognition," in *International Conference on Computer Vision and Graphics*. Springer, 2016, pp. 594–603.

[269] J. M. H. M. Jayamaha and H. M. M. Naleer, "Feature extraction technique based character recognition using artificial neural network for sinhala characters," 2016.

[270] T. N. Kumara and R. Ragel, "A systematic feature selection process for a sinhala character recognition system," in *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*. IEEE, 2016, pp. 1–6.

[271] B. R. Jayawickrama, L. Ranathunga, K. L. Mahaliyanaarachchi, L. G. B. Subhagya, and W. H. A. Nimasha, "Letter segmentation and modifier detection in printed sinhala signage," in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2018, pp. 203–208.

[272] S. Gunawardhana and L. Ranathunga, "Segmentation and identification of presence of sinhala characters in facebook images," in *2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2018, pp. 77–82.

[273] K. L. N. D. Liyanage, "Improving Sinhala OCR using Deep Learning," 2018.

[274] I. Anuradha, C. Liyanage, and R. Weerasinghe, "Estimating the effects of text genre, image resolution and algorithmic complexity needed for sinhala optical character recognition," *International*

Journal on Advances in ICT for Emerging Regions (ICTer), vol. 14, no. 3, 2021.

[275] I. Anuradha, C. Liyanage, H. Wijayawardhana, and R. Weerasinghe, "Deep learning based sinhala optical character recognition (ocr)," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 298–299.

[276] R. Smith, "An overview of the tesseract ocr engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.

[277] N. de Silva and S. Liyanage, "Sinhala braille character recognizer."

[278] H. C. Fernando, N. D. Kodikara, and S. Hewavitharana, "A database for handwriting recognition research in sinhala language." in *ICDAR*, 2003, pp. 1262–1264.

[279] M. L. M. Karunanayaka, C. A. Marasinghe, and N. D. Kodikara, "Thresholding, noise reduction and skew correction of sinhala handwritten words." in *MVA*, 2005, pp. 355–358.

[280] B. Jayasekara and L. Udawatta, "Non-cursive sinhala handwritten script recognition: A genetic algorithm based alphabet training approach," in *Proceedings of the International Conference on Information and Automation*, 2005.

[281] N. P. T. I. Nilaweera, H. L. Premeratne, and D. U. J. Sonnadara, "Comparison of projection and wavelet based techniques in recognition of sinhala handwritten scripts," in *Proceedings of the 25th National IT Conference*, 2007.

[282] C. Silva and C. Kariyawasam, "Segmenting sinhala handwritten characters," *International Journal of Conceptions on Computing and Information Technology*, vol. 2, no. 4, pp. 22–26, 2014.

[283] C. M. Silva, N. D. Jayasundere, and C. Kariyawasam, "State of handwriting recognition of modern sinhala script," 2014.

[284] ——, "Contour tracing for isolated sinhala handwritten character recognition," in *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2015, pp. 25–31.

[285] K. A. K. N. D. Dharmapala, W. P. M. V. Wijesooriya, C. P. Chandrasekara, U. K. A. U. Rathnapriya, and L. Ranathunga, "Sinhala handwriting recognition mechanism using zone based feature extraction," 2017.

[286] K. S. A. Walawage and L. Ranathunga, "Segmentation of overlapping and touching sinhala handwritten characters," in *2018 3rd International Conference on Information Technology Research (ICITR)*. IEEE, 2018, pp. 1–6.

[287] C. M. Silva and N. Jayasundere, "Character modifier combinations recognition in sinhala handwriting."

[288] J. Mariyathas, V. Shanmuganathan, and B. Kuhaneswaran, "Sinhala handwritten character recognition using convolutional neural network," in *2020 5th International Conference on Information Technology Research (ICITR)*. IEEE, 2020, pp. 1–6.

[289] W. Wasalthilake and T. Kartheeswaran, "Sinhala handwritten character recognition using convolution neural networks," 2020.

[290] K. A. M. P. Rathnasena, K. M. S. J. Kumarasinghe, D. T. P. Paranavitharana, D. V. A. U. Dayarathne, and L. Ranathunga, "Summarization based approach for old sinhala text archival search and preservation," in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2018, pp. 182–188.

[291] T. M. T. H. Peiris, "Recognition of inscriptions in ancient sri lanka," 2012.

[292] K. G. N. D. Karunarathne, K. V. Liyanage, D. A. S. Ruwanmini, K. Dias, and S. Nandasara, "Recognizing ancient sinhala inscription characters using neural network technologies," *Internationa Journal of Scientific Emgineering and Applied Sciences*, vol. 3, no. 1, 2017.

[293] S. Chanda, S. Pal, and U. Pal, "Word-wise sinhala tamil and english script identification using gaussian kernel svm," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[294] J. Liyanapathirana and R. Weerasinghe, "English to sinhala machine translation: Towards better information access for sri lankans," in *Conference on Human Language Technology for Development*, 2011, pp. 182–186.

[295] J. U. Liyanapathirana, "A statistical approach to english and sinhala translation," 2013.

[296] L. Wijerathna, W. L. S. L. Somaweera, S. L. Kaduruwana, Y. V. Wijesinghe, D. I. De Silva, K. Pulasinghe, and S. Thellijjagoda, "A translator from sinhala to english and english to sinhala (sees)," in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*. IEEE, 2012, pp. 14–18.

[297] D. De Silva, A. Alahakoon, I. Udayangani, V. Kumara, D. Kolonnage, H. Perera, and S. Thelijjagoda, "Sinhala to english language translator," in *2008 4th International Conference on Information and Automation for Sustainability*. IEEE, 2008, pp. 419–424.

[298] A. M. Silva and R. Weerasinghe, "Example based machine translation for english-sinhala translations," in *Proceedings of the 09th International IT Conference*, 2008, pp. 27–28.

[299] A. J. Vidanaralage, A. U. Illangakoon, S. Y. Sumanaweera, C. Pavithra, and S. Thelijjagoda, "Sinhala language decoder," in *2018 National Information Technology Conference (NITC)*. IEEE, 2018, pp. 1–5.

[300] J. Joseph, W. Chathurika, A. Nugaliyadde, and Y. Mallawarachchi, "Evolutionary algorithm for sinhala to english translation," *arXiv preprint arXiv:1907.03202*, 2019.

[301] R. Pushpananda, R. Weerasinghe, and M. Niranjan, "Statistical machine translation from and into morphologically rich and low resourced languages," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2015, pp. 545–556.

[302] A. Fernando, S. Ranathunga, and G. Dias, "Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation," *arXiv preprint arXiv:2011.02821*, 2020.

[303] T. Fonseka, R. Naranpanawa, R. Perera, and U. Thayasivam, "English to sinhala neural machine translation," in *IALP*, 2020.

[304] R. Naranpanawa, R. Perera, T. Fonseka, and U. Thayasivam, "Analyzing subword techniques to improve english to sinhala neural machine translation," *International Journal of Asian Language Processing*, vol. 30, no. 04, p. 2050017, 2020.

[305] A. Fernando, G. Dias, and S. Ranathunga, "Data augmentation and list integration for improving domain-specific sinhala-english-tamil statistical machine translation," 2021.

[306] K. Epaliyana, S. Ranathunga, and S. Jayasena, "Improving Back-Translation with Iterative Filtering and Data Selection for Sinhala-English NMT," in *2021 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2021, pp. 438–443.

[307] R. Perera, T. Fonseka, R. Naranpanawa, and U. Thayasivam, "Improving english to sinhala neural machine translation using part-of-speech tag," *arXiv preprint arXiv:2202.08882*, 2022.

[308] S. Goonetilleke, Y. Hayashi, Y. Itoh, and F. Kishino, "Srishell primo: A predictive sinhala text input system," in *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 2008.

[309] H. Priyadarshani, M. Rajapaksha, M. Ranasinghe, K. Sarveswaran, and G. Dias, "Statistical machine learning for transliteration: Transliterating names between sinhala, tamil and english," in *2019 International Conference on Asian Language Processing (IALP)*. IEEE, 2019, pp. 244–249.

[310] W. M. P. Liwera and L. Ranathunga, "Combination of trigram and rule-based model for singlish to sinhala transliteration by focusing social media text," in *2020 From Innovation to Impact (FITI)*, vol. 1. IEEE, 2020, pp. 1–5.

[311] A. D. De Silva, "Singlish to sinhala converter using machine learning," 2020.

[312] L. de Silva and S. Ahangama, "Singlish to sinhala transliteration using rule-based approach," in *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, pp. 162–167.

[313] P. Tennage, P. Sandaruwan, M. Thilakarathne, A. Herath, S. Ranathunga, S. Jayasena, and G. Dias, "Neural machine translation for sinhala and tamil languages," in *Asian Language Processing (IALP), 2017 International Conference on*. IEEE, 2017, pp. 189–192.

[314] P. N. Tennage, M. W. D. P. Sandaruwan, J. K. M. M. Thilakarathne, A. N. Herath, S. Ranathunga, S. Jayasena, and G. Dias, "Neural machine translation for sinhala-tamil," 2017.

[315] P. Tennage, A. Herath, M. Thilakarathne, P. Sandaruwan, and S. Ranathunga, "Transliteration and byte pair encoding to improve tamil to sinhala neural machine translation," in *2018 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2018, pp. 390–395.

[316] P. Tennage, P. Sandaruwan, M. Thilakarathne, A. Herath, and S. Ranathunga, "Handling rare word problem using synthetic training data for sinhala and tamil neural machine translation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[317] S. Ranathunga, F. Farhath, U. Thayasivam, S. Jayasena, and G. Dias, "Si-ta: Machine translation of sinhala and tamil official

documents," in *2018 National Information Technology Conference (NITC)*. IEEE, 2018, pp. 1–6.

[318] F. Farhath, S. Ranathunga, S. Jayasena, and G. Dias, "Integration of bilingual lists for domain-specific statistical machine translation for sinhala-tamil," in *2018 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2018, pp. 538–543.

[319] R. Weerasinghe, "A statistical machine translation approach to sinhala-tamil language translation," *Towards an ICT enabled Society*, p. 136, 2003.

[320] S. Sripirakas, A. R. Weerasinghe, and D. L. Herath, "Statistical machine translation of systems for sinhala-tamil," in *Advances in ICT for Emerging Regions (ICTer), 2010 International Conference on*. IEEE, 2010, pp. 62–68.

[321] M. Jeyakaran and R. Weerasinghe, "A novel kernel regression based machine translation system for sinhala-tamil translation," in *Proceedings of 4th Annual UCSC Research Symposium*, 2013.

[322] R. Pushpananda, R. Weerasinghe, and M. Niranjan, "Towards sinhala tamil machine translation," in *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*. IEEE, 2013, pp. 288–288.

[323] ——, "Sinhala-tamil machine translation: Towards better translation quality," in *Proceedings of the Australasian Language Technology Association Workshop 2014*, 2014, pp. 129–133.

[324] S. Rajpirathap, S. Sheeyam, K. Umasuthan, and A. Chelvarajah, "Real-time direct translation system for sinhala and tamil languages," in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2015, pp. 1437–1443.

[325] W. S. N. Dilshani, S. Yashothara, R. T. Uthayasanker, and S. Jayasena, "Linguistic divergence of sinhala and tamil languages in machine translation," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 13–18.

[326] T. Mokanarangan, "Translation of named entities between sinhala and tamil for official government documents," 2019.

[327] A. Arukgoda, A. R. Weerasinghe, and R. Pushpananda, "Improving sinhala-tamil translation through deep learning techniques," 2019.

[328] A. S. Arukgoda, "Improving sinhala–tamil translation through deep learning techniques," Ph.D. dissertation, 2021.

[329] A. Pramodya, R. Pushpananda, and R. Weerasinghe, "A comparison of transformer, recurrent neural networks and SMT in Tamil to Sinhala MT," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 155–160.

[330] L. Nissanka, B. Pushpananda, and A. Weerasinghe, "Exploring neural machine translation for sinhala-tamil languages pair," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 202–207.

[331] S. Thillainathan, S. Ranathunga, and S. Jayasena, "Fine-Tuning Self-Supervised Multilingual Sequence-To-Sequence Models for Extremely Low-Resource NMT," in *2021 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2021, pp. 432–437.

[332] S. Thelijjagoda, Y. Imai, and T. Ikeda, "Japanese-sinhalese machine translation system jaw/sinhalese," *Journal of the National Science Foundation of Sri Lanka*, vol. 35, no. 2, 2007.

[333] R. M. M. Shalini and B. Hettige, "Dictionary based machine translation system for pali to sinhala," in *SLAAI-International Conference on Artificial Intelligence*, 2017, p. 23.

[334] R. C. Childers, *A dictionary of the Pali language*. Trübner, 1875.

[335] S. Salaville, *An introduction to the study of eastern liturgies*. Sands & Company, 1938.

[336] A. J. Liddicoat, "Choosing a liturgical language," *Australian Review of Applied Linguistics*, vol. 16, no. 2, pp. 123–141, 1993.

[337] A. Wasala, R. Weerasinghe, R. Pushpananda, C. Liyanage, and E. Jayalatharachchi, "A data-driven approach to checking and correcting spelling errors in sinhala," *Int. J. Adv. ICT Emerg. Reg*, vol. 3, no. 01, 2010.

[338] R. A. Wasala, R. Weerasinghe, R. Pushpananda, C. Liyanage, and E. Jayalatharachchi, "An open-source data driven spell checker for sinhala," *ICTer*, vol. 3, no. 1, 2011.

[339] E. Jayalatharachchi, A. Wasala, and R. Weerasinghe, "Data-driven spell checking: the synergy of two algorithms for spelling error detection and correction," in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*. IEEE, 2012, pp. 7–13.

[340] L. G. B. Subhagya, L. Ranathunga, W. H. A. Nimasha, B. R. Jayawickrama, and K. L. Mahaliyanaarchchi, "Data driven approach to sinhala spellchecker and correction," in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2018, pp. 01–06.

[341] U. Liyanapathirana, K. Gunasinghe, and G. Dias, "Sinspell: A comprehensive spelling checker for sinhala," *arXiv preprint arXiv:2107.02983*, 2021.

[342] L. Sithamparanathan and T. Uthayasanker, "A sinhala and tamil extension to generic environment for context-aware correction," in *2019 National Information Technology Conference (NITC)*. IEEE, 2019, pp. 102–106.

[343] H. Pabasara and S. Jayalal, "Computational model for detecting grammatical mistakes in sinhala text," in *9TH YSF SYMPOSIUM*, 2020, p. 255.

[344] ——, "Grammatical error detection and correction model for sinhala language sentences," in *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*. IEEE, 2020, pp. 17–24.

[345] S. Gunasekara, D. Chathura, C. Jeewantha, and G. Dias, "Using annotation projection for semantic role labeling of low-resourced language: Sinhala," 2020.

[346] B. Gamage, R. Pushpananda, and R. Weerasinghe, "The impact of using pre-trained word embeddings in sinhala chatbots," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 161–165.

[347] S. Harshani, "Sinhala chatbot for train information," Ph.D. dissertation, 2021.

[348] J. Chandrasena, A. Gunasekara, and G. Uwanthika, "Sinhala chatbot with recommendation system for sri lankan traditional dancers," 2021.

[349] U. Kumanayake, "A sinhala chatbot for user inquiries regarding degree programs at university of ruhuna," Ph.D. dissertation, 2021.

[350] M. Punchimudiyanse and R. G. N. Meegama, "Computer interpreter for translating written sinhala to sinhala sign," *OUSL Journal*, vol. 12, no. 1, pp. 70–90, 2017.

[351] ——, "Animation of fingerspelled words and number signs of the sinhala sign language," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 16, no. 4, p. 24, 2017.

[352] T. Dissanayake and B. Hettige, "Thematic relations based qa generator for sinhala," *13th International Research Conference General Sir John Kotelawala Defence University*, 2020.

[353] S. C. Fernando, "Inexact matching of proper names in sinhala," 2011.

[354] A. B. P. Kanduboda and K. Tamaoka, "Priority information in determining canonical word order of colloquial sinhalese sentences," in *Proceedings of the 139th Conference of the Linguistic Society of Japan*, vol. 1, 2009, pp. 32–37.

[355] ——, "Priority information for canonical word order of written sinhala sentences," in *Proceedings of the 140th Conference of the Linguistic Society of Japan*, 2010, pp. 358–363.

[356] K. Tamaoka, P. B. A. Kanduboda, and H. Sakai, "Effects of word order alternation on the sentence processing of sinhalese written and spoken forms," *Open Journal of Modern Linguistics*, vol. 1, no. 02, pp. 24–32, 2011.

[357] A. B. P. Kanduboda and K. Tamaoka, "Priority information determining the canonical word order of written sinhalese sentences," *Open Journal of Modern Linguistics*, vol. 2, no. 01, p. 26, 2012.

[358] J. Jayakody, T. Gamlath, W. Lasantha, K. Premachandra, A. Nugaliyadde, and Y. Mallawarachchi, ""mahoshadha", the sinhala tagged corpus based question answering system," in *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*. Springer, 2016, pp. 313–322.

[359] S. Basnayake, H. Wijekoon, and T. K. Wijayasiriwardhane, "Plagiarism detection in sinhala language: A software approach."

[360] L. Rajamanthri and S. Thelijjagoda, "Sinhala language plagiarism tool with internet resources using natural language processing."

[361] ——, "Plagiarism detection tool for sinhala language with internet resources using natural language processing," in *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*. IEEE, 2021, pp. 156–160.

[362] V. Liyanage and S. Ranathunga, "A multi-language platform for generating algebraic mathematical word problems," in *2019 14th Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2019, pp. 332–337.

[363] ——, "Multi-lingual mathematical word problem generation using long short term memory networks with enhanced input features," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 4709–4716.

[364] I. Smith and U. Thayasivam, “Language detection in sinhala-english code-mixed data,” in *2019 International Conference on Asian Language Processing (IALP)*. IEEE, 2019, pp. 228–233.
[365] ——, “Sinhala-english code-mixed data analysis: A review on data collection process,” in *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, vol. 250. IEEE, 2019, pp. 1–6.
[366] L. Sandathara, S. Tissera, R. Sathsarani, H. Hapuarachchi, and S. Thelijjagoda, “Arunalu: Learning ecosystem to overcome sinhala reading weakness due to dyslexia,” in *2020 2nd International Conference on Advancements in Computing (ICAC)*, vol. 1. IEEE, 2020, pp. 416–421.
[367] C. Rajitha, L. Piyarathne, D. Sachintha, and S. Ranathunga, “Metric learning in multilingual sentence similarity measurement for document alignment,” *arXiv preprint arXiv:2108.09495*, 2021.
[368] R. Kumari and S. Hettiarachchi, “Sintm-lda and rake based topic modelling for sinhala language,” in *2021 Asian Conference on Innovation in Technology (ASIANCON)*. IEEE, 2021, pp. 1–5.
[369] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
[370] S. Rose, D. Engel, N. Cramer, and W. Cowley, “Automatic Keyword Extraction from Individual Documents,” *Text mining: applications and theory*, vol. 1, pp. 1–20, 2010.
[371] A. Arambewela, S. Ahangama, and D. Dissanayake, “Real-time sinhala writing assistant for kids,” in *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, pp. 152–156.
[372] N. A. Malsha, K. D. Heshani, R. K. Ransara, D. A. Bandara, P. S. Kumari, and T. A. Kuruppu, “Automated sinhala news platform based on machine learning and deep learning,” in *2021 3rd International Conference on Advancements in Computing (ICAC)*. IEEE, 2021, pp. 134–139.
[373] A. Jayaweera, Y. Senanayake, and P. S. Haddela, “Dynamic stopword removal for sinhala language,” in *2019 National Information Technology Conference (NITC)*. IEEE, 2019, pp. 1–6.
[374] L. Jayasekara and S. Ahangama, “Trend detection in sinhala tweets using clustering and ranking algorithms,” in *2020 From Innovation to Impact (FITI)*, vol. 1. IEEE, 2020, pp. 1–6.
[375] U. Sandamini, K. Rathnakumara, P. Pramuditha, M. Dissanayake, D. Sriyaratna, H. De Silva, and D. Kasthurirathna, “A singlish supported post recommendation approach for social media,” 2022.
[376] J. B. Disanayake, *Basaka Mahima: 2 - Akuru ha Pili*. Godage & brothers, 2000.
[377] ——, *Basaka Mahima: 6 - Prakurthi*. Godage & brothers, 2004.
[378] ——, *Sinhala Reethiya: 7 - Pada Nirmanaya*. Sumitha Books, 2014.
[379] ——, *Basaka Mahima: 8 - Badditha*. Godage & brothers, 2000.
[380] ——, *Basaka Mahima: 10 - Nama Padaya*. Godage & brothers, 2008.
[381] ——, *Basaka Mahima: 11 - Kriya Padaya*. Godage & brothers, 2001.
[382] ——, *The structure of spoken Sinhala: Sounds and their patterns*. National Institute of Education, 1991.
[383] ——, *Sinhala Akshara Vicharaya (Sinhala Graphology)*. Sumitha Publishers, 2006.
[384] ——, *The Usage of Dental and Cerebral Nasals*. Sumitha Publishers, 2007.
[385] ——, *Bashavaka rata samudaya*. Lake house investment Co. Ltd. Colombo 2, 1969.
[386] ——, *Grammar of Contemporary Literary Sinhala-Introduction to Grammar, Structure of Spoken Sinhala*. Godage & Bros, 1995, vol. 661.
[387] D. A. Indrasena, *Sinhala Akshara Malava*, 2001.
[388] K. Jayathilake, *Modern Sinhalese lingustics*. Pradeepa Publications, 1991.
[389] D. K. Henadeerage, “Topics in sinhala syntax,” Ph.D. dissertation, The Australian National University, 2002.
[390] A. E. S. Dasanayaka, *kumara rachanaya; Grade 4*. M D Gunasena Publishers, Colombo, 1990.
[391] ——, *kumara rachanaya; Grade 5*. M D Gunasena Publishers, Colombo, 2005.
[392] S. O. Fernando, *Wara nonamena Nipatha*. Sammana, January 1994.
[393] ——, *Kriya pada igena ganimu*. Sammana, January 1994.
[394] ——, *Sinhala Nouns year 11*. Sammana, March 1994.
[395] E. Ranawake, *Spoken Sinhalese for Foreigners*. MD Gunasena & Co. Ltd., 1986.
[396] A. M. Gunasekara, *A comprehensive grammar of the Sinhalese language*. Asian Educational Services, 1999.
[397] S. Koparahewa, *Dictionary of Sinhala Spelling*. S. Godage and Brothers, Colombo, 2006.
[398] J. W. Gair and W. S. Karunatillake, *The Sinhala Writing System, A Guide to Transliteration*. Sinhamedia, PO Box 1027, Trumansburg, NY 14886, 2006.
[399] W. S. Karunatillake, *An Introduction to Spoken Sinhala*. M D Gunasena & Company, 1990.
[400] M. Inman, *Duration and Stress in Sinhala*. Stanford University, 1986.
[401] K. Munidasa, *Vyakarana Vivaranaya*. MD Gunasena Publishers, Colombo, 1938.
[402] National Institute of Education, *Sinhala Laekhana Reethiya*, 1989.
[403] K. Jayathilake, *Nuthana Sinhala Vyakaranaye Mul Potha*. Pradeepa Publications, 1991.
[404] U. S. Sannasgala and A. Perera, *Viyakarana Vimansawa*, 1995.
[405] V. G. Balagalle, *Basha Adauanayasaha Sinhala Vivaharaya*, 1995.
[406] W. S. Karunatilaka, *Sinhala Basha Viharanaya*. M D Gunasena Publishers, Colombo, 2004.
[407] S. Karunarathna, *Sinahala Viharanaya*. Washana prakasakayo, Dankotuwa, Sri Lanka, 2004.
[408] P. Alwis, *Niwaeradi Wahara*. Suriya Prakashakayo, 2006, vol. 109.
[409] ——, *Niwaeradi Wahara -2*. Suriya Prakashakayo, 2007.
[410] K. C. Perera, *Prayogika Sinhla Viyakaranaya*.
[411] K. Munidasa, *Kriya Vivaranaya*. MD Gunasena Publishers, Colombo, 1993.
[412] T. Jayawardana, *The surface case system in Sinhala*. University of Kelaniya, 1989.
[413] D. Rajapaksha, *Sinhala bhashave pada bedima saha virama lakshana bhavithaya*, 2008.
[414] S. M. Kariyakarawana, *The syntax of focus and wh-questions in Sinhala*. Karunaratne & Sons Limited, 1998.
[415] S. L. Kekulawala, *The future tense in Sinhalese – an ‘unorthodox’ point of view*. Vidyalankara University of Ceylon, 1972.

Appendix A
Co-author Graph

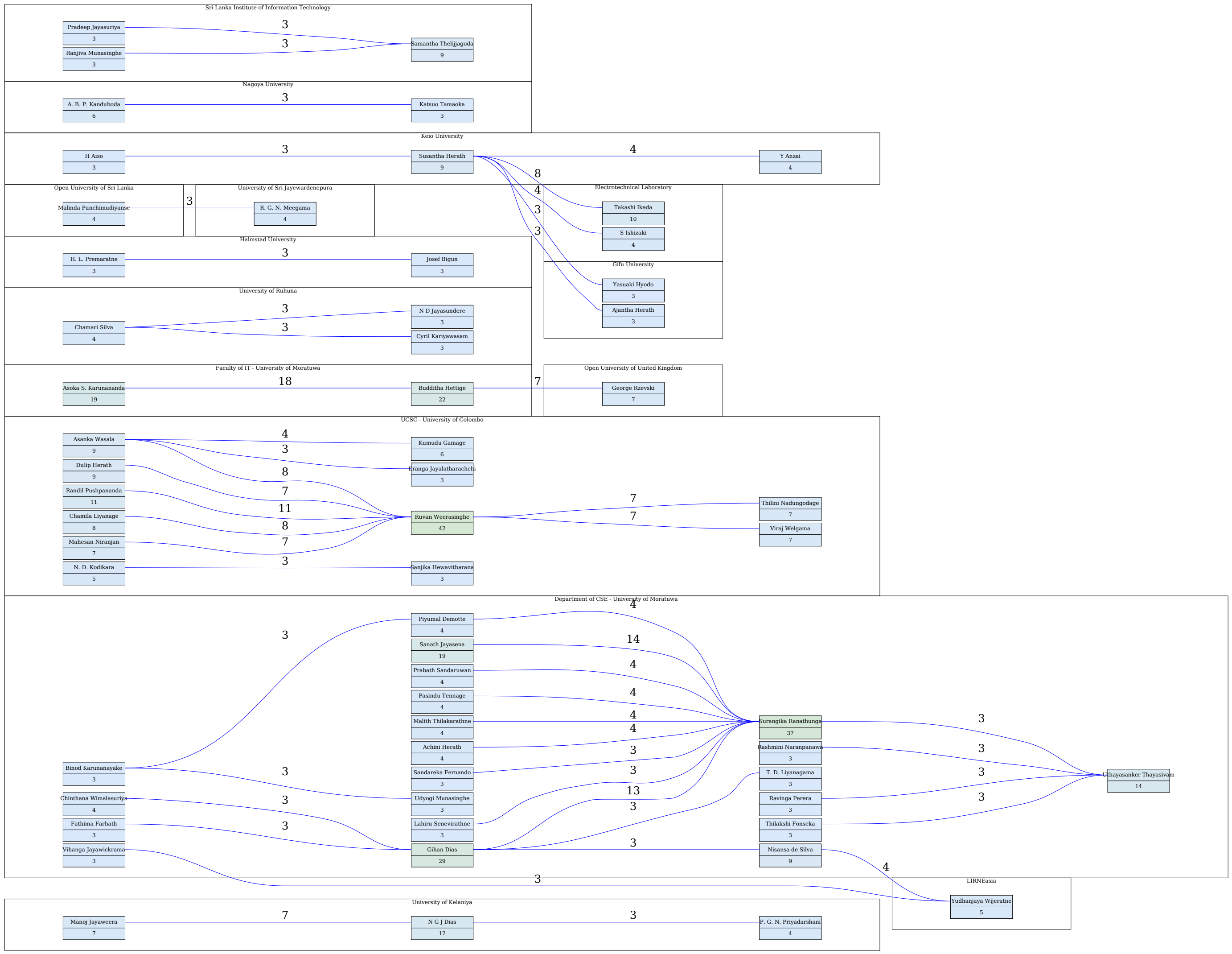


Fig. 3: Co-author graph of the most prolific researchers in the the Sinhala NLP domain (Selected at the threshold of at least 3 coauthored publications)