Building a Linguistic Resource: A Word Frequency List for Sinhala

Aloka Fernando
University of Moratuwa
Sri Lanka
alokaf@uom.lk

Gihan Dias
University of Moratuwa
Sri Lanka
gihan@uom.lk

Abstract

A word frequency list is a list of unique words in a language along with their frequency count. It is generally sorted by frequency. Such a list is essential for many NLP tasks, including building language models, POS taggers, spelling checkers, word separation guides, etc., in addition to assisting language learners. Such lists are available for many languages, but a large-scale word list is still not available for Sinhala. We have developed a comprehensive list of words, together with their frequency and part-of-speech (POS), from a large textbase. Unlike many other such lists, our list includes a large number of low-frequency words (many of which are erroneous), which enables the analysis of such words, including the frequencies of errors. In addition to the main list, we have also prepared a list of linguistically verified words. The word frequency list and the verified word list are the largest collections of words lists that are available for the Sinhala language.

1 Introduction

Word frequency lists are useful for analysing the vocabulary of a language (Nation and Waring, 1997). They generally comprise a list of lexical words sorted by frequency of occurrence in a corpus, as a ranked list. Such lists can be used to build a directory, for language learning and for teaching. They are also useful in downstream applications such as part-of-speech (POS) tagging, syntactic parsing, machine translation, speech processing etc.

In this paper, first, we present a word frequency list for Sinhala. We compiled a list of 2 million unique words by considering published documents from multiple sources, representing diverse domains. The raw corpus contained 127 million word tokens. In addition to frequency, the list provides the POS Tag of each word. Second, we have compiled a linguistically verified word list of over 280 thousand unique words.

These two word lists are currently the largest available word lists for the Sinhala language. Therefore, they would be very useful linguistic resources for many downstream applications.

Further, both the word frequency list and verified word list are publicly released [1], for the progression of future research.

1.1 Objectives

In our work, we often encountered a word, and needed to know if it is frequently used, a rare word, or a misspelling. Therefore, we decided to take a large dataset including common-crawl data and other available data, and construct a list of words seen in the wild. Rather than a curated corpus, we wanted to use the largest dataset available. Instead of limiting ourselves to correctly spelled and separated words, we decided to build a list which includes incorrect words. As we often needed to know the POS of each word, we also tagged each word with a POS, even though the tagging may sometimes be inaccurate.

We also needed lists of correct words, therefore we decided to compile a list of linguistically verified words.

In this paper, Section 2 covers related work and Section 3 describes how the lists were compiled. Our results are analysed in Section 4, and Section 5 gives our conclusions and planned continuations.

2 Related Work

Word frequency lists have long been in use for languages such as Spanish, French, German and English. Before the advent of computing, they were generated manually, but have become much easier

[1] https://github.com/nlpcuom/Word-Frequency-List-for-Sinhala

Proceedings of the 18th International Conference on Natural Language Processing, pages 606–610
Silchar, India. December 16 - 19, 2021.

to generate using a computer (Davies and Davies, 2017; Tschirner et al., 2019; Lonsdale and Le Bras, 2009; Leech et al., 2014).

The main objectives of many word frequency lists based on written or spoken corpora (Dang et al., 2017) were for language teaching and learning.

However, word lists are valuable resources for computational linguistics as well. They provide annotated datasets for downstream applications such as POS taggers (Kucera and Francis, 1967), dependency parsers (Silveira et al., 2014) etc.

More recently, frequency lists were generated for non-Latin languages such as Tamil (Kumar, 2019) and Russian (Sherstinova et al., 2020). Frequency lists were also generated for specific domains such as literary genres and historical periods, and even individual books and authors.

For Sinhala, a corpus based lexicon were done by Weerasinghe et al. (2009). This contains 35K unique words. They have constructed a word frequency list extracted from publicly available Sinhala text documents from multiple domains. However, the list does not appear to be publicly accessible.

More recently, Google has released a Sinhala lexicon with about 42,000 entries (Jansche, 2017).

Text documents were crawled from the web at scale (Schwenk et al., 2019; Wenzek et al., 2020) for the progress of research in diverse domains. Such crawled corpora are also available for Sinhala. To the best of our knowledge, the generated word frequency lists had not considered these recently web crawled data.

3 Methodology

3.1 Dataset for the Word Frequency List

As our dataset, we used publicly available Sinhala text crawled from the web by the Common Crawl project (Wenzek et al., 2020). We also included government documents (Fernando et al., 2020) and news sites (Isuranga et al., 2020). These documents contain text corresponding to modern usage of the language. Hence the word list would be suitable for many downstream applications.

The corpus statistics of the data is in Table 1.

Domain	Total Sents.	Total Tokens
Govt. Documents	77,694	0.75M
News	332,793	20M
Common crawl	5,000,324	106M

Table 1: Corpus Statistics

3.2 Data Cleaning and Pre-processing

The government documents dataset had been compiled manually and the text was of good quality. Since the news data and common crawl data had been crawled from the web, they needed to be cleaned before use.

The data were tokenised using the Sinhala tokenizer (Farhath et al., 2018), to separate punctuation and the words. As compound nouns, verbs or particles used as suffixes, may be written with or without white space, such words were not combined or split, but considered them as they appeared in the corpus.

The first cleaning step was to filter out words with non-Sinhala characters, including English and other foreign language words and numerals. Scripts were developed using rules and regular expressions to remove such invalid words.

The data contained Unicode errors such as duplicated modifiers, misplaced modifiers, separation of modifiers into parts, etc. These were corrected with the Unicode Error Corrector[2], a rule-based tool developed for Sinhala Unicode error correction. Some corrections were:

අාකර්ෂනය → ආකර්ෂනය

ොඩාලර් → ඩොලර්

ේ + ා → ෝ

නිරීක්ෂණය → නිරීක්ෂණය

The Sinhala yansaya and rakaransaya symbols are formed using the Unicode zero-width joiner (ZWJ) character. Use of these symbols is mandatory in Sinhala. However, some systems erroneously delete ZWJ characters, giving incorrect words, e.g. ක්‍රිකට් (cricket) is erroneously depicted as ක්රිකට්. Others divide a word into two, e.g., ක්‍රිකට් → ක් රිකට් by replacing the ZWJ with a space.

These errors were corrected using our Zero-Width Joiner Fix [3], which uses lists of valid words and sub-words to identify where the ZWJ has been deleted or replaced, and re-inserts the character.

As there remained further invalid words in the list, they were removed based on a POS filtering. A Sinhala POS Tagger (Fernando and Ranathunga, 2018) was used to tag the words. We removed

[2]https://nlp-tools.uom.lk/uniec/

[3]https://nlp-tools.uom.lk/zwjfix/

the words tagged as Full Stop (FS), Punctuation (PUNC), Foreign Word (FRW) and Unknown (UNK). Subsequently, some of the obvious erroneous words were removed manually.

Sinhala contains many single character words and particles, such as ඒ (that) ද (and) ලී (wood). We included such words in our list as well.

3.3 Word Frequency list

After cleaning the final corpus statistics are shown in Table 2.

	Total Words
Tokens in original documents	127M
Total word count after cleaning	122,998,105
Total unique words	2,170,052

Table 2: Corpus statistics in-terms of token counts

We extracted the lemma of each word using SinMorphy (Kumarasinghe et al., 2021), a morphological parser for Sinhala. The top 20 frequent words are listed along with the lemma, POS and frequency in Table 3.

Words	**Lemma**	**Frequency**	**POS Tag**
මේ	මේ	1067105	DET
ඒ	ඒ	946049	ABB
ඇති	ඇති	567890	NIP
සහ	සහ	518043	CC
හා	හා	511663	CC
එක	එක	491217	NUM
මම	ම	485869	PRP
බව	බව	469031	POST
ද	ද	453787	RP
නම්	නම	438329	POST
කර	කර	423842	VNF
වන	වන	387674	VP
කරන	කර	376601	VP
අතර	අතර	353691	POST
මට	ම	342623	PRP
ගැන	ගැන	339464	POST
මෙම	මෙම	326924	DET
නෑ	නෑ	323017	NIP
නිසා	නිසා	306021	POST
වූ	ව	298454	VP

Table 3: Top 20 Frequent words

The most frequent words are mainly determiners, particles, pronouns, etc. However, we see that the POS tagging is sometimes not accurate.

3.4 Verified Words List

The words in the Word Frequency list were run through a spelling checker (Liyanapathirana et al., 2021), and the words accepted by it were taken as correct. As it is infeasible to manually check all the words which failed the spelling check, it was decided to manually check the 3555 highest frequency words which failed the spelling check. Of these, 1836 were manually verified to be correct, and added to the verified words list. This list comprises of 280,603 words. This is the largest list of verified Sinhala words available.

4 Analysis

4.1 Analysis of the Word Frequency List

When the distribution of words was analysed, we observed that 50% of the words in the corpus are covered by 17% of the words in the word frequency list. This means 17% words can be identified as the most commonly used words in Sinhala language.

Subsequently we analysed the word counts based on the POS Tag. The outcome is shown in Table 4.

Gram. Category	**POS Categories**	**Total Words**	**Per. %**
Noun	NNC,NNP, PRP,NNJ, VNN,NNP	67.8M	55.1
Verb	VNF,VP,VFM	18.7M	15.2
Adjective	JJ	9.0M	7.4
Adverb	RB	1.4M	1.2
Other		26.1M	21.2

Table 4: Word counts based on POS Tag.

Words tagged as Common Noun (NNC), Proper Noun(NNP), Pronoun(PRP), Adjectival Noun(NNJ) and Verbal Noun(VNN) were grouped into the Noun grammatical category. Those tagged as Verb Non Finite(VNF), Verb Finite(VFM) and Verb Participle(VP) were considered as Verbs.

From the statistics in Table 4, we see that the majority of words in the list are nouns. Sinhala is a morphologically rich language, which means the words are inflected based on the gender, number, case, etc. As nouns have more morphological variants, they account for more entries in the list.This emphasises the importance for linguistic tool sup-

port for nouns in morphologically rich languages such as Sinhala.

On the other hand number of verbs, adjectives and adverbs in Sinhala is limited. Therefore the representation of such words in the list is lower.

When we further analysed the words under each POS category, we came across some issues.

Some words were incorrectly classified by the POS tagger, eg: අකකර (acres) is a noun but was incorrectly tagged as a non-finite verb (VNF) . We plan to implement a better POS tagger at a later date.

Further, some words contained spelling and word separation issues, e.g. නිලධාරී → නිලධාරි (official), අංකගණිතය → අංක ගණිතය (arithmetic). These may or may not be considered errors depending on the point of view of the user.

4.2 Analysis of the Verified Word List

Each word in the linguistically verified word list was parsed by the morphological parser (Kumarasinghe et al., 2021) and the most frequent lemmas were analysed. The morphological parser is rule-based, and covers both morphological rules as well as sandhi-rules. Further these rules have been linguistically verified. Therefore we believe the morphological parser can provide reliable morphological information for our word list.

Of the 280,603 unique words in the linguistically verified word list, we obtained morphological information for 256,083 words, which is a coverage of 91%. A total of 43,313 unique lemmas were found. The information corresponding to the top most 10 frequent lemmas can be found in Table 5. The Total Words, in Table 5 corresponds to the total number of words from the word frequency list, with the lemma.

It was an interesting observation that 50% of total words in the word list were covered by only 10% of unique lemmas.

Lemma	Total Words ('000)	Sample Words
කර	1,674	කර, කරන, කරමින්...
ම	1,124	මම, මට, මා, මමත්, මටත්...
ඒ	1,072	ඒ, ඒත්, ඒයි, ඒය, ඒවල...
මේ	1,067	මේ, මේට, මේගේ, මේටත්...
ව	902	වූ, වන්නේ, වෙයි, නොවන...
අප	865	අපි, අපේ, අප, අපිට...
ය	738	ය, ගිය, ගිහින්, ගියා, යයි...
එක	723	එක, එකේ, එකෙන්, එකත්...
බව	649	බව, බවයි, බැවිනි, බැවිණි...
නම	625	නම්, නමුත්, නම, නමක්...

Table 5: Top 10 Most Frequent Lemmas

5 Conclusion and Future work

The word frequency list, comprising over 2 million words, is by far the largest word list for the Sinhala language. It is also the only one supplemented with POS information. Although the list contains many incorrect words and incorrect word separations, this is a feature, not a drawback, as it allows us to analyse the frequencies of variant spellings and compound words.

The verified word list of over 280 thousand words is also the largest such list for Sinhala. Although many words and word inflections are not in this list, it does cover most of the words in common use.

As future work, we plan to compile a frequency lists of morphosyntactic suffixes for Sinhala. We also plan to compile lists of word bi-grams, tri-grams, etc.

We also plan to compile domain-specific word lists for government documents, news, textbooks, web, etc.

These lists have been used to develop several other tools, including a spelling checker and a sentence generator, and are being used to identify spelling error patterns, etc. It will be a useful tool in many other areas of NLP.

6 Acknowledgments

This research was supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Education funded by the World Bank.

References

Thi Ngoc Yen Dang, Averil Coxhead, and Stuart Webb. 2017. The academic spoken word list. Language Learning, 67(4):959--997.

Mark Davies and Kathy Hayward Davies. 2017. A frequency dictionary of Spanish: Core vocabulary for learners. Routledge.

Fathima Farhath, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. 2018. Integration of bilingual lists for domain-specific statistical machine translation for sinhala-tamil. In 2018 Moratuwa

Engineering Research Conference (MERCon), pages 538--543. IEEE.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. arXiv preprint arXiv:2011.02821.

Sandareka Fernando and Surangika Ranathunga. 2018. Evaluation of different classifiers for sinhala pos tagging. In 2018 Moratuwa Engineering Research Conference (MERCon), pages 96--101. IEEE.

Udhan Isuranga, Janaka Sandaruwan, Udesh Athukorala, and Gihan Dias. 2020. Improved cross-lingual document similarity measurement. In 2020 International Conference on Asian Language Processing (IALP), pages 45--49. IEEE.

Martin Jansche. 2017. A pronunciation dictionary for sinhala.

H Kucera and W Nelson Francis. 1967. The brown university standard corpus of present day american english.

LR Prem Kumar. 2019. Word frequency in language teaching--a case study of tamil textbooks of tamilnadu. Language in India.

Kalindu Kumarasinghe, Gihan Dias, and Indu Herath. 2021. Sinmorphy: A morphological analyzer for the sinhala language. In 2021 Moratuwa Engineering Research Conference (MERCon), pages 681--686. IEEE.

Geoffrey Leech, Paul Rayson, et al. 2014. Word frequencies in written and spoken English: Based on the British National Corpus. Routledge.

Upuli Liyanapathirana, Kaumini Gunasinghe, and Gihan Dias. 2021. Sinspell: A comprehensive spelling checker for sinhala.

Deryle Lonsdale and Yvon Le Bras. 2009. A frequency dictionary of French: Core vocabulary for learners. Routledge.

Paul Nation and Robert Waring. 1997. Vocabulary size, text coverage and word lists. Vocabulary: Description, acquisition and pedagogy, 14:6--19.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. arXiv preprint arXiv:1911.04944.

Tatiana Sherstinova, Alexander Grebennikov, Tatiana Skrebtsova, Anna Guseva, Mary Gukasian, Irina Egoshina, and Maria Turygina. 2020. Frequency word lists and their variability (the case of russian fiction in 1900-1930). In Conference of Open Innovations Association, FRUCT, 27, pages 366--373. FRUCT Oy.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A gold standard dependency corpus for english. In LREC, pages 2897--2904. Citeseer.

Erwin Tschirner, Jupp Möhring, and Elisabeth Muntschick. 2019. A frequency dictionary of German: Core vocabulary for learners. Routledge.

Ruvan Weerasinghe, Dulip Herath, and Viraj Welgama. 2009. Corpus-based sinhala lexicon. In Proceedings of the 7th Workshop on Asian Language Resources (ALR7), pages 17--23.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4003--4012.