SiDiaC: Sinhala Diachronic Corpus

Nevidu Jayatilleke, Nisansa de Silva

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

SiDiaC: Sinhala Diachronic Corpus

Nevidu Jayatilleke and Nisansa de Silva
Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{nevidu.25, NisansaDdS}@cse.mrt.ac.lk

Abstract

SiDiaC, the first comprehensive *Sinhala Diachronic Corpus*, covers a historical span from the 5th to the 20th century CE. SiDiaC comprises 58k words across 46 literary works, annotated carefully based on the written date, after filtering based on availability, authorship, copyright compliance, and data attribution. Texts from the *National Library of Sri Lanka* were digitised using Google Document AI OCR engine, followed by post-processing to correct formatting and modernise the orthography. The construction of SiDiaC was informed by practices from other corpora, such as *FarPaHC*, particularly in syntactic annotation and text normalisation strategies, due to the shared characteristics of low-resourced language status. This corpus is categorised based on genres into two layers: primary and secondary. Primary categorisation is binary, classifying each book into Non-Fiction or Fiction, while the secondary categorisation is more specific, grouping texts under Religious, History, Poetry, Language, and Medical genres. Despite challenges including limited access to rare texts and reliance on secondary date sources, SiDiaC serves as a foundational resource for Sinhala NLP, significantly extending the resources available for Sinhala, enabling diachronic studies in lexical change, neologism tracking, historical syntax, and corpus-based lexicography.

1 Introduction

Sola lingua bona est lingua mortua[1]; given that all languages that are in use, evolve with a gradual process of linguistic change over time, proposed to have originated from an initially gestural communication system (Corballis, 2017). Factors affecting this complex evolution include cultural influences, which drive irregular word meaning shifts through phenomena such as new technologies (e.g., *cell* to mean *cell phone* in addition to *prison cell*) or community-specific vernaculars (e.g., *gay* to mean *homosexual* in addition to *carefree*). These cultural shifts often impact nouns more significantly. Conversely, regular linguistic processes, such as subjectification (e.g., *actually* shifting from objective to subjective usage) or grammaticalisation (e.g., *promise* undergoing rich changes), cause more predictable semantic changes and tend to affect verbs, adjectives, and adverbs more readily (Hamilton et al., 2016).

The Sinhala language is an Indo-European language, which possesses a rich and diverse literary heritage that has developed over the course of several millennia, with its origins tracing back to between the 3rd and 2nd centuries BCE (de Silva, 2025). This language has undergone significant evolution and transformation throughout its history, resulting in the form of modern Sinhala that we engage with today. Sinhala is spoken as L1 by approximately 16 million people, primarily located on the island of Sri Lanka (de Silva, 2025). The Sinhala script, which is unique to the language, descends from the Indian Brahmi script (Fernando, 1949; De Mel et al., 2025). Sinhala is classified as a lower-resourced language (Category 02) according to the criteria presented by Ranathunga and de Silva (2022).

In this study, we introduce a novel diachronic Sinhala dataset, SiDiaC[2], which covers the period from 426 CE to 1944 CE. This dataset is based on distinct identifications of written years or specific time frames of the recognised Sinhala literature.

2 Existing Work

The development of historical corpora has garnered significant attention due to its importance beyond the creation of general-purpose corpora. It enables researchers to investigate the evolution of language, taking into account changes in semantics, lexicon,

[1] **Latin:** The only good language is a dead language.

[2] https://github.com/NeviduJ/SiDiaC

morphology, and syntax. As a result, studies have been conducted to develop diachronic corpora for different languages.

2.1 LatinlSE

McGillivray and Kilgarriff (2013) introduced LatinlSE, a 13-million-word historical Latin corpus developed for the Sketch Engine[3], a leading corpus query tool. Covering an extensive 22-century period from the 2nd Century BCE to the 21st Century CE, LatinlSE is equipped with detailed metadata, including author, title, genre, era, date, and century.

The methodology for creating this corpus involved gathering texts from various online digital libraries, such as *LacusCurtius*[4], *IntraText*[5], and *Musisque Deoque*[6]. This process ensured a broad classification of genres as prose and poetry, and the texts were converted into a verticalized format while preserving their metadata. A significant aspect of the creation process was the automatic linguistic annotation using advanced NLP tools. This included lemmatisation with the PROIEL[7] project's morphological analyser, complemented by *Quick Latin*[8] for unrecognised forms. Part-of-Speech (POS) tagging was achieved by training TreeTagger (Schmid, 1999) on existing Latin treebanks, including the *Index Thomisticus Treebank*[9], the *Latin Dependency Treebank* (Bamman and Crane, 2006), and the PROIEL project's Latin treebank (Haug and Jøhndal, 2008). This training helps disambiguate analyses and assign the most likely lemma and POS to each token in context. This comprehensive dataset allows users to perform sophisticated searches based on lemmas, POS, and context, facilitating the study of shifts in word meanings over time.

2.2 IcePaHC and FarPaHC

The *Icelandic Parsed Historical Corpus* (IcePaHC) (Rögnvaldsson et al., 2012) is a one-million-word parsed historical corpus of Icelandic, spanning from the late 12th century to the early 21st century. But more relevant to our work in this study is the *Faroese Parsed Historical Corpus* (FarPaHC), a syntactically annotated corpus of Faroese historical texts, that is presented as a *spin-off* of IcePaHC. The reason for this relevance is that, according to Ranathunga and de Silva (2022), Faroese also belongs to Category 02, similar to Sinhala. The FarPaHC corpus has 53,000 words.

It's given that FarPaHC is an extension of IcePaHC; the primary sources included narrative and religious texts that have parallel texts in IcePaHC. A key step in the process was the conversion of all texts to modern spelling using the IceNLP package[10] (which includes a tokeniser, POS tagger, and lemmatiser), which was necessary for preprocessing and for facilitating searches. The annotation process involved manually dividing clauses, semi-automatically preprocessing texts with IceNLP and CorpusSearch[11] for partial annotations, and extensive manual parsing carried out by one annotator using a custom-developed visual tree editor, Annotald[12].

2.3 Other Historical Corpora

Pettersson and Borin (2019) provides a comprehensive survey on existing diachronic and historical corpora. The work by Keersmaekers and Van Hal (2024) presents a case study demonstrating how large-scale automated parsing of Greek papyri can create richly annotated diachronic resources. Chen and Liu (2025) have created a Chinese corpus from the last 30 years of news articles on land usage. Even the corpora in higher-resourced languages such as DIAKORP (Kučera et al., 2015) (Czech), ARCHER (Biber et al., 1994), COHA (Davies, 2012)[13] (English), DTA (Geyken et al., 2011) and GerManC (Scheible et al., 2011) (German) differ in size, balance, annotation depth, and access models. DIAKORP offers seven centuries of Czech texts, though it lacks linguistic annotation. ARCHER samples English registers across four centuries in 50-year intervals, while COHA spans two centuries of American English with lemmatisation and POS tagging. *Penn Parsed Corpora of Historical English* (Taylor and Kroch, 1994) (PPCHE) and SRCMF (Stein and Prévost, 2013) (Old French) are similar to FarPaHC in the sense that they, too, are manually annotated corpora which provide syntactic analyses suitable for structural studies.

[3] https://www.sketchengine.eu/
[4] https://penelope.uchicago.edu/Thayer/E/Roman/Texts/
[5] https://www.intratext.com/
[6] https://www.mqdq.it/
[7] https://www.hf.uio.no/ifikk/english/research/projects/proiel/
[8] https://www.quicklatin.com/
[9] https://itreebank.marginalia.it/
[10] https://sourceforge.net/projects/icenlp/
[11] https://corpussearch.sourceforge.net/
[12] https://github.com/Annotald/annotald
[13] https://www.english-corpora.org/coha/

PPCHE, in particular, has influenced corpora in other languages through its *Penn-Helsinki* annotation scheme, facilitating cross-linguistic comparison. Similarly, the PROIEL treebank family (Eckhoff et al., 2018) extends such comparisons to some Indo-European languages via aligned New Testament translations.

ReM (Klein and Dipper, 2016) (Middle High German), RIDGES (Odebrecht et al., 2017) (German-Science), and the *Swedish Culturomics Gigaword* corpus (Eide et al., 2016), offer layered annotation or harmonised spellings for OCR quality control. While all corpora use some metadata scheme to provide critical contextual information such as date, genre, region, and authorship, beyond that, the metadata coverage varies widely. However, it can be noted that TEI-based[14] metadata schemes are popular among European language corpora.

3 Methodology

In this section, we describe the methodology used to create this dataset from the ground up. The process involved careful attention to detail at every stage, from planning to the final presentation, ensuring that the data are valid and of high quality. The procedure included addressing copyright laws in Sri Lanka, acquiring data, extracting text, and performing post-processing and formatting of the data as shown in Figure 1.

3.1 Dataset Assembly

At first, we began acquiring Sinhala literature, including both fiction and non-fiction books, from the *Internet Archives*. However, the amount of data we were able to gather was quite limited. As a result, we decided to turn to the primary institution dedicated to Sinhala literature: the National Library (Natlib) of Sri Lanka[15], which has its own digital repository[16].

In the digital repository, we were able to organise all available content chronologically by issue date, allowing us to see publications printed dating back to 1800 CE. We carefully selected the book title, author name, identifier number, and collection name for each book from that point onward. This process required careful filtering, as most of the available content consisted of gazettes and police reports.

[14] Text Encoding Initiative (TEI) guidelines
[15] https://www.natlib.lk/
[16] https://diglib.natlib.lk/

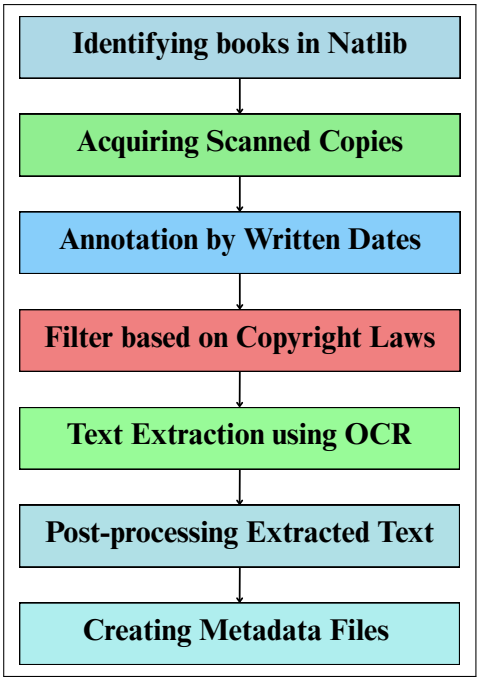


Figure 1: Summary of the Methodology Used in the Creation of SiDiaC.

We identified 233 unique books printed between 1800 CE and 1955 CE in the Natlib digital repository (this is based on the issued date, not the written date). Of these, only 12 books were available for open access; we had to request access to the remainder. Our initial plan was to collect 100 sentences per year. To achieve this, we estimated that obtaining five pages of text from each book, excluding content pages and the preface, would provide us with more than 100 sentences, which amounts to approximately 1500 to 2000 word tokens, assuming there are about 15 to 20 words per sentence. Therefore, for the closed-access books, we requested five pages from each one. The process of obtaining access to these books was very difficult because most of them were part of the Rare Books Collection.

3.2 Annotation by Written Date

The issue date of the identified books was clearly stated in the digital repository at Natlib. However, this does not imply that the books were actually written during those specified dates. In fact, a book could have been written centuries earlier, while the printed version was released much later.

Document dating has become extensively recognised in computational sociology and studies within digital humanities (Ren et al., 2023; Baledent et al., 2020; Hellwig, 2020). When compared to other dating tasks, historical text dating is more

complex due to the absence of explicit temporal indicators (such as time expressions) that aid in determining the date a document was written (Toner and Han, 2019; Baledent et al., 2020; Hellwig, 2020). It is clear that text dating, or the process of annotating the written date of a document, is an important task in diachronic studies (Ansari et al., 2023; Ren et al., 2023; Favaro et al., 2022).

Therefore, a comprehensive analysis was conducted to ensure that the written year of each book was accurately represented, ensuring that the resulting SiDiaC dataset accurately represents a proper diachronic corpus.

Upon the recommendation of experts in Sinhala linguistics, we identified a comprehensive book on Sinhala literature that claims to encompass literature information from its inception until 1994 CE (Sannasgala, 2015). This text served as the primary reference for the establishment of the respective date anchors, employing both time periods and specific years as outlined. The date ranges identified in Sannasgala (2015) correspond either to the period during which the book was authored or to the time period in which the author lived.

The process of determining the written dates for the books became more complicated because some books in the dataset include commentaries and discourses on earlier works. In this version of the dataset, these cases are tied to the original earlier book's written date, as they contain both the information from the original book and its corresponding commentary (often written centuries prior, with extensive sections given as direct quotes without paraphrasing).

3.3 Challenges from Copyright Laws

During the planning stage, one of the biggest challenges we faced was managing copyright issues. To address this, we conducted a thorough analysis of copyright laws in Sri Lanka, which are governed by the Intellectual Property Act No. 36 of 2003[17].

According to this act, copyright in Sri Lanka is generally protected for the life of the author, plus an additional 70 years after their death. In cases where the author is unknown, copyright protection lasts for 70 years from the date of first publication. As a result, we focused on literature where the author passed away before 1955, as well as works by unknown authors that were published before 1955.

3.4 Data Filtration

We initially identified 233 unique books, but after careful consideration of several factors, we ultimately selected only 46. Our selection process was influenced by the availability of scanned copies, the written dates of the works, and compliance with copyright laws as illustrated in Figure 2.

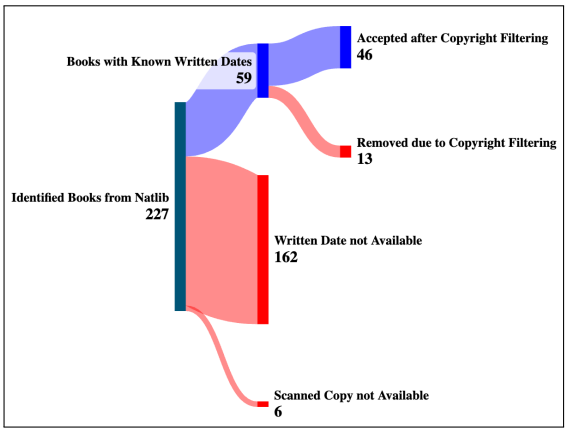


Figure 2: Sequential Data Filtration Procedure

Our first limitation was the availability of scanned copies at the digital repository of Natlib, which restricted us to 221 books. Additionally, we were only able to determine the written dates or periods for 65 of the 233 books. Taking both the availability of scanned copies and the accessibility of written dates into account, the number of selected records was reduced to 59. We then further refined this selection following the copyright law based filtering process we discussed, resulting in a final total of 46 books as presented in Table 3 in Appendix A.

3.5 Text Extraction using OCR

As the digital repository of Natlib shared the scanned copies of the requested books, we were forced to extract text information from the documents. Therefore, we selected an Optical Character Recognition (OCR) engine to get this task done.

In the comparative study conducted by Jayatilleke and de Silva (2025), five different OCR engines were analysed for text extraction using a synthetically created image-text dataset for Sinhala. Based on this study, we identified two standout OCR engines: Google Document AI[18] and

[17] https://www.gov.lk/wordpress/wp-content/uploads/2015/03/IntellectualPropertyActNo.36of2003Sectionsr.pdf

[18] https://cloud.google.com/document-ai/

Surya[19], with Surya being reported to outperform all other systems compared. However, during our text extraction process, we found that, under realistic conditions (unlike the synthetic conditions used in the study), Google Document AI provided more accurate results as shown in Appendix B.

Document AI is a service provided by Google Cloud Platform (GCP)[20]. In this platform, we created a processor and utilised its Application Programming Interface (API) key to conduct OCR. The processor can handle a maximum of 15 pages at a time; however, this was not an issue since all of our scanned copies contained 5 to 8 pages, as shown in Figure 3. Throughout the procedure, we ensured that we obtained the model confidence for every page of each processed document. We then calculated the average confidence score, which is included in the metadata file of each book folder.

This OCR processor has demonstrated that it can perform text recognition that goes beyond simple extraction. It adapts effectively and generates words in modern Sinhala spelling while also taking into account morphology, where morphemes are formed accordingly, as explained in the section 4.1.

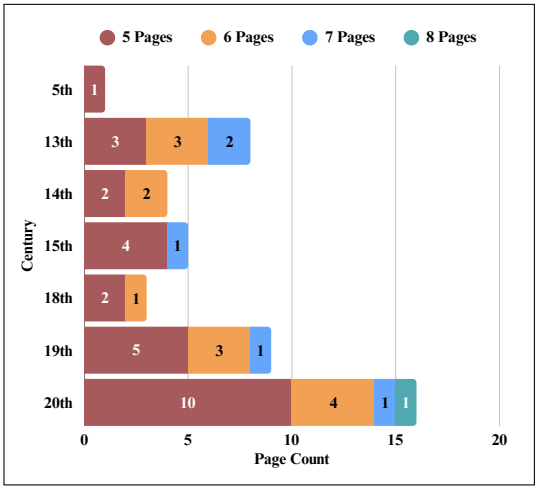


Figure 3: Distribution of Page Counts of Scanned Copies

3.6 Post-Processing Extracted Text

Although the OCR accuracy averaged 96.84% across all documents, the formatting issues were significant enough to require manual adjustments. The Document AI's advanced performance helped to streamline manual post-processing tasks, significantly reducing the time required for that work.

[19] https://github.com/VikParuchuri/surya
[20] https://cloud.google.com/

The post-processing includes correcting the following text formatting issues:

- **Spacing Errors**: This involves fixing incorrect or inconsistent spacing between words and sentences, such as missing spaces, uneven spaces and extra spaces.
- **Multi-column Text**: This refers to texts containing errors where two columns were treated as a single column, resulting in entire horizontal lines being extracted without properly traversing each column separately.
- **Misplaced Words/Phrases**: This addresses instances where words or phrases are out of order, leading to illogical text.
- **Paragraph and Line Indentation**: This involves standardising the indentation of paragraphs and individual lines. This could mean adding consistent indents to new paragraphs (e.g., poem blocks) and removing incorrect indents.
- **Removal of Seal Context**: This involves identifying and eliminating specific phrases or watermarks that represent seals or official stamps.
- **Page Number Removal**: This focuses on identifying and deleting page numbers that appear within the body of the text, as they are part of the document's structure rather than its content.

While language understanding was not a strict requirement for addressing these formatting-related factors, all manual post-processing procedures were carried out by the authors using a human-in-loop strategy (Lamba and Madhusudhan, 2023). This approach involved correcting formatting errors within a single window that contained both page scans and editable transcripts (Christy et al., 2017). The authors responsible for these corrections are native Sinhala speakers. The post-processing steps applied, along with examples, are further discussed and illustrated in Appendix C.

An important finding of this study was the presence of Pali, Sanskrit, and minimal English in certain records of SiDiaC. The inclusion of Pali and Sanskrit can be attributed to the fact that most historical texts are related to religion. Notably, both Pali and Sanskrit are written in the Sinhala script in all cases. In this study, we chose not to remove the content in these languages to avoid losing context.

3.7 Creation of Metadata Files

The dataset consisted of folders, each dedicated to a specific book. Within each folder, there is a text file along with a metadata file. The metadata files contain information such as the title and author names in both Sinhala and romanised forms, as well as the genre, issue date, written date, and the OCR confidence level for each particular book. Most of these information fields were identified through the LatinlSE corpus (McGillivray and Kilgarriff, 2013). Following the conventions of Davies (2012) and Rögnvaldsson et al. (2012), we maintain a consistent metadata annotation method throughout the corpus without changing it across the centuries. The overall composition of the SiDiaC corpus, including folder and file level examples, is illustrated in Figure 4.

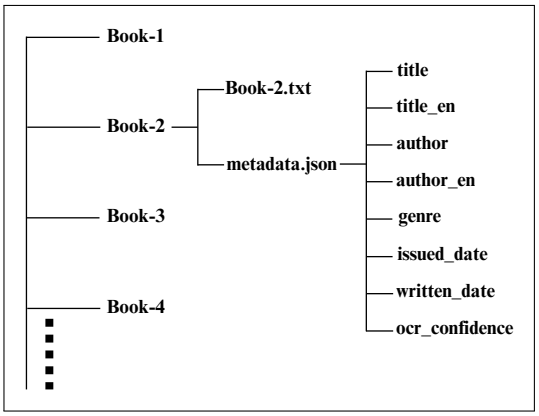


Figure 4: Composition of the SiDiaC Corpus

The title of each book was consistently provided. When known, the authors' names were included; if the authors were unknown, they were labelled as *unknown*. The issued date corresponds to the published year as listed in the digital repository of Natlib, while the written date was determined by referring to Sannasgala (2015), as explained in the section 3.2.

The genres of the books were selected based on the details provided by Sannasgala (2015), as well as the content evaluated by authors who are native Sinhala speakers. The classification process occurs at two levels. The primary level is broad and divides the books into two categories: 'Fiction' and 'Non-Fiction.' The secondary level is more specific, categorising the content of the books into five distinct classes: religious, history, poetry, language, and medical. This approach was inspired by the methodologies followed in IcePaHC and DIAKORP (Rögnvaldsson et al., 2012; Kučera et al., 2015) corpora to ensure diverse genres. It is important to note that the first level of categorisation was applied to all documents, while the second level of categorisation was applied only to the books that fell into the selected specific categories.

Furthermore, the average OCR confidence per page is included, as mentioned in the section 3.5. In addition, we romanised the titles and author names for each book, which involved transliterating Sinhala content into the Latin alphabet. An example of a metadata record is shown in Figure 5.


```
{
  "title": "අධිමාස දීපනය",
  "title_en": "Adhimasa Dheepanaya",
  "author": "මාදම්පේ ධම්මතිලක හිමි",
  "author_en": "Madhampe Dhammathilaka Himi",
  "genre": "Non-Fiction; Religious",
  "issued_date": "1896",
  "written_date": "1850 - 1896",
  "ocr_confidence": 0.9984
}
```


Figure 5: An example of a metadata record in SiDiaC

4 Analysis of SiDiaC

4.1 OCR Performance

As previously noted, the performance of Document AI extended beyond simple text extraction as shown in Figure 6. These improvements helped to streamline manual post-processing tasks, significantly reducing the time required for that work.

Two significant additional steps were clearly observable to be performed by Document AI.

1. **Text Modernisation**: The historical development of the Sinhala language has resulted in various eras of syntax being linguistically equivalent but different in grapheme representation (Nandasara and Mikami, 2016). It is clear that Document AI adjusts to generate text in modern Sinhala syntax, ensuring a consistent and unified syntax throughout the dataset. Given that the change is only at the grapheme level, this does not violate the syntactic or semantic properties of the word.
2. **Morpheme Segmentation**: In Historical Sinhala, certain words are combined without spaces, forming a closed compound (Gaikwad and Saini, 2024). This phenomenon was accurately identified by Document AI, which effectively performs morpheme segmentation.

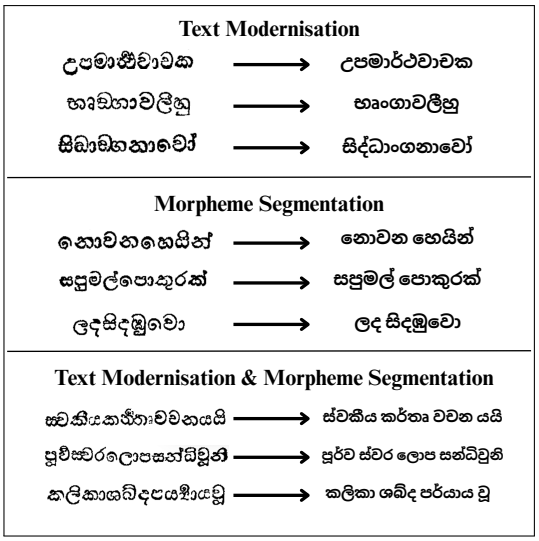


Figure 6: Examples of Sinhala Text Modernisation and Morpheme Segmentation in Document AI

Some characters in SiDiaC literature do not exist in the Sinhala Unicode. Therefore, mapping old characters to modern ones was an essential task performed by Document AI. Morpheme segmentation is crucial for maintaining consistency among words. If this issue is not addressed, combined multi-word expressions may be treated as unique words during the word embedding process, which can lead to significant differences in results when analysing semantic meaning.

4.2 Evaluation of Metadata

The dataset spans from the 5th to the 20th century CE, making it the longest continuous diachronic Sinhala corpus created to date. It covers many significant time periods, from the Anuradhapura era (377 BCE – 1017 CE) to just after Sri Lanka gained independence from Britain in 1948. This extensive timeframe allows for a representation of various changes in the language over the centuries.

Assuming that the books with specified date ranges are attributed to the upper bound year, an analysis of the number of books per century was conducted, as illustrated in Table 1. The analysis reveals that the distribution of books in the corpus is heavily skewed toward the 20th century, with 28 out of 46 records originating after the 18th century. This trend may largely be attributed to the introduction of the printing press to Sri Lanka by the Dutch in 1737, which thereafter popularised book printing in the country (Wickremasuriya, 1978; Nandasara and Mikami, 2016).

In the first level of genre classification between fiction and non-fiction, it is evident that there are more non-fiction books than fiction books in the corpus. At the second level of genre classification, religious texts and poetry dominate among the five categories. This predominance is largely due to the close relationship between Sinhala literary culture and Theravada Buddhism, which provided both subjects and a framework for preserving texts. Additionally, the influence of Sanskrit *kavya* traditions and courtly patronage, which valued literary artistry and prestige, also played a significant role (Hallisey, 2003).

The author of the book is known for 32 out of 46, while the remaining books are labelled as "Unknown." Only three authors have published more than one book in this dataset: two authors each have two books, and one author has four.

The OCR confidence levels are extremely high, with an average of 96.84% across all books and a minimum confidence score of 85.53%. Despite these encouraging figures, it is clear that Document AI encountered various types of errors, which we largely addressed during the post-processing phase as discussed in Appendix C. The accurate identification of characters and words likely contributes to these strong confidence scores; however, most errors appear to arise from the challenges presented by complex content formats.

4.3 Evaluation of the Corpus

SiDiaC consists of 58,027 word tokens that were filtered using regex, retaining only Sinhala and Latin characters, and subsequently tokenised by whitespace. The corpus contains 833 words in Latin script, which accounts for just 1.42% of the entire dataset. Also, the complete dataset comprises 22,837 unique word tokens in Sinhala script, which accounts for 39.36% unique word coverage of all words. This total word token count, while not in the range of millions, such as the COHA corpus for English, comfortably passes the 53,000 token count of FarPaHC for Faroese, which is in the same language resource category as Sinhala according to Ranathunga and de Silva (2022).

In the 5th century, 72.98% of words were unique, while the 13th century had 52.12% unique words. The 14th century saw 54.1%, the 15th century 55.97%, and the 18th century 55.05%. The 19th century featured 54.57%, but by the 20th century, the percentage dropped to 44.49%. As illustrated in Figure 7, generally a higher word count correlates with a lower percentage of unique words across the

	Primary Category		Secondary Category					Total
	Non-Fiction	Fiction	Religious	History	Poetry	Language	Medical	
5th	1	0	0	0	0	0	1	1
13th	7	1	5	0	1	2	0	8
14th	2	2	3	1	0	0	0	4
15th	1	4	2	0	3	0	0	5
18th	3	0	1	0	0	1	1	3
19th	6	3	2	2	3	2	0	9
20th	12	4	5	2	5	3	0	*16
Total	32	14	18	5	12	8	2	46

Table 1: Distribution of Books Across Centuries and Genres. *The total count for the secondary category in the 20th century amounts to 15, while the overall number of books is 16. This discrepancy arises because the book 'Hithopadhesha Sannaya', which offers advice, was not classified under any of the five secondary categories.

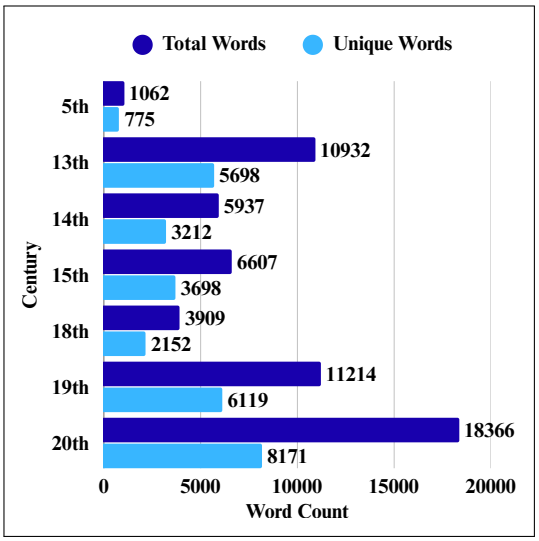


Figure 7: Total vs Unique Word Counts per Century

centuries.

Additionally, we conducted an analysis to identify the stopwords in the corpus by examining word tokens at the century level. Following the method described by Wijeratne and de Silva (2020) for their contemporary Sinhala corpus, we used a combination of word frequency analysis and manual vetting to identify appropriate stopwords from the corpora. To achieve this, we first counted the frequency of each unique word for each century and then converted these frequencies into z-scores.

$$z_{w,c} = \frac{f_{w,c} - \mu_c}{\sigma_c} \quad (1)$$

where, $z_{w,c}$ is the z-score of the word w in the century c. The term $f_{w,c}$ refers to the frequency of the word w during that century, μ_c denotes the mean frequency of all words in century c, and σ_c indicates the standard deviation of the frequencies of all words in that century.

Next, we calculated $-\infty < Z < 6.1027$ for a 99.80% threshold. In the 20th century, we observed the highest number of word tokens, which was used to establish the threshold, assuming that it provided adequate coverage of stop words throughout the entire corpus. After sorting the words in the 20th century by z-score in descending order, we manually inspected the words to determine the ideal z-score as the upper limit.

Over the centuries, the number of identified stopwords has varied accordingly. In the 5th century, there were 47 stopwords, followed by 42 in the 13th century, 39 in the 14th century, 28 in the 15th century, 44 in the 18th century, 65 in the 19th century, and finally, 61 in the 20th century.

The stopword that ranked highest across six of the seven centuries was ' වූ \vu: '(was, became), while ' හෝ \ho: '(or, either) ranked highest in the fifth century. Additionally, other frequently occurring stopwords include ' නම් \nʌm '(that, if), ' යන \ jʌnʌ '(that[named]), ' ඒ \e: '(that, those), ' මේ \me: '(this, these), ' ද \ðʌ '(or, interrogative particle), ' කොට \ kɒtʌ '(while), ' විසින් \visin '(by), and ' ඇති \æθɪ '(be [exists]). However, it was observed that certain words, such as ' වහන්සේ \vahɛnse: '(honorific for a revered person) and ' ලක්ෂණය \lʌkʃʌnʌjʌ '(Quality) were also receiving high z-scores, potentially due to the corpus's strong connection to Buddhist literature. The methodology for identifying the top 48 stopwords in the entire corpus and their depiction is presented in Appendix D.

5 Future Work

SiDiaC, being the first resource of its kind, paves the way for diachronic linguistic studies of the Sinhala language. We encourage researchers to explore areas such as lexical semantic change, tracking neologisms, grammatical change, historical language modelling, and corpus-based lexicography. Additionally, since the data is annotated by genre, it also allows for synchronic studies focusing on domain-based differences.

The corpus, as mentioned earlier, has comfortably surpassed the 53,000-token count of FarPaHC for Faroese, which falls into the same language resource category according to Ranathunga and de Silva (2022). However, the token count of SiDiaC, currently at 58,027, could certainly be increased by adding more literary works and additional pages from the existing 46 books. This enhancement would support more accurate research studies.

It is important to note that the OCR post-processing conducted in this study focuses only on formatting. However, as discussed in Appendix E, there are clear issues present at the word and character levels that need to be meticulously addressed. Furthermore, as mentioned earlier, the corpus is code-mixed with Pali, Sanskrit, and English. This highlights the need for a processing step to identify and remove irrelevant text, ensuring that the corpus is entirely focused on Sinhala.

We also identified that books called 'සන්න \sʌnnʌ' (meaning *commentaries*) may include two dates: one for the original (quoted) text and another for the commentary. However, this study did not consider this phenomenon, and such instances were attributed to the well-known original version. Therefore, in future studies, it would be beneficial to identify the two dates and include the text sections originating from the different time periods at their correct positions in the corpus.

The SiDiaC corpus did not undergo any lexical annotations during this study. Typically, the most recognised method for creating diachronic corpora involves parsing the entire corpus using POS tagging. However, this approach was not feasible with Sinhala POS taggers due to their limited performance. In future research, it would be highly beneficial to have the entire corpus parsed manually by Sinhala linguists who understand the evolution of language structure in Sinhala.

6 Conclusion

In this study, we introduced SiDiaC, a diachronic corpus of the Sinhala language. The corpus contains approximately 58k tokens, categorised into genres at two levels, and spans from the 5th century to the 20th century. This makes it the first diachronic Sinhala corpus ever created, which can serve as a foundational dataset for enhancing historical corpora in the Sinhala language. The entire process involved carefully identifying literature from the Natlib of Sri Lanka, followed by data filtering, date annotation, text extraction from PDF images, and post-processing. We also created metadata files containing important information about each book.

The complete corpus was thoroughly analysed, highlighting the powerful OCR performance of Document AI beyond simple text extraction. This was followed by a detailed evaluation of the dataset based on the metadata of all the books. Additionally, a comprehensive analysis was conducted at the word token level to ensure the identification of important findings within the corpus. Finally, we discussed potential future studies and approaches that could enhance the dataset, as well as the research opportunities that this corpus provides.

Limitations

The creation of the corpus went through different types of limitations due to various challenges we faced.

Literature Identification: While we recognised the *Department of National Archives*[21] of Sri Lanka also as a credible source, data acquisition was conducted only from the *National Library of Sri Lanka* due to permission constraints.

Data Filtration: Out of the 221 scanned copies acquired, we were able to identify the written dates or periods for only 59 of them. The written dates of the books were annotated based on the lifespans of well-known authors, while the majority of the remainder were annotated relying heavily on the work by Sannasgala (2015), which represents an over-reliance on a single source.

Post-Processing after OCR: Under this process, while corrections were initiated to address identified formatting issues, possible identification errors at the word or character level discussed in Appendix E were not addressed.

[21] https://websnew.lithium.lk/archives/

Code-Mixed Data: It was noted that the corpus contains code mixing of Pali, Sanskrit, and English languages, but the removal of text in these languages from the corpus has not been done.

Commentary Books: The identified books primarily named with the term ' සන්න \sʌnnʌ ' (meaning commentaries) will include two written dates for the original and the commentary. However, these instances were anchored to the well-known original version without removing the commentary.

Lexical Annotation: Unlike the LatinlSE, IcePAHC, COHA, and Google N-gram corpora, which have undergone lexical annotations specifically for POS tagging, we were unable to conduct similar annotations due to the unavailability of Sinhala POS taggers (de Silva, 2025).

Acknowledgments

The creation of the SiDiaC corpus was made possible through the valuable contributions of several individuals. We extend our sincere gratitude to Padma Bandaranayake, *Director of the National Library & Documentation Centre*, for her assistance with data acquisition. We also acknowledge Uthpala Nimanthi and Charani Palangasinghe for their efforts in the post-processing of the data, and the expertise of Nalaka Jayasena, a Sinhala Linguist, which was important in the identification of the book by Sannasgala (2015). Finally, we would like to thank Jayath de Silva, Savin Madapatha, and Thushan Bawantha for their dedicated work on the written date annotation.

References

Marjan Ansari, Bahram Hadian, and Vali Rezaei. 2023. Diachronic study of information structure in Persian. *Journal of Researches in Linguistics*, 15(2):65–76.

Anaëlle Baledent, Nicolas Hiebel, and Gaël Lejeune. 2020. Dating ancient texts: an approach for noisy French documents. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 17–21, Marseille, France. European Language Resources Association (ELRA).

David Bamman and Gregory Crane. 2006. The design and use of a latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78.

Douglas Biber, Edward Finegan, and Dwight Atkinson. 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. *Creating and using English language corpora*, pages 1–14.

Cheng Chen and Renping Liu. 2025. How administrative powers have impacted land-use development in China during the last 30 years: A diachronic corpus-based news values analysis. *Cities*, 159:105786.

Matthew Christy, Anshul Gupta, Elizabeth Grumbach, Laura Mandell, Richard Furuta, and Ricardo Gutierrez-Osuna. 2017. Mass digitization of early modern texts with optical character recognition. *Journal on Computing and Cultural Heritage (JOCCH)*, 11(1):1–25.

Michael C Corballis. 2017. The evolution of language.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157.

Yomal De Mel, Kasun Wickramasinghe, Nisansa de Silva, and Surangika Ranathunga. 2025. Sinhala transliteration: A comparative analysis between rule-based and Seq2Seq approaches. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 166–173, Abu Dhabi. Association for Computational Linguistics.

Nisansa de Silva. 2025. Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. *arXiv preprint arXiv:1906.02358v25*.

Hanne Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish culturomics gigaword corpus: A one billion word Swedish reference dataset for NLP. In *Proceedings of the From Digitization to Knowledge workshop at DH*, pages 8–12.

Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi, and Simonetta Montemagni. 2022. Towards the creation of a diachronic corpus for Italian: A case study on the GDLI quotations. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 94–100, Marseille, France. European Language Resources Association.

P E E Fernando. 1949. Palaeographical Development of the Brahmi Script in Ceylon from 3rd Century BC to 7th Century AD. *University of Ceylon Review*, 7(4):282–301.

Hema Gaikwad and Jatinderkumar R. Saini. 2024. Identification of closed compound words in devanagari scripted and non-devanagari scripted corpora. In *Proceedings of Fifth Doctoral Symposium on Computational Intelligence*, pages 411–418, Singapore. Springer Nature Singapore.

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. *Digitale Wissenschaft*, 157.

Charles Hallisey. 2003. Works and persons in sinhala literary culture. *Literary cultures in history: Reconstructions from South Asia*, pages 689–746.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34. Prague.

Oliver Hellwig. 2020. Dating and stratifying a historical corpus with a Bayesian mixture model. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 1–9, Marseille, France. European Language Resources Association (ELRA).

Nevidu Jayatilleke and Nisansa de Silva. 2025. Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil. *arXiv preprint arXiv:2507.18264*.

Alek Keersmaekers and Toon Van Hal. 2024. Creating a large-scale diachronic corpus resource: Automated parsing in the Greek papyri (and beyond). *Natural Language Engineering*, 30(5):1035–1064.

Thomas Klein and Stefanie Dipper. 2016. Handbuch zum Referenzkorpus Mittelhochdeutsch. Technical report, Ruhr-Universität Bochum, Sprachwissenschaftliches Institut.

Karel Kučera, Anna Řehořková, and Martin Stluka. 2015. DIAKORP: diachronic corpus of Czech, version 6. *Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague*.

Manika Lamba and Margam Madhusudhan. 2023. Exploring ocr errors in full-text large documents: a study of lis theses and dissertations. *Library Philosophy and Practice (e-journal)*, 7824.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of latin. *New methods in historical corpus linguistics*, 1(3):247–257.

S T Nandasara and Yoshiki Mikami. 2016. 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)*, 8(1).

Carolin Odebrecht, Malte Belz, Amir Zeldes, Anke Lüdeling, and Thomas Krause. 2017. RIDGES Herbology: designing a diachronic multi-layer corpus. *Language Resources and Evaluation*, 51(3):695–725.

Eva Pettersson and Lars Borin. 2019. Characteristics of diachronic and historical corpora. *Features to consider in a Swedish diachronic corpus.[online].[cit. 29. 1. 2022]. Dostupné z.*

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Han Ren, Hai Wang, Yajie Zhao, and Yafeng Ren. 2023. Time-aware language modeling for historical text dating. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13646–13656, Singapore. Association for Computational Linguistics.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic parsed historical corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey. European Language Resources Association (ELRA).

Punchibandara Sannasgala. 2015. *Sinhala Sahithya Wanshaya*. S. Godage saha Sahodarayo.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A gold standard corpus of early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 124–128, Portland, Oregon, USA. Association for Computational Linguistics.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts: the syntactic reference corpus of medieval french (srcmf). *New methods in historical corpora*, 3:275.

Ann Taylor and Anthony S Kroch. 1994. The pennhelsinki parsed corpus of middle english. *MS. University of Pennsylvania*, page 30.

Gregory Toner and Xiwu Han. 2019. *Language and chronology: text dating by machine learning*, volume 84. Brill.

Sarathchandra Wickremasuriya. 1978. The beginnings of the sinhalese printing press. In *Senarat Paranavitana commemoration volume*, pages 283–300. Brill.

Yudhanjaya Wijeratne and Nisansa de Silva. 2020. Sinhala language corpora and stopwords from a decade of sri lankan facebook. *arXiv preprint arXiv:2007.07884*.

A Utilized Literary Works

During this study, we collected 46 literary works from the Natlib of Sri Lanka, including 32 by recognised authors and the rest by 'unknown' authors. Munidhasa Kumarathunga contributed four books, Madhampe Dhammathilaka Himi and Hikkaduwe Sri Sumangala Himi contributed two each, and the remaining authors each had one book, totalling 27 unique authors.

The metadata includes the title in Sinhala, the romanised title, the author's name, the romanised author's name, the genre, the issue date, the writing date, and the OCR confidence level.

The complete metadata for each literary work used in this compilation of the dataset can be found in Table 3 with the titles and authors' names presented in romanised Sinhala. This diachronic spread ensures coverage of Sinhala evolution across medieval, pre-modern, and modern stages. Religious texts dominate, reflecting both preservation biases and the centrality of Buddhism in Sinhala literary culture.

B Comparison of Document AI & Surya

The quantitative analysis conducted by Jayatilleke and de Silva (2025) indicates that Surya outperforms Document AI when evaluated on a synthetically created Sinhala dataset. However, during our text extraction process, we found that Document AI actually surpasses Surya. We believe this discrepancy stems from the synthetic data used in their study, which does not accurately reflect the challenges presented by real scanned documents.

Table 4 highlights three examples that clearly demonstrate why Document AI is the superior OCR engine. The errors indicated in red boxes for both systems demonstrate that Document AI excels in character identification, particularly with diacritics and similar-looking letters. Additionally, Document AI appears to be the only system effectively implementing morpheme segmentation, which is crucial for maintaining consistent word forms over time. Lastly, Document AI's text modernisation feature provides another significant advantage, making it the ideal choice for integration into the OCR pipeline used in this study.

C Post-Processing Extracted Text

During this phase, we addressed six types of formatting issues. This careful task was carried out by the authors of this study, who are native Sinhala speakers.

Certain literary works contained unwanted text referred to as seal context, which did not relate to the books' actual content. As a result, this text was identified and removed from the book files in the dataset. Some examples of these seal contexts can be seen in Table 2.

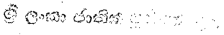
Seal Contexts	OCR Extractions
	ශ්‍රී ලංකා ජාතික ක
	ප්‍රමමිහි කෙළ මමහිසෙ
	වර ඇනි. වික්‍රමණි,
	BATAXABARA Y.C.T.
	RRalanazára

Table 2: Examples of seal contexts found in the PDF image documents and the corresponding extractions by Document AI that were removed during post-processing.

It was notable to see errors in spacing in poems, especially when the last word or letter of a word is separated by multiple spaces that were not correctly detected by Document AI, as shown in Table 5. Although correcting these errors may not hold significant semantic importance, accurately replicating the structure of the poems is crucial for preserving the original form of the books in case any downstream task using our dataset requires it. But for any task that does not require the original structure and would work on the lexical or semantic properties of the writing, the Document AI output is adequate. There were also multi-column texts, particularly evident in poetry books such as *'Kavya Wajrayudhaya - Palamu Kotasa'*. These texts contained errors where two columns were treated as a single column, resulting in entire horizontal lines being extracted without properly traversing each column, as shown in Table 6. To address these errors, corrections were made by mimicking the

Title	Author	Genre		Issued Date	Written Date	OCR Confidence↑
		Primary	Secondary			
Adhimasa Dheepanaya	Madhampe Dhammathilaka Himi	Non-Fiction	Religious	1896	1850 - 1896	0.9984
Adhimasa Winishchaya	Walikande Sri Sumangala Himi	Non-Fiction	Religious	1904	1850 - 1904	0.9971
Adhimasa Sangrahawa	Madhampe Dhammathilaka Himi	Non-Fiction	Religious	1903	1850 - 1903	0.9692
Anagathawanshaya: Methe Budu Siritha	Watadhdhara Medhanandha Himi; Siri Parakumabahu; Wilgammula Sangaraja Himi	Fiction	Religious	1934	1325 - 1333	0.9992
Ashoka Shilalipi saha Prathimakarana Winishchaya	D. E. Wickramasuriya	Non-Fiction	History	1919	1916	0.9989
Okandapala Sannaya hewath Balawathara Liyana Sanna	Don Andhris Silva	Non-Fiction	Language	1888	1760 - 1778	0.9884
Kavya Wajrayudhaya - Palamu Kotasa	Engalthina Kumari	Fiction	Poetry	1889	1825 - 1893	0.9254
Kavyashekaraya	Thotagamuwe Rahula Himi	Fiction	Poetry	1872	1408 - 1491	0.9813
Kudusika	Unknown	Non-Fiction	Poetry	1894	1270 - 1293	0.9988
Kusajathaka Wiwaranaya (Prathama Bagaya)	Munidhasa Kumarathunga	Non-Fiction	Religious	1932	1887 - 1932	0.9701
Gadaladeni Sannayai Prasidhdha wu Balawathare Purana Wyakyanaya	Hikkaduwe Sri Sumangala Himi	Non-Fiction	Language	1877	1827 - 1911	0.9970
Jubili Warnanawa	John de Silva	Non-Fiction	Language	1887	1857 - 1922	0.9957
Dhaham Sarana	Unknown	Fiction	Religious	1931	1220 - 1293	0.9891
Dhaladha Pujawaliya	Unknown	Non-Fiction	History	1893	1325 - 1333	0.9978
Dhurwadhi Hardhaya Widharanaya	Sri Dhanudhdharacharya	Non-Fiction	Religious	1899	1854 - 1899	0.9919
Dhampiya Atuwa Gatapadaya	D.B. Jayathilaka	Non-Fiction	Religious	1932	1868 - 1932	0.9241
Dharma Pradheepikawa hewath Mahabodhiwansha Parikathawa	Unknown	Non-Fiction	Religious	1906	1187 - 1225	0.9682
Dharmapradeepikawa	Gurulu Gomeen	Non-Fiction	Religious	1951	1187 - 1225	0.9786
Nikam Hakiyawa	Munidhasa Kumarathunga	Fiction	Poetry	1941	1887 - 1941	0.8932
Nikaya Sangrahaya hewath Shasanawatharaya	Unknown	Non-Fiction	Religious	1922	1390	0.9754
Nidhahase Manthraya	S Mahinda Himi	Non-Fiction	Poetry	1938	1901 - 1938	0.8997
Pansiya Panas Jathaka Potha	Unknown	Fiction	Religious	1881	1303 - 1333	0.9987
Parawi Sandheshaya	Unknown	Fiction	Poetry	1873	1430 - 1440	0.9902
Parani Gama	Galpatha Kemanandha Himi	Non-Fiction	History	1944	1944	0.9846
Budhdha Sikka hewath Kudu Sika	Unknown	Non-Fiction	History	1898	1270 - 1293	0.9699
Mage Malli	G. H Perera	Non-Fiction	Poetry	1938	1886 - 1938	0.8659
Mahawansa Teeka	Hikkaduwe Sri Sumangala Himi	Non-Fiction	History	1895	1827 - 1895	0.9978
Muwadew da Wiwaranaya	Munidhasa Kumarathunga	Non-Fiction	Religious	1949	1887 - 1944	0.8710
Moggalalayanawyakaranan	Moggallana Himi	Non-Fiction	Language	1890	1070 - 1232	0.9051
Moggallana Panchika Pradeepaya	Unknown	Non-Fiction	Language	1896	1070 - 1232	0.9918
Liyanora Nadagama	Unknown	Fiction	Poetry	1936	1852 - 1927	0.9935
Wibath Maldhama	Kirama Dhammarama Himi	Non-Fiction	Language	1906	1821	0.9986
Waidya Chinthamani Baishadhya Sangrahawa	Unknown	Non-Fiction	Medical	1909	1706 - 1739	0.9965
Wyakarana Wiwarana hewath Sinhala Bashawe Wyakaranaya	Munidhasa Kumarathunga	Non-Fiction	Language	1937	1887 - 1937	0.9029
Sadhdharma Rathnawaliya -Prathama Bagaya	Dharmasena Himi	Non-Fiction	Religious	1930	1220 - 1293	0.9962
Sanna sahitha Abhisambodhi Alankaraya	Waliwita Saranankara Sangaraja Himi	Non-Fiction	Religious	1897	1698 - 1778	0.9989
Sanna sahitha Salalihini Sandheshaya	Unknown	Fiction	Religious	1859	1450	0.9909
Sanskrutha Shabdhamalawa hewath Sanskrutha Nama Waranagilla	Rathmalane Dharmaloka Himi	Non-Fiction	Language	1876	1828 - 1887	0.9671
Sarartha Sangrahawa: Prathama Bhagaya	Srimadh Budhdhadhasa Rajathuma	Non-Fiction	Medical	1904	398 - 426	0.9997
Sithiyam sahitha Mahiyangana Warnanawa	Unknown	Fiction	Poetry	1898	1878	0.9989
Sithiyam sahitha Sadhdharmalankaraya	Unknown	Non-Fiction	Religious	1954	1398 - 1410	0.9810
Sithiyam sahitha Siyabas Maldhama	Kirama Dhammanandha Himi	Fiction	Poetry	1894	1820	0.9256
Sithiyam sahitha Sinhala Mahawanshaya	D.H.S Abhayarathna	Non-Fiction	History	1922	1874	0.9549
Sinhala Wyakaranaya enam Sidath Sangarawa	Hikkaduwe Sri Sumangala Himi	Non-Fiction	Language	1884	1827 - 1911	0.9886
Hansa Sandheshaya	C.E. Godakumbure	Fiction	Poetry	1953	1457 - 1465	0.8553
Hithopadhesha Sannaya	Waligama Sri Sumangala Himi	Non-Fiction	-	1884	1825 - 1905	0.9871

Table 3: The metadata information for all the literature used in the creation of this dataset.

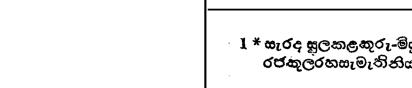
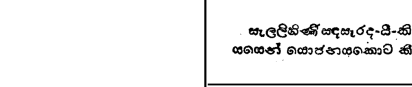
Image Example	Surya	Document AI
	1 * සැරද සුලකළකුරු-මියුරුකෙපලෙන් රදනා, රජකුලරහසැමැතිනිය-සියනිහිසැලලිහිණිසඳ.	1 * සැරද සුලකළකුරු - මියුරු කෙපලෙන් රදනා, රජකුල රහසැමැතිනිය - සියනිහි සැලලිහිණි සඳ.
	සැලලිහිණි සඳසැරද-යි-කියා ධනුර්ගුණසම්බන්ධනා යගෙන් යොජ්පායනොට කීනයි දහසුතු.	සැලලිහිණි සඳ සැරද - යි - කියා ධනුර්ගුණ සම්බන්ධන්‍යා යයෙන් යොජනය කොට කීහයි දන යුතු.
	යහනිය ... නවෙකොලොස්වසමු... බැගි නෙනොලොස්සෙ මෙහිලා, ලුහුබහලිනිබලුයේ-යහනිය (මේනිදසුන්)..... සි.	* යහගීය - නවෙකොලොස් වස්මි - බැගි නෙ කොලොස්සෙ මෙහිලා, ලුහු බහලිනි බදුයේ - යහ ගිය (මේ නිදසුන්) යි.

Table 4: Examples of sentences along with their corresponding text extractions from `Surya` and `Document AI` for comparison. Note that the characters and phrases highlighted in red boxes contain errors. † This character is known as ‘ කුණ්ඩලිය \ kuːndʌlijʌ ’, a punctuation mark that indicates the end of a text or section in historical Sinhala.

original structure of the books to preserve their intended format.

The corrections provided in the table demonstrate that OCR outputs can be restructured faithfully only through column-aware preprocessing (for example, layout analysis, region detection, or image segmentation).

The misplaced words and phrases appeared mul-

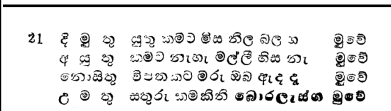
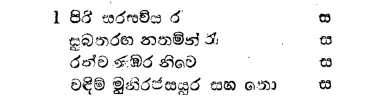
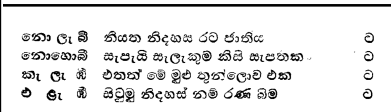
Image Example	OCR Extracted Text	Corrected Extracted Text
	21 දිමුතු යුතු කමට මිස නිල බල හ මුවේ අයුතුකමට නැහැ මල්ලි හිස නැ මුවේ නොසිතු විපතකට මරු ඔබ ඇද දැ මුවේ උ ම තු සතුරු කමකිනි බොරලැස්ග මුවේ *	21 දිමුතු යුතු කමට මිස නිල බල හ මුවේ අයුතුකමට නැහැ මල්ලි හිස නැ මුවේ නොසිතු විපතකට මරු ඔබ ඇද දැ මුවේ උමතු සතුරු කමකිනි බොරලැස්ග මුවේ
	1 පිරි සරසවිය ර ස සුබතරඟ නතමින් රැ ස රන්වණඟර නිවෙ ස වඳිමි මුනි රජසයුර සහ තො ස *	1 පිරි සරසවිය ර ස සුබතරඟ නතමින් රැ ස රන්වණඟර නිවෙ ස වඳිමි මුනි රජසයුර සහ තො ස
	නොලැබී නියත නිදහස රට ජාතිය ට නොහොබී සැපැයි සැලැකුම කිසි සැපතක ට කැලැඹී එතත් මේ මුළු තුන්ලොව එක ට ළැඹී සිටුමු නිදහස් නම් රණ බිම ට *	නොලැබී නියත නිදහස රට ජාතිය ට නොහොබී සැපැයි සැලැකුම කිසි සැපතක ට කැලැඹී එතත් මේ මුළු තුන්ලොව එක ට ළැඹී සිටුමු නිදහස් නම් රණ බිම ට

Table 5: Examples of spacing errors after OCR using `Document AI` on images and their corresponding corrections. *Note that the OCR extractions depicted were not exact; some final words were completely unidentified, which were added manually, and some had line breaks in awkward places.

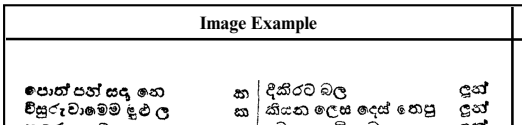
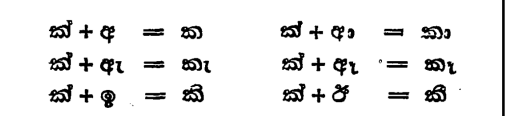
Image Example	OCR Extracted Text	Corrected Extracted Text
	පොත් පත් සදා නෙ ක දී කිරට බල ලුන් විසුරුවාමෙම මුළු ල ක කියන ලෙස දෙස් නෙපු ලුන් සතර ආගම් දෙ ක මෙලක පඬි කබ ලුන් නසන්නට තැන්කරති නොවැසැ වැණුව හැටි බල විලස සලෙලුන් *	පොත් පත් සදා නෙ ක දී කිරට බල ලුන් විසුරුවාමෙම මුළු ල ක කියන ලෙස දෙස් නෙපු ලුන් සතර ආගම් දෙ ක මෙලක පඬි කබ ලුන් නසන්නට තැන්කරති නොවැසැ වැණුව හැටි බල විලස සලෙලුන්
	ක් + අ = කා ක් + ආ = කා ක් + ඇ = කැ ක්+ ඈ = කෑ ක් + ඉ = කි ක් + ඊ = කී *	ක් + අ = කා ක් + ආ = කා ක් + ඇ = කැ ක් + ඈ = කෑ ක් + ඉ = කි ක් + ඊ = කී

Table 6: Examples of errors in multi-column text after OCR using `Document AI` on images and their corresponding corrections. *Note that the OCR extractions shown here are not exact, as we could not fully represent an entire page that experienced this type of error in real case scenarios.

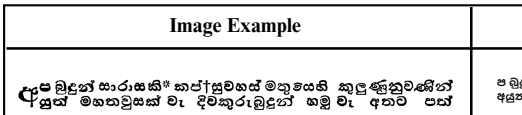
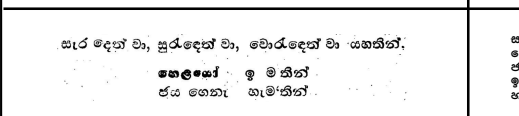
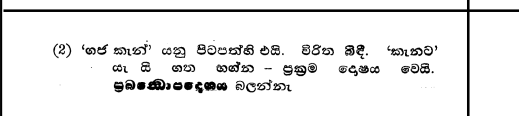
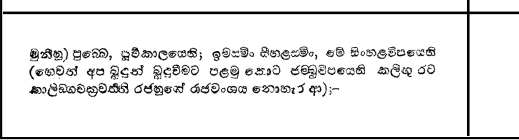
Image Example	OCR Extracted Text	Corrected Extracted Text
	ප බුදුන් සාරාසකි* කප්සුවහස් මතුයෙහි කුලුණු නුවණින් අයුත් මහත වූ සක් වැ දිවකුරු බුදුන් හමු වැ අතට පත්	අප බුදුන් සාරාසකි* කප්සුවහස් මතුයෙහි කුලුණු නුවණින් යුත් මහත වූ සක් වැ දිවකුරු බුදුන් හමු වැ අතට පත්
	සැර දෙන් වා, සුරැදෙන් වා, වොරැදෙන් වා යහනින්, හෙළයෝ ඉමනින් ජය ගෙනැ ඉමනින් හැමතින්. *	සැර දෙන් වා, සුරැදෙන් වා, වොරැදෙන් වා යහනින්, † හෙළයෝ ඉමනින් ජය ගෙනැ හැමතින්.
	(2) 'ගජ කැන්' යනු පිටපත්හි එසි. * විරිත බිඳී. 'කැනට' ප්‍රකුම දොෂය වෙසි. යැ සී ගත හඟින ප්‍රබන්ධොපදේශය බලන්නැ	(2) 'ගජ කැන්' යනු පිටපත්හි එසි. † යැ සී ගත හඟින - ප්‍රකුම දොෂය වෙසි. ප්‍රබන්ධොපදේශය බලන්නැ
	මුනිහු) පුබිකේ, පූර්ව කාලයෙහි; ඉම සම්ං (හෙවත් අප බුදුන් බුදු වීමට පළමු කාලිංග චක්රවර්ති රජහුගේ රාජ වංශය සිහළ ස්ම්ං, මේ සිංහළ ද්වීපයෙහි කොට ජම්බුද්වීපයෙහි කලිඟු රට නොහැර ආ);- *	මුනිහු) පුබිකේ, පූර්ව කාලයෙහි; ඉම සම්ං සිහළ ස්ම්ං, මේ සිංහළ ද්වීපයෙහි (හෙවත් අප බුදුන් බුදු වීමට පළමු කොට ජම්බුද්වීපයෙහි කලිඟු රට කාලිංග චක්රවර්ති රජහුගේ රාජ වංශය නොහැර ආ);-

Table 7: Examples of misplaced words and phrases, along with †errors in paragraph and line indentation that had occurred after using `Document AI` for OCR on images. Additionally, the corrections for these errors are provided. *Note that the OCR extractions displayed were not precise, as the errors were shown together rather than individually. The other errors were corrected to highlight the specific error being focused on.

Stopword	Meaning [in Context]	5th	13th	14th	15th	18th	19th	20th
වූ \ʊu:	[that which came to] be	X	X	X	X	X	X	X
නම් \nʌm	that, if	X	X	X	X	X	X	X
යන \ jʌnʌ	that [named]	X	X	X	X	X	X	X
ඒ \e:	that, those		X	X		X	X	X
මේ \me:	this, these	X	X	X	X	X	X	X
ද \ðʌ	or, interrogative particle	X	X		X	X	X	X
කොට \kɒtʌ	while		X	X	X	X	X	X
විසින් \ʋisin	by	X	X	X		X	X	X
ඇති \æθɪ	be [exists]		X	X	X	X	X	X
ලද \lʌðʌ	that [which was]	X	X	X	X		X	X
වේ \ve:	be [affirmed]		X		X	X		X
හෝ \hɔ:	or	X	X					X
හා \hɑ:	and		X	X	X	X	X	X
න \nʌ	by		X				X	X
යනු \ jʌnʊ	is [so named]		X			X	X	X
ච \tʃʌ	with, by		X				X	X
ය \ jʌ	be [affirmed]		X				X	X
වහන්සේ \ʋahɐnse: *	Honourable [suffix]		X	X			X	
කළ \kʌlʌ	do [end of action]					X	X	X
හට \hʌtʌ	to	X				X		
හෙයින් \hɛjɪn	because		X	X		X		X
ලක්ෂණය \lʌkʃʌnʌjʌ *	Quality	X						
මහා \mʌhɑ: *	Great [Prefix]		X	X			X	X
බව \bʌvʌ	[the fact] that					X	X	X
යයි \ jʌjɪ	so [called]			X		X	X	X
යි \ jɪ	be [reported]		X	X				X
කියා \kɪjɑ:	is [reported]					X	X	X
කරණ \kʌrʌnʌ	by [means of]				X		X	X
කල්හි \kʌlhɪ	while, when		X	X		X		
සේ \se:	like [in manner of]		X					X
ස \sʌ	[poetic suffix]				X		X	
නි \θɪ	be [affirmed]		X				X	
යැ \ jæ	be [reported]		X					X
මහ \mʌhʌ *	Great [Prefix]		X	X				
ර \rʌ	[poetic suffix]				X			X
වන \vʌnʌ	be [known to exist]		X					X
නන්හි \θʌnhɪ	at, after [so declared]		X			X		
එක \ɛkʌ	a/the/one						X	X
තෙම \θɛmʌ	him	X				X		
නො \nɒ	not		X					X
මෙන් \mɛn	also [in similar manner]				X			X
එහි \ɛhɪ	therein			X				X
පිණිස \pɪnɪsʌ	for						X	
බුද්ධ \bʊððʰʌ *	Buddha			X		X		
ගුණ \gʊnʌ *	Quality				X			X
වැ \væ	[after] be		X					X
මෙහි \mɛhɪ	herein							X
කර \kʌrʌ	do [end of action]				X			X

Table 8: The top 48 stopwords and their presence in each century after applying the threshold $-\infty < Z < 6.1027$. *Note that some words, which are not technically stopwords, were included in this analysis. This may have occurred due to the limited availability of literary works in certain centuries and a bias toward Theravada Buddhism in the selected literature.

-tiple times due to varying spacing and text positioning styles used in different books. Unlike simple character substitutions, these errors dislocate words or phrases from their expected syntactic and semantic slots. For example, text fragments are merged across lines, while key markers or paragraph boundaries vanish. This was particularly noticeable at the beginning of the first paragraph in texts with drop caps. Similarly, in poetry, irregular line breaks caused by extra spaces before the last letter or word were among these issues. A misplaced suffix or dislocated phrase could invert rhetorical emphasis or obscure reference chains. Importantly, these errors also impact computational parsing, where algorithms expecting consistent lineation and phrase boundaries will misinterpret discourse structure. Another one of the most common issues addressed during post-processing was the indentation errors in paragraphs and lines. These errors were often found in the first line of paragraphs, with certain poems being misaligned, and page headings being centred incorrectly. A few examples of these issues are presented in Table 7.

`Document AI` also captured meta information on the physical book, such as page numbers. During the text extraction, these were removed from the books because they do not add any value to the presented corpus. These numbers were typically placed at the top or bottom, and they were centred or right-aligned in different literary works.

D Analysis of Stop Words in `SiDiaC`

The stopword analysis was conducted using the z-score calculation, as detailed in section 4.3. During this analysis, we identified a union set of words that exceeded the established threshold, resulting in a total of 194 unique words. The words were sorted by their average z-score for each century in descending order. The list of words that were above the set threshold was cross-checked with the union set, illustrating their availability in each century. The top 48 words with the highest mean z-scores are displayed in Table 8.

The continued inclusion of particles and suffixes shows continuity of core grammatical function words. Their presence across the 5th–20th centuries suggests remarkable diachronic stability in Sinhala morphosyntactic scaffolding. However, the table also reveals anomalies: certain lexical items not traditionally classified as stopwords appear in the stopword list. This distortion is likely due to corpus composition (religious texts with Buddhist themes dominate some centuries, inflating the relative frequency of doctrinal terms). Another observation is the persistence of poetic suffixes in older centuries, gradually tapering in modern texts.
This diachronic shift may point to a movement away from poetry and towards prose. The presence/absence patterns also reveal data sparsity in some centuries, as marked by gaps where certain stopwords do not appear due to limited surviving texts. To wit, it is as interesting (mayhap more) to note what is missing in the 20th century, given that it seems to count a majority of the candidate words among its stopwords. Note how the archaic forms of honorifics and some direct references to Buddhism have dropped out of the list.

In earlier centuries, the preponderance of monastic authorship and the dominance of canonical or exegetical works meant that words indexing reverence and religious entities were unavoidable high-frequency items. Their disappearance, or at least their reduced prominence, in the most recent century may be reflecting how the subjects covered in the text have shifted from esoteric religious communication to comparatively more secular discourse. Equally important is the fact that the persistence of other grammatical function words across all centuries stands in stark contrast to this attrition of religiously marked lexemes. This points to a kind of lexical stratification: the unmarked syntactic scaffolding of Sinhala remains stable over time (only being replaced by synonyms when they do), while the culturally bound vocabulary tied to ritual, doctrine, or honorific practice is more vulnerable to historical change.

E Word & Character Level Errors

During the post-processing of extracted text from scanned PDF files, we carefully conducted formatting level corrections as mentioned in section 3.6. However, these adjustments did not resolve all the necessary corrections at the word and character levels.

It became clear that character identification issues persisted throughout the documents. As illustrated in Table 9, diacritics in Sinhala posed significant challenges. Some characters were completely unrecognised, resulting in character deletions. Additionally, some identified diacritics were incorrect substitutions for different diacritics. There were also instances where a diacritic was erroneously

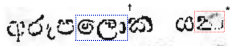
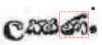
Image Example	OCR Extracted Text
	අරූපාලක යන්
	ලක්ෂණ.
ස න් න ස හි ත	සන්න සහිත ත
	කණිමුලෙ
	පූර්වග

Table 9: Examples of word deformation caused by character-level identification errors, including *incorrect identifications of diacritics/letters and **a complete deletion of a character at the marker. † Note that this is not an error, but rather a text modernisation step.

added even when no corresponding character existed, indicative of character insertion errors. In simpler terms, most of the errors in this category are related to spelling issues.