SiDiaC-v.2.0: Sinhala Diachronic Corpus Version 2.0

Nevidu Jayatilleke♣, Nisansa de Silva♣, Uthpala Nimanthi♠,
Gagani Kulathilaka♠, Azra Safrullah♠, Johan Sofalas♠

♣Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
♠Research Department, Informatics Institute of Technology, Sri Lanka

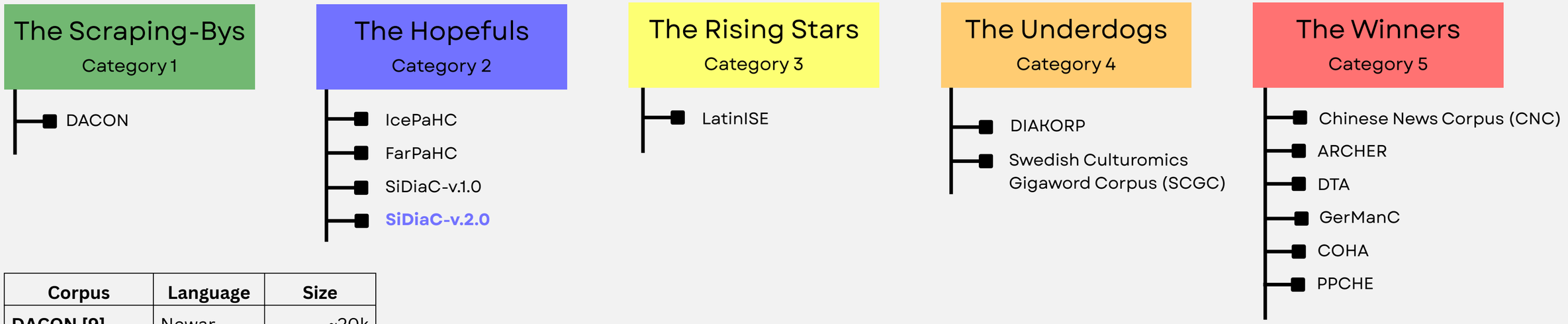
Presented By Nevidu Jayatilleke

Introduction

- SiDiaC-v.2.0 [1] is the largest comprehensive Sinhala Diachronic Corpus to date, covering a period from 1800 CE to 1955 CE in terms of publication dates, and a historical span from the 5th to the 20th century CE in terms of written dates.
- The corpus consists of 241k words across 185 literary works that underwent thorough filtering, preprocessing, and copyright compliance checks, followed by extensive post-processing.
- Additionally, a subset of 59 documents totalling 67k words was annotated based on their written dates.
- SiDiaC-v.2.0 was developed using informed practices from other corpora, such as FarPaHC [2], SiDiaC-v.1.0 [3], and CCOHA [4]. It emphasises syntactic annotation and text normalisation, which reflect the low-resource status of Faroese [5] and the cleaning strategies of CCOHA.

[1] Jayatilleke, N., de Silva, N., Nimanthi, U., Kulathilaka, G., Safrullah, A. and Sofalas, J., 2026. SiDiaC-v. 2.0: Sinhala Diachronic Corpus Version 2.0. arXiv preprint arXiv:2603.10861.
[2] Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J., 2012, May. The Icelandic Parsed Historical Corpus (IcePaHC). In LREC (pp. 1977-1984).
[3] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.
[4] Alatrash, R., Schlechtweg, D., Kuhn, J. and Im Walde, S.S., 2020, May. CCOHA: Clean corpus of historical American English. In Proceedings of the twelfth language resources and evaluation conference (pp. 6958-6966).
[5] Ranathunga, S. and De Silva, N., 2022, November. 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) (pp. 823-848).

Related Works

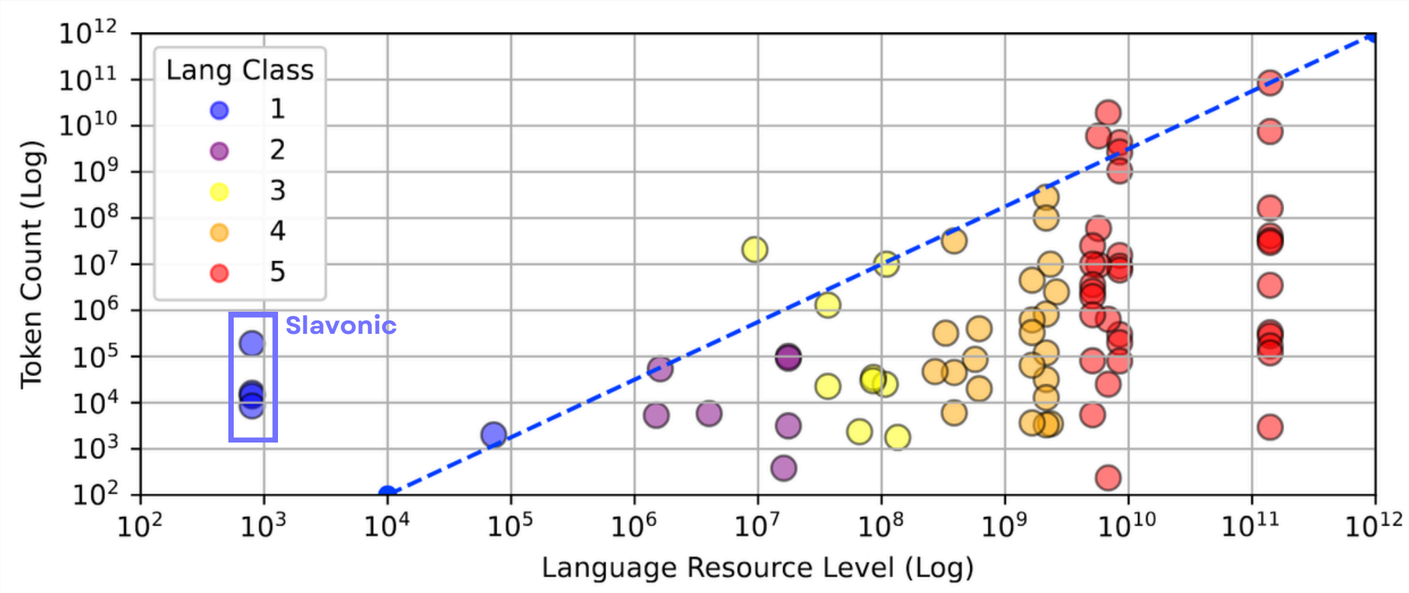


Corpus	Language	Size
DACON [9]	Newar	~20k
IcePaHC [2]	Icelandic	~1 mil
FarPaHC [2]	Faroese	~53k
SiDiaC-v.1.0 [3]	Sinhala	~58k
SiDiaC-v.2.0 [1]	Sinhala	~245k
LatinISE [10]	Latin	~13 mil
DIAKORP [11]	Czech	~4 mil
SCGC [12]	Swedish	~1 bil
CNC [13]	Chinese	~541k*
ARCHER [14]	English	~3.3 mil
DTA [15]	German	~155 mil
GerManC [16]	German	~800k
COHA [7]	English	~410 mil
PPCHE [17]	English	~4.5 mil

- The language categorisations introduced by Joshi et al. [6] were further updated by Ranathunga and de Silva [5] based on data resources, which were used to classify the corpora identified during the literature review.
- It is important to note that existing corpora vary significantly in size. For example, the FarPaHC [2] contains about 53,000 words, while the COHA [7] has approximately 400 million words.
- The size of a corpus is influenced not only by the textual resources available for a given language, but also by the level and quality of annotation provided within the corpus [8].
- We identified 80 corpora with their respective sizes and the language classes during the literature review conducted for SiDiaC-v.2.0.

[2] Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J., 2012, May. The Icelandic Parsed Historical Corpus (IcePaHC). In LREC (pp. 1977-1984).
[5] Ranathunga, S. and De Silva, N., 2022, November. 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) (pp. 823-848).
[6] Joshi, P., Santy, S., Budhiraja, A., Bali, K. and Choudhury, M., 2020. The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint arXiv:2004.09095.
[7] Davies, M., 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. Corpora, 7(2), pp.121-157.
[8] Pettersson, E. and Borin, L., 2022. Swedish diachronic corpus.

Related Works Contd.



Log token counts of the existing Diachronic corpora against the log of language resource level calculated by Ranathunga and de Silva [5]

- We observe that 10^{-2} of the resource level seems to be the soft-cap for the size of the diachronic corpora for any language class, as shown by the dashed line.
- Note the extremely low-resourced language *Slavonic* punching way above its weight due to it being included in DIACU [18], PROIEL [19] and TOROT [20].

[5] Ranathunga, S. and De Silva, N., 2022, November. 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) (pp. 823-848).
[18] Cassese, M., Puccetti, G., Napolitano, M. and Esuli, A., 2025, July. DIACU: A dataset for the DIAchronic analysis of Church Slavonic. In Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025) (pp. 101-107).
[19] Haug, D.T. and Jøhndal, M., 2008, June. 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) (pp. 27-34). Prague.
[20] Eckhoff, H.M. and Berdicevskis, A., 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS treebank.

SiDiaC-v.1.0: Sinhala Diachronic Corpus

- The first phase of the created Sinhala Diachronic Corpus, covers a historical span from the 5th to the 20th century CE.
- SiDiaC [3] comprises 58k words across 46 literary works, which comfortably passes the 53,000 token count of FarPaHC for Faroese [2], which is in the same language resource category as Sinhala according to Ranathunga and De Silva [5].
- It was annotated based on the written date or period of the literature.
- The methodology involved identifying books, followed by several filtration steps up to creating metadata files.

[3] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.
[2] Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J., 2012, May. The Icelandic Parsed Historical Corpus (IcePaHC). In LREC (pp. 1977-1984).
[5] Ranathunga, S. and De Silva, N., 2022, November. 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) (pp. 823-848).

SiDiaC-v.1.0: Sinhala Diachronic Corpus Contd.

Identifying books in Natlib

Acquiring Scanned Copies

Annotation by Written Dates

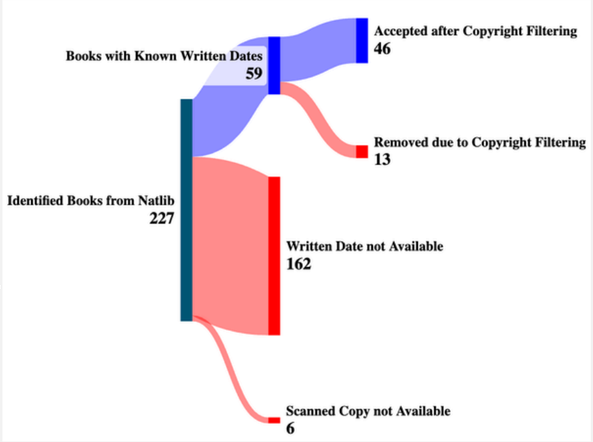
Filter based on Copyright Laws

Text Extraction using OCR

Post-processing Extracted Text

Creating Metadata Files

We identified written dates from a detailed book on Sinhala literature that covers information from Sinhala's inception up until 1994 CE [21].



The analysis showed Surya as the best OCR engine, but we found Document AI OCR provided more accurate results in realistic conditions compared to the synthetic ones used in the analysis.

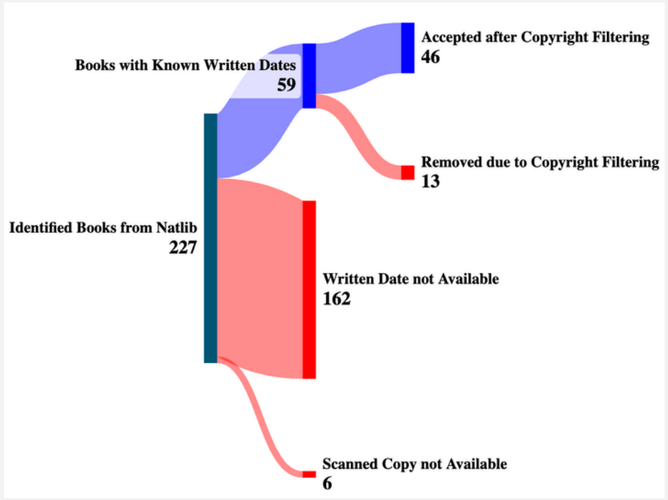
The post-processing includes correcting the following text formatting issues: Spacing Errors, Multi-column Text, Misplaced Words/Phrases, Paragraph and Line Indentation, Removal of Seal Context, and Page Number Removal.

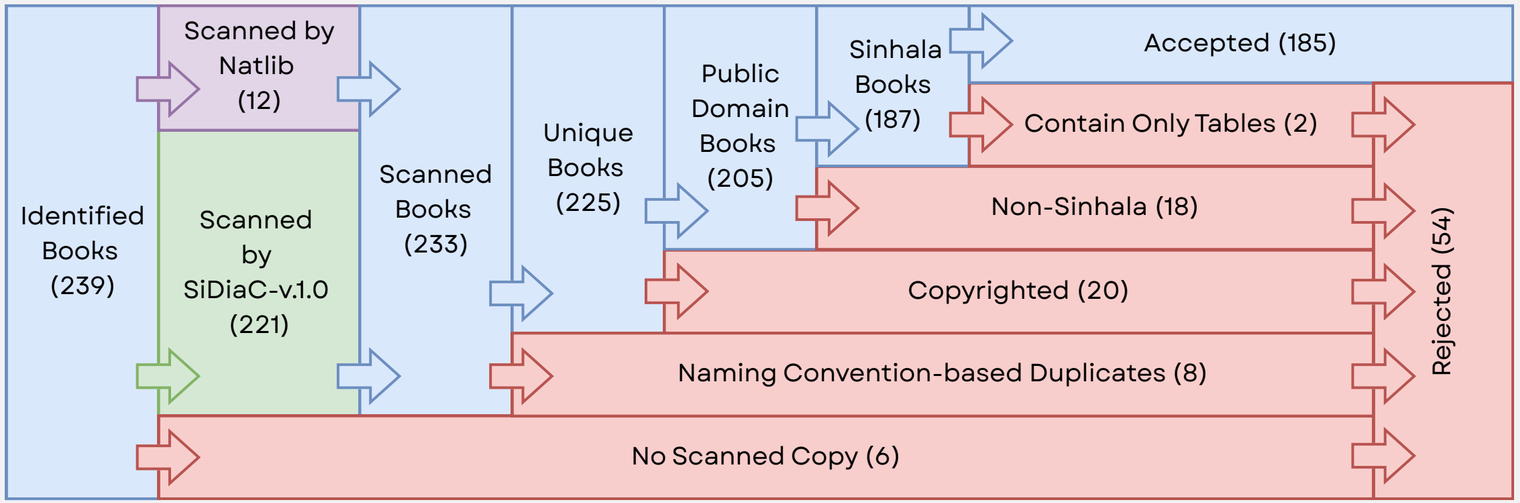
The metadata files contain information such as the title and author names in both Sinhala and romanised forms, as well as the genre, issue date, written date, and the OCR confidence level for each particular book.

[21] Sannasgala, P., 2015. Sinhala Sahithya Wanshaya. Colombo: Lake House Book.

Data Filtration of SiDiaC-v.2.0

- SiDiaC-v.1.0 filtered data, selecting only 46 books from 221 scanned copies, and the date annotations are based on a single resource.
- A crucial step in this process involved applying a written date annotation, which eliminated 168 out of the original 233 books from the identified literature.
- Some corpora, like PROIEL [19] and TOROT [20], annotate texts via manuscript dating, while others, such as COHA and RSC [22], mainly use publication years. COHA [7] also considers authors' lifespans when applicable.





[7] Davies, M., 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. Corpora, 7(2), pp.121-157.
[19] Haug, D.T. and Jøhndal, M., 2008, June. 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) (pp. 27-34). Prague.
[20] Eckhoff, H.M. and Berdicevskis, A., 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS treebank.
[22] Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and Teich, E., 2016, May. The royal society corpus: From uncharted data to corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1928-1931).

Challenges from Copyright Laws

- According to Intellectual Property Act No. 36 of 2003[1], 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.

Text Extraction using OCR

- We applied the same mechanism as SiDiaC-v.1.0 for OCR, utilising Document AI OCR [29]. This ensures consistency in modern Sinhala syntax and facilitates morpheme segmentation of closed compounds without spaces.

Annotation by Written Date

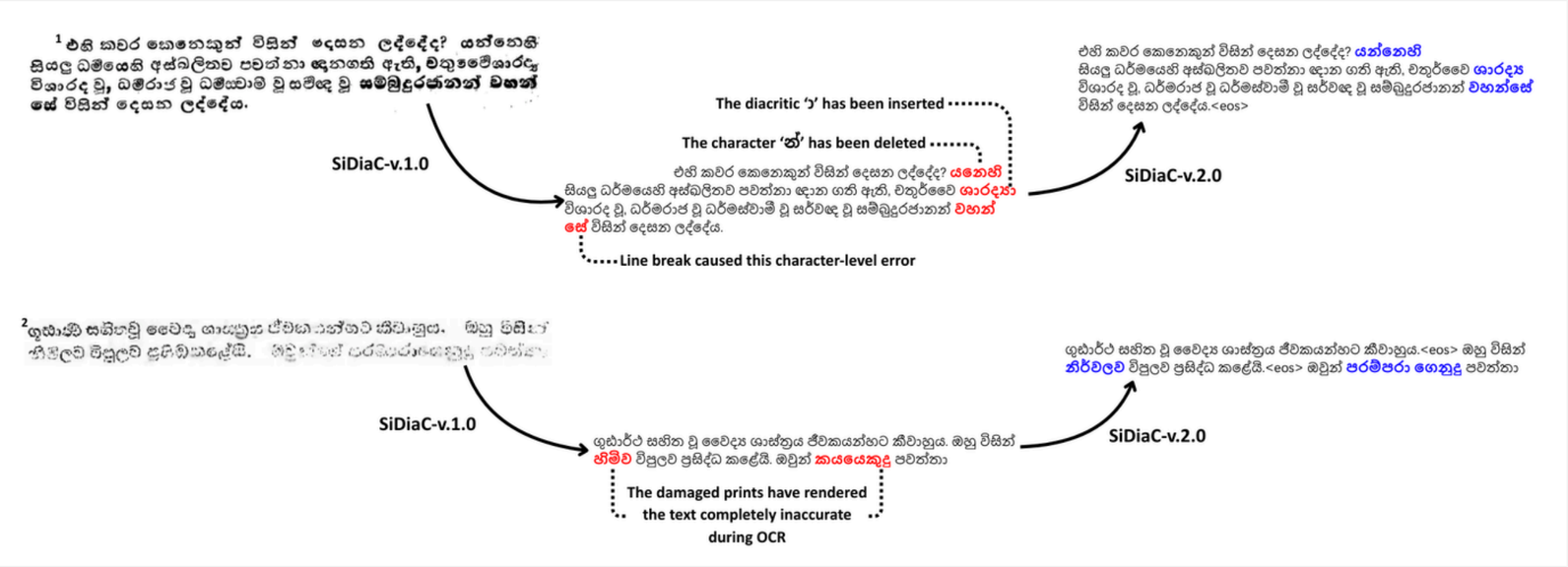
- The issue date of the identified books in the digital repository at Natlib does not reflect their actual writing dates, which could be centuries earlier.
- Unlike SiDiaC-v.1.0, which relied solely on Sinhala Sahithya Wanshaya [21], we also utilised Sri Sumangala Shabdakoshaya [23] in our current version.
- This resource lists various written periods of many books used to create the dictionary, helping us identify anchors for our corpus.

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

[21] Sannasgala, P., 2015. Sinhala Sahithya Wanshaya. Colombo: Lake House Book.
[23] Pandit Weliwitiye Soratha Thera. 2011. Sri Sumangala Shabdakoshaya: A Sinhalese-Sinhalese Dictionary. S. Godage and Brothers (Pvt) Ltd, Colombo, Sri Lanka
[29] Jayatilleke, N. and De Silva, N., 2025, September. Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil. In Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era (pp. 471-480).

Limitations of SiDiaC-v.1.0 and Post-processing of SiDiaC-v.2.0

- SiDiaC-v.1.0 filtered data, selecting only 46 books from 221 scanned copies, and the date annotations are based on a single resource.
- In the post-processing of SiDiaC-v.1.0, only five formatting issues were resolved, leaving the most pertinent word and character-level errors unaddressed.
- SiDiaC-v.1.0 comprises a mixture of Pali, Sanskrit, and English, while also containing specific works, like the Adhimasa Dheepanaya, that are entirely in Pali.



English

[1] When the whole world is like unto a feast, darling, be joyous and sing songs earnestly.

Pali

[2] රාජග්ගහ පුරාසන්නෙ ගිජ්ඣකූටම්හි පබ්බතෙ, වසන්තෙ ලොකනාථම්හි සංයමීහි යතීහිතු.

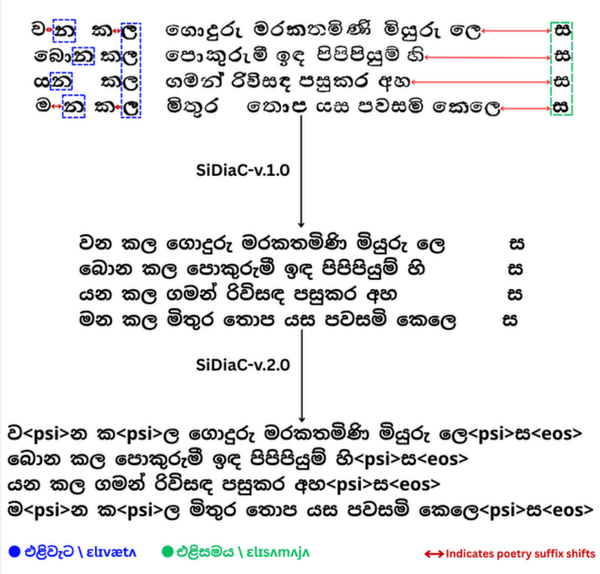
Sanskrit

[3] යථා ගන්ධවශාස්ථුල්ල මලිනො යාන්ති ශාඛ නම් තථොඝාත වශාච්ඡාස්ත්‍ර මුපගච්ඡන්ති සාධවඃ

Limitations of SiDiaC-v.1.0 and Post-processing of SiDiaC-v.2.0 Contd.

- In SiDiaC-v.1.0, books titled 'සන්න' are later commentaries on earlier works, such as the Sanna Sahitha Salalihini Sandeshaya, which has been annotated by SiDiaC-v.1.0 with the original composition date of the Salalihini Sandeshaya.
- The poems in SiDiaC-v.1.0 and SiDiaC-v.2.0 emphasise rhyming through එළිසමය and එළිවැට, which isolate and enhance rhyming sounds at line starts, middles and ends for greater auditory appeal.
- The SiDiaC-v.1.0 documents still contain page title information in the first pages and headers, which does not hold any contextual relevance.

To address code-mixing, we removed Pali and Sanskrit originals, dating the remaining Sinhala commentaries according to their specific era (e.g., anchoring the 5th-century Wishudhdhi Margaya to the 13th-century commentary). However, for dual-Sinhala texts like Salalihini Sandheshaya, we retained the original text's date to avoid chronological inconsistencies caused by anonymous or multi-era commentators.



We aimed to ensure proper word formation for future studies while preserving the original text's splitting information. To achieve this, we added a special token <psi> to indicate poetry suffix shifts.

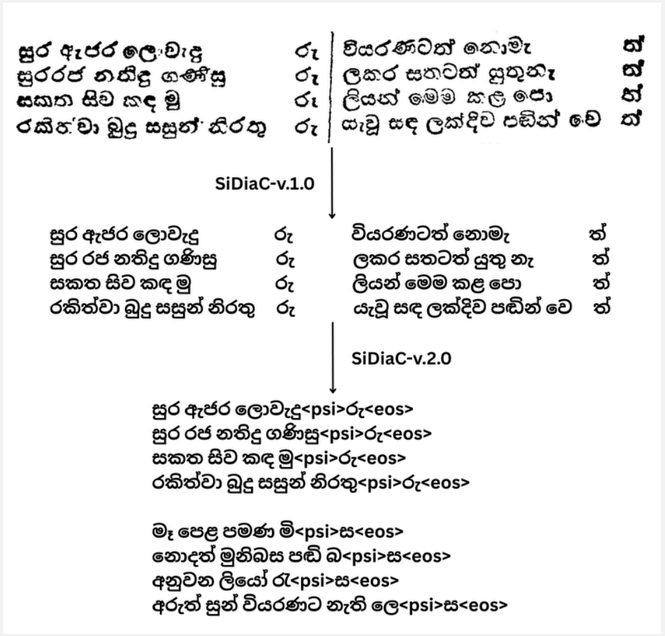
එසේ ආදියෙහි අවශ්‍ය වූ එම කරුණු විමසා බැලීමෙන් තොරව සිංහළ භාෂාවගේ නියම් ඉතිහාසාවබොධය අතිශයින් දුෂ්කරය.**<eos>** එහෙයින් යථොක්ත මාතෘකාවන් අනුව දැක්විය යුතු කරුණුද සැකෙවින් පළමුව ප්‍රකාශ කොට අනතුරුව ක්‍රමයෙන් මෙහි දක්වන මාතෘකාවන්ට ඇතුළත් වන කරුණු අනුසාරයෙන්ම දතයුතු සිංහළ භාෂා ඉතිහාසය විජයගේ අවතරණ ප්‍රවෘත්ති ආදි කථා මාත්‍රයකින් ප්‍රකාශ නොවෙයි.**<eos>**

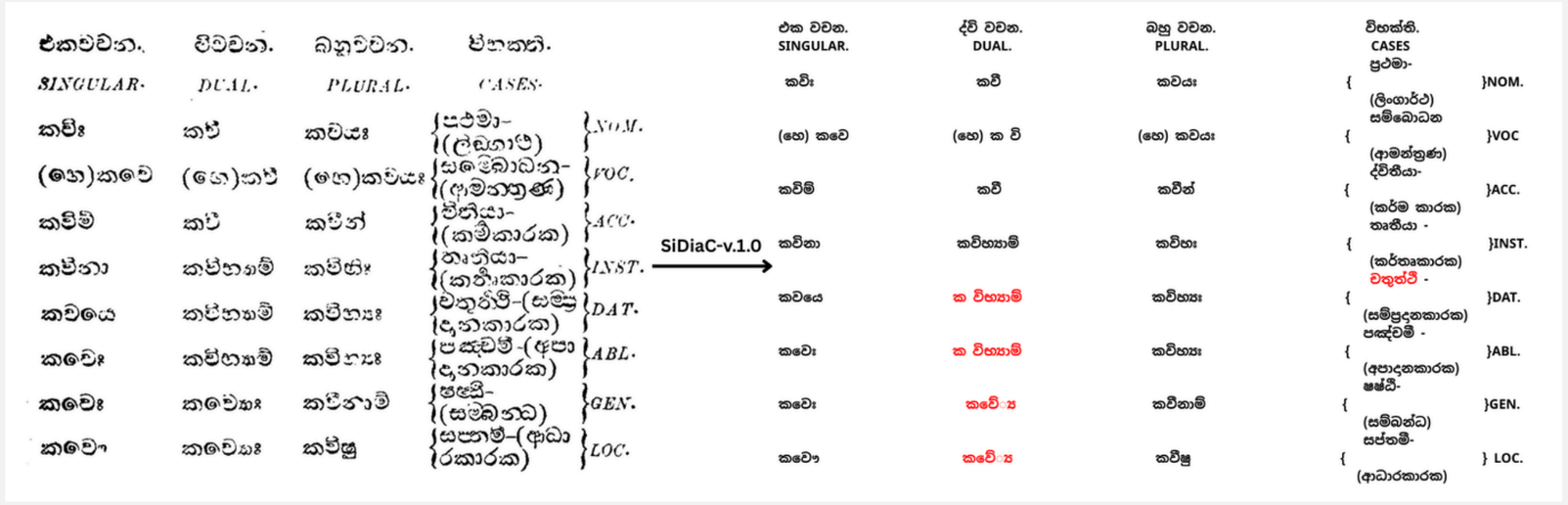
Inspired by CCOHA [4], we included an end-of-sentence-indicator <eos> to facilitate sentence-level diachronic analysis.

[4] Alatrash, R., Schlechtweg, D., Kuhn, J. and Im Walde, S.S., 2020, May. CCOHA: Clean corpus of historical American English. In Proceedings of the twelfth language resources and evaluation conference (pp. 6958-6966).

Limitations of SiDiaC-v.1.0 and Post-processing of SiDiaC-v.2.0 Contd.

- OCR in SiDiaC-v.1.0 struggled with multi-column text, leading to incorrect rendering. The use of a two-column format to mimic the original layout imposed major limitations on information retrieval from the text files.
- Some content tables in SiDiaC-v.1.0 used indentation to replicate format instead of line separations, and letters or words were spaced out by multiple spaces.





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.

Metadata Tag	Example
title [24]	**පානදුරා වාදය**
title_en [25] ∆	**Paanadhuraa Waadhaya**
author [24] ◇	**පී.ඒ. පීරිස්; පී.ජේ. දියෙස්**
author_en [25] ∆◇	**P.A. Peris; P.J. Dhiyes**
genre [2, 11]	**Non-Fiction; Religious**
issued_date [26] ✯	**1903**
written_date [27] †	**1873**
ocr_confidence [3]	**0.9945**

∆ Transliterated (Romanised) by Native Sinhala speakers
◇If the authors were unknown, they were labelled as Unknown
✯ Always included as listed in the digital repository of Natlib
† Included when applicable, sourced from Sinhala Sahithya Wanshaya and Sri Sumangala Shabda Koshaya.

The classification of genres occurs at two levels. The primary level is broad and divides books into two main 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.

[3] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.
[2] Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J., 2012, May. The Icelandic Parsed Historical Corpus (IcePaHC). In LREC (pp. 1977-1984).
[24] He, S., Zou, X., Xiao, L. and Hu, J., 2014, May. Construction of Diachronic Ontologies from People's Daily of Fifty Years. In LREC (pp. 3258-3263).
[25] Hellwig, O., Scarlata, S., Ackermann, E. and Widmer, P., 2020, May. The treebank of vedic sanskrit. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 5137-5146).
[11] Kučera, K., Řehořková, A. and Stluka, M. (2015) 'DIAKORP: diachronic corpus of Czech, version 6', Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague [Preprint]. Available at: http://www.korpus.cz/.
[26] Bouma, G., Coussé, E., Dijkstra, T. and van der Sijs, N., 2020, May. The EDGeS diachronic bible corpus. In Proceedings of the Twelfth language resources and evaluation conference (pp. 5232-5239).
[27] McGillivray, B. and Kilgarriff, A., 2013. Tools for historical corpus research, and a corpus of Latin. New methods in historical corpus linguistics, 1(3), pp.247-257.

Evaluation of the SiDiaC-v.2.0 Corpus

- SiDiaC-v.2.0 is the largest Sinhala diachronic corpus to date, comprising 241k word tokens. After filtering for written dates, it contains 67k word tokens, making it a comprehensive resource for temporal linguistic studies.
- The corpus spans the years from 1800 to 1955 CE based on publication dates and from the 5th century to the 20th century CE based on written dates.
- The dataset featuring written dates is called SiDiaC-v.2.0-filtered, while SiDiaC-v.1.0-filtered represents the SiDiaC-v.1.0 corpus that underwent the same filtration process as in SiDiaC-v.2.0.

Corpus	Post-Processing	Date-based Filtering	Documents	Total Words	Unique Words	Total Sentences
SiDiaC-v.1.0	V.1.0	✓	46	58027	22837	-
	V.2.0	✓	40	45571	16025	2970
SiDiaC-v.2.0	V.2.0	✗	185	241491	58173	11806
	V.2.0	✓	59	67005	21776	4363

- These statistics were achieved through regex-based filtering to isolate Sinhala text while excluding punctuation and Latin characters, followed by whitespace word tokenisation.

Evaluation of the SiDiaC-v.2.0 Corpus Contd.

	Primary Category		Secondary Category					Total
	Fiction	Non-Fiction	History	Language	Medical	Poetry	Religious	
1800 - 1820	1	0	0	0	0	1	0	1
1820 - 1840	0	1	0	1	0	0	0	1
1840 - 1860	2	1	0	0	0	1	2	3
1860 - 1880	2	6	1	2	0	2	3	8
1880 - 1900	12	51	3	7	1	17	34	*62
1900 - 1920	13	27	3	2	3	11	18	*37
1920 - 1940	15	35	6	4	1	17	21	*49
1940 - 1955	5	14	5	1	0	5	8	19
Total	50	135	18	17	5	54	86	185

Distribution of Books Across Issued Dates and Genres in SiDiaC-v.2.0.

*The total count for the secondary category between 1880 - 1900, 1900 - 1920, and 1920 - 1940 CE is 63, 40, and 50, respectively, while the overall number of books in those periods is 64, 43, and 51. This discrepancy arises because the books ‘Hithopadhesha Sannaya’, ‘Dhrawya Gunadharpana Sannaya’, ‘Asabandhi Sabandi’, ‘Ajuudha Neethiya’ and ‘Maadhanaa’ were not classified under any of the five secondary categories.

	Primary Category		Secondary Category					Total
	Fiction	Non-Fiction	History	Language	Medical	Poetry	Religious	
5th	0	1	0	0	1	0	0	1
12th	1	0	0	0	0	1	0	1
13th	1	12	1	3	0	1	8	13
14th	2	2	1	0	0	0	3	4
15th	4	4	0	2	0	3	3	8
16th	0	1	0	0	0	1	0	1
18th	0	4	0	1	1	0	2	4
19th	3	7	1	2	0	3	4	10
20th	5	12	2	2	0	6	6	*16
Total	16	43	5	10	2	15	26	59

Distribution of Books Across Centuries and Genres in SiDiaC-v.2.0-filtered.

*The total count for the secondary category in the 20th century amounts to 16, while the overall number of books is 17. This discrepancy arises because the book ‘Hithopadhesha Sannaya’, which offers advice, was not classified under any of the five secondary categories.

- Next, we identified consistent terms across the 13th–20th centuries and applied a z-score threshold ($-\infty < Z < 5.293$) at 99.80% confidence for the entire corpus to isolate and eliminate statistical stopwords [3, 28].
- However, due to corpus bias toward Buddhism and history, we manually audited the results to preserve domain-specific keywords before final stopword removal.

[3] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.
[28] Wijeratne, Y. and de Silva, N., 2020. Sinhala language corpora and stopwords from a decade of sri lankan facebook. arXiv preprint arXiv:2007.07884.

Evaluation of the SiDiaC-v.2.0 Corpus Contd.

- For the SiDiaC-v.2.0-filtered stopwords removed corpus, we conducted a BoW analysis with a window span of +/- 10 following the methodology by Davies [7].
- We identified 80 consistent words across eight centuries (13th-20th), a figure limited by the absence of high-accuracy morphological analysers for lemmatisation, particularly for historical Sinhala.
- Based on the identified consistent words, we selected a couple of words that have multiple meanings and conducted a manual inspection of their neighbouring words.

Neighbour	Related Meaning(s)	13th	14th	15th	16th	18th	19th	20th
ඉගෙන \ ɪgɛnʌ	Learn, Educate				2			
සිප් \ sɪp								2
ගුරුකු \ gurʊkʊ								3
ඉගෙණ \ ɪgɛnʌ							1	
සිප්සතරා \ sɪpsʌθʌrɑ:								1
ශික්ෂායි \ ʒɪkʃɑ:j		1						
විද්‍යාස්ථාන \ vɪðjɑ:sθʰɑ:nʌ		1						
ගණ \ gʌnʌ	Value, Math	5					1	
තුන් \ θʊn		1						1
පංච \ pʌŋʧʌ						3		
අටකින්ද \ ʌtʌkɪnðʌ							1	
පණසකින්ද \ pʌnʌsʌkɪnðʌ							1	
සතරක් \ sʌθʌrʌk					1			
ඩැහැගෙණ \ dæhægɛnʌ	Thief						1	
අපාය \ ʌpɑ:jʌ	Hell	1		1				
ඥාන \ dʒʰɑ:nʌ	Wisdom, Knowledge					4		
නුවණ \ nʊvʌnʌ						1		
ඥානය \ dʒʰɑ:nʌjʌ				1				
දත් \ ðʌθ							1	
දැනගන්ට \ ðænʌgʌntʌ					1			
නැණින් \ nænɪn							1	

Frequency Distribution of Neighbour Words for "සතර" from the 13th to the 20th Century.

- The word "සතර" demonstrates polysemy throughout the corpus, representing various meanings such as the number four, skills, or thief. Notably, the meaning of "thief" appears only in the 19th century, while the education-related meanings reemerge after the 16th century, having shown up only twice in the 13th century.
- Most remaining senses, including wisdom, direction, and mathematics, align with the number four, likely reflecting Buddhist concepts such as the four types of wisdom and the four hells (15th to 19th centuries), given the term's high religious significance.

[7] Davies, M., 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. Corpora, 7(2), pp.121-157.

Acknowledgements

- We acknowledge the support from the LK Domain Registry for funding granted to publish this paper.
- We sincerely thank Padma Bandaranayake, director of the National Library and Documentation Centre, for her assistance with data acquisition during the development of both SiDiaC-v.1.0 and SiDiaC-v.2.0.
- We appreciate the expertise of Ven. Thimbiriwawa Sirisumana, a senior lecturer at the Bhiksu University of Sri Lanka, along with Nalaka Jayasena as Sinhala linguists, whose insights were crucial for validating our text dating and genre classification procedures.

References

[1] Jayatilleke, N., de Silva, N., Nimanthi, U., Kulathilaka, G., Safrullah, A. and Sofalas, J., 2026. SiDiaC-v. 2.0: Sinhala Diachronic Corpus Version 2.0. arXiv preprint arXiv:2603.10861.

[2] Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J., 2012, May. The Icelandic Parsed Historical Corpus (IcePaHC). In LREC (pp. 1977-1984).

[3] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.

[4] Alatrash, R., Schlechtweg, D., Kuhn, J. and Im Walde, S.S., 2020, May. CCOHA: Clean corpus of historical American English. In Proceedings of the twelfth language resources and evaluation conference (pp. 6958-6966).

[5] Ranathunga, S. and De Silva, N., 2022, November. 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) (pp. 823-848).

[6] Joshi, P., Santy, S., Budhiraja, A., Bali, K. and Choudhury, M., 2020. The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint arXiv:2004.09095.

[7] Davies, M., 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. Corpora, 7(2), pp.121-157.

[8] Pettersson, E. and Borin, L., 2022. Swedish diachronic corpus.

[9] O'NEILL, A.J. and Meelen, M., 2024. The Diachronic Annotated Corpus of Newar: From manuscript to morphosyntax. Cahiers de Linguistique Asie Orientale, 1(aop), pp.1-30.

[10] McGillivray, B. and Kilgarriff, A., 2013. Tools for historical corpus research, and a corpus of Latin. New methods in historical corpus linguistics, 1(3), pp.247-257.

[11] Kučera, K., Řehořková, A. and Stluka, M. (2015) 'DIAKORP: diachronic corpus of Czech, version 6', Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague [Preprint]. Available at: http://www.korpus.cz/.

[12] Eide, S.R., Tahmasebi, N. and Borin, L., 2016, July. 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 (pp. 8-12).

[13] Chen, C. and Liu, R., 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, p.105786.

[14] Biber, D., Finegan, E. and Atkinson, D., 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. Creating and using English language corpora, pp.1-14.

[15] Geyken, A., Haaf, S., Jurish, B., Schulz, M., Steinmann, J., Thomas, C. and Wiegand, F., 2011. Das deutsche textarchiv: Vom historischen korpus zum aktiven archiv. Digitale Wissenschaft, 157.

[16] Scheible, S., Whitt, R.J., Durrell, M. and Bennett, P., 2011, June. A gold standard corpus of Early Modern German. In Proceedings of the 5th linguistic annotation workshop (pp. 124-128).

[17] Taylor, A. and Kroch, A.S., 1994. The Penn-Helsinki Parsed Corpus of Middle English. MS. University of Pennsylvania, p.30.

[18] Cassese, M., Puccetti, G., Napolitano, M. and Esuli, A., 2025, July. DIACU: A dataset for the DIAchronic analysis of Church Slavonic. In Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025) (pp. 101-107).

[19] Haug, D.T. and Jøhndal, M., 2008, June. 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) (pp. 27-34). Prague.

[20] Eckhoff, H.M. and Berdicevskis, A., 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS treebank.

[21] Sannasgala, P., 2015. Sinhala Sahithya Wanshaya. Colombo: Lake House Book.

[22] Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and Teich, E., 2016, May. The royal society corpus: From uncharted data to corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1928-1931).

[23] Pandit Weliwitiye Soratha Thera. 2011. Sri Sumangala Shabdakoshaya: A Sinhalese-Sinhalese Dictionary. S. Godage and Brothers (Pvt) Ltd, Colombo, Sri Lanka

[24] He, S., Zou, X., Xiao, L. and Hu, J., 2014, May. Construction of Diachronic Ontologies from People's Daily of Fifty Years. In LREC (pp. 3258-3263).

[25] Hellwig, O., Scarlata, S., Ackermann, E. and Widmer, P., 2020, May. The treebank of vedic sanskrit. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 5137-5146).

[26] Bouma, G., Coussé, E., Dijkstra, T. and van der Sijs, N., 2020, May. The EDGeS diachronic bible corpus. In Proceedings of the Twelfth language resources and evaluation conference (pp. 5232-5239).

[27] McGillivray, B. and Kilgarriff, A., 2013. Tools for historical corpus research, and a corpus of Latin. New methods in historical corpus linguistics, 1(3), pp.247-257.

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

[29] Jayatilleke, N. and De Silva, N., 2025, September. Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil. In Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era (pp. 471-480)

Thank You!

Presented by Nevidu Jayatilleke