The Dynamics of Meaning: Towards the Evaluation of Diachronic Semantic Drift in Sinhala Language

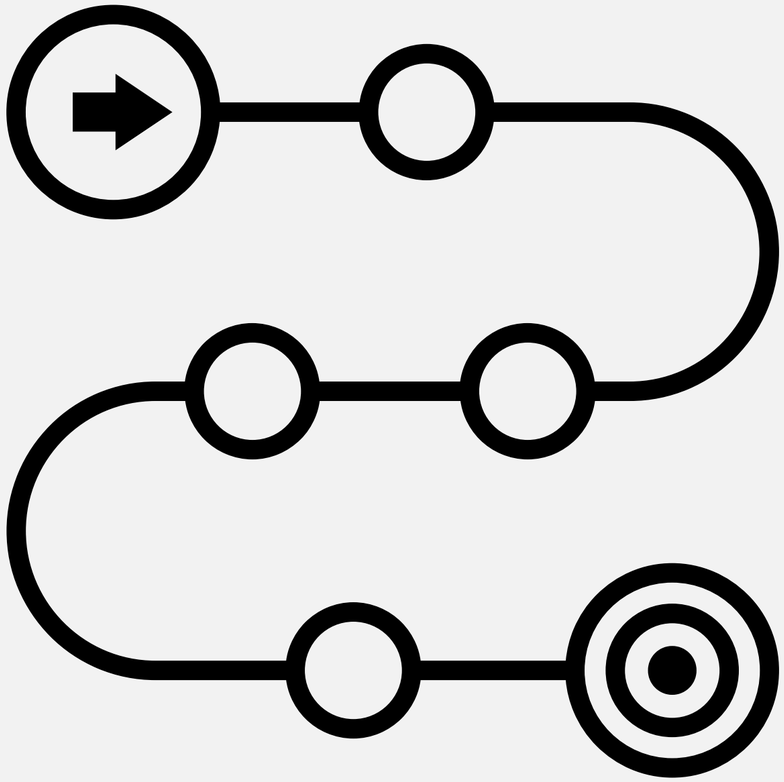
Progress Review 2

Presented By Nevidu Jayatilleke

Reg No: 258027C

Supervised By Dr. Nisansa de Silva

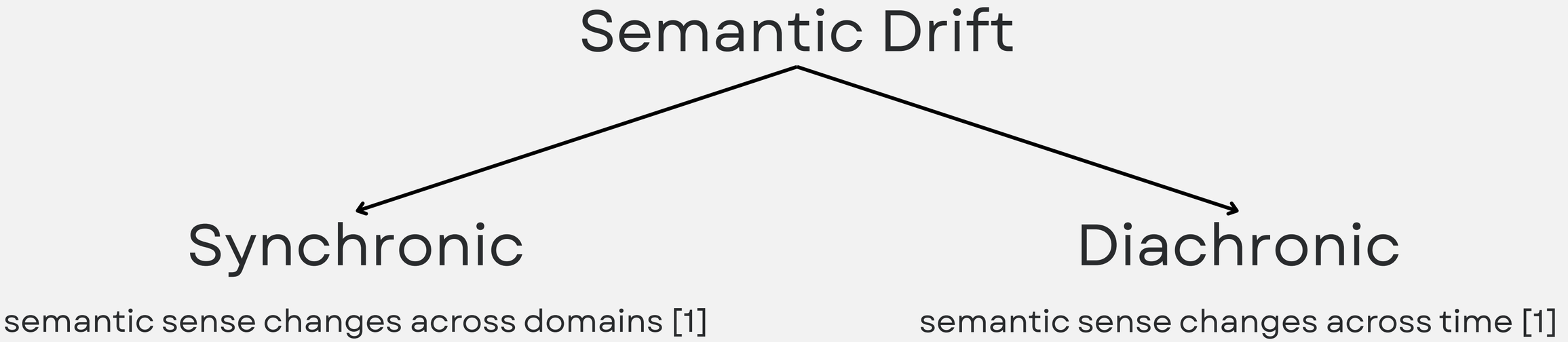
Agenda



- Introduction
- Recap of the Previous Review
- Completed Research Progress
  - Summary of Papers Published, Presented and Submitted
  - SiDiaC-v.2.0
- Ongoing and Planned Work
  - SiDiaC-v.2.5 (Ongoing)
  - Semantic Drift Analysis using SiDiaC-v.2.5 (Ongoing)
  - SiDiaC-v.3.0 (Planned)
  - Semantic Drift Analysis using SiDiaC-v.3.0 (Planned)
- References

What is Semantic Drift?

- The concept of semantic drift refers to a change in the meaning of certain words for a variety of reasons and in different contexts.
- In technical terms, it involves a shift in a word's position within the latent space.

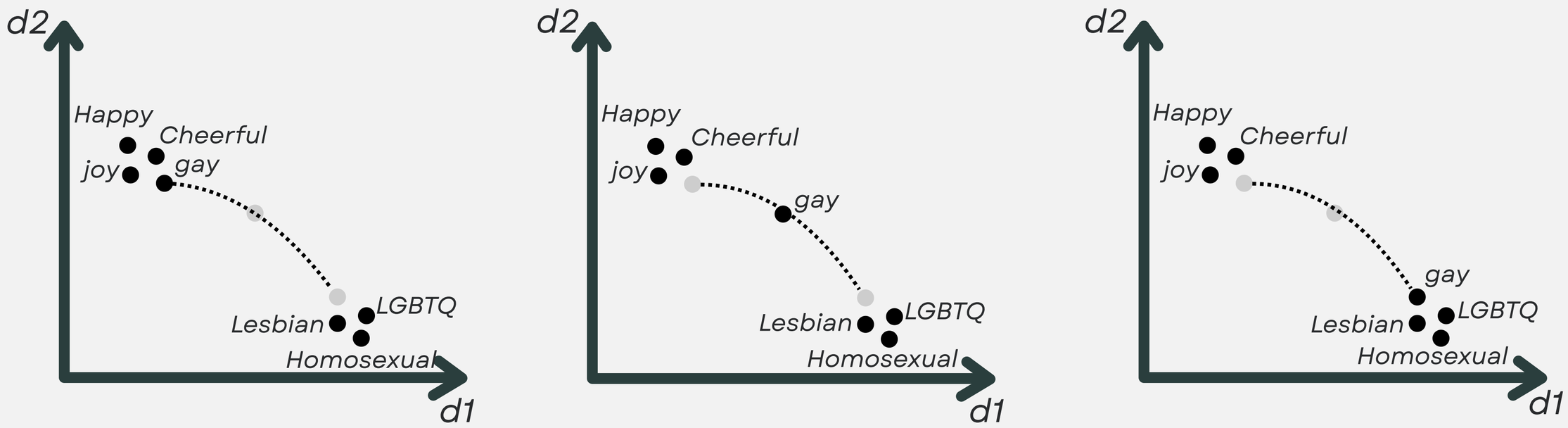


[1] Schlechtweg, D., Hätty, A., Del Tredici, M. and im Walde, S.S., 2019, July. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 732-746).

Diachronic Semantic Drift

- Semantic drift is primarily identified in diachronic studies, highlighting the evolution of meaning over time.

A well-known example is the transformation of the word "gay," which has shifted in meaning from cheerful → homosexual over the years [2]



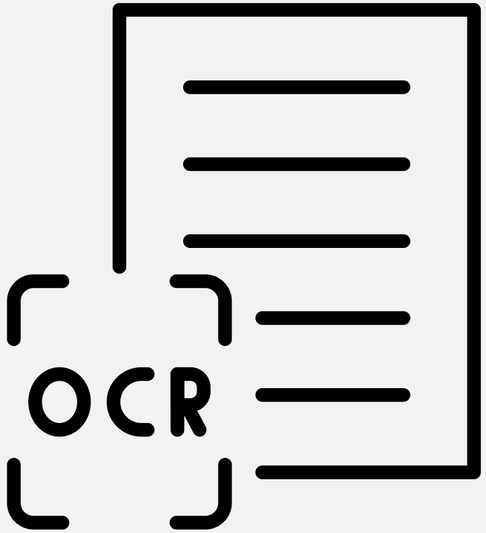
[2] Beinborn, L. and Choenni, R., 2020. Semantic Drift in Multilingual Representations. Computational Linguistics, 46(3), pp.571-603.

Summary of the Work Presented during the Previous Progress Review.

Previous Progress Review Work

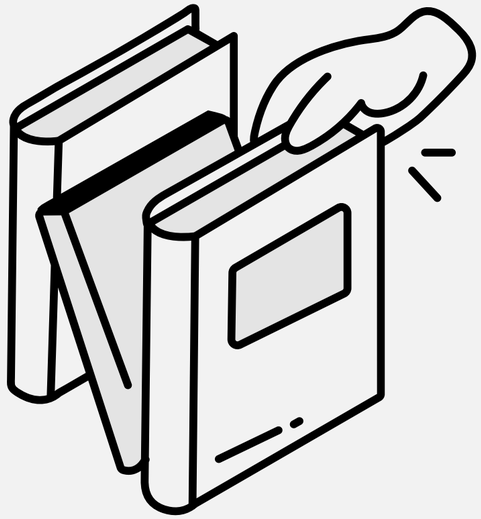
- The entire focus of the progress was on creating a Sinhala Diachronic Corpus (SiDiaC-v.1.0).
- The research update consisted of two main phases (followed by research paper submissions for each):

Identifying the best OCR engine to extract text from scanned documents.



Paper 1

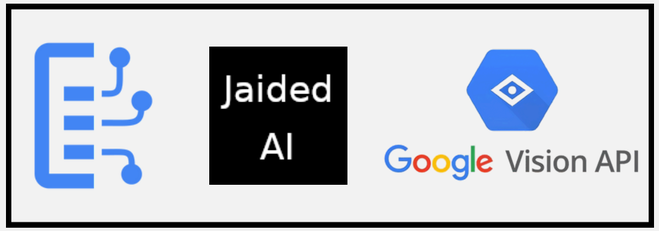
Acquiring literary works spanning multiple centuries.



Paper 2

Conclusion of the Comparative Analysis (Paper 1)

- We conducted a zero-shot comparative analysis on Sinhala and Tamil OCR engines.
- We identified Surya [3] as the best-performing OCR engine for Sinhala and Document AI [4] as the best for Tamil (we presented an original synthetic dataset for Tamil).
- When evaluating the overall results for both languages, the Document AI and Cloud Vision API [5] have achieved very similar outcomes, ranking first and second, respectively.



- This comparative study was compiled as a research paper and accepted at RANLP 2025 as a long paper [6].

[3] Vikas Paruchuri, & Datalab Team. (2025). Surya: A lightweight document OCR and analysis toolkit.
[4] Google Codelabs. (2020). Optical Character Recognition (OCR) with Document AI (Python) | Google Codelabs. [online] Available at: https://codelabs.developers.google.com/codelabs/docai-ocr-python#0 [Accessed 17 Apr. 2025].
[5] Google Cloud. (2019). Detect Text (OCR) | Cloud Vision API Documentation | Google Cloud. [online] Available at: https://cloud.google.com/vision/docs/ocr.
[6] 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).

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 [7] comprises 58k words across 46 literary works, which comfortably passes the 53,000 token count of FarPaHC for Faroese [8], which is in the same language resource category as Sinhala according to Ranathunga and De Silva (2022) [9].
- 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.

[7] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.
[8] Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J., 2012, May. The Icelandic Parsed Historical Corpus (IcePaHC). In LREC (pp. 1977-1984).
[9] 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

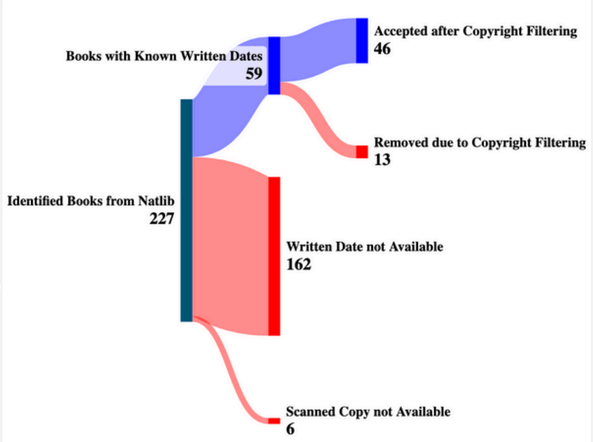
Filter based on Copyright Laws

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 [10].



We used Document AI OCR, which, under realistic conditions, yielded more accurate results than the synthetic conditions 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.

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

Let's Discuss the Progress of the Completed Research.

Summary of Papers Published, Presented and Submitted

- The comparative analysis paper titled “Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil” was presented and is included in the proceedings of RANLP 2025.
- The resource paper titled “SiDiaC: Sinhala Diachronic Corpus” was accepted and presented at PACLIC 2025.
- The dataset proposal compiled by my supervisor, titled “Sinhala Diachronic Corpus”, was accepted at the AI4Science workshop in collaboration with NeurIPS 2025 as a poster presentation.
- A resource paper titled “SiDiaC-v.2.0: Sinhala Diachronic Corpus Version 2.0” was submitted to LREC 2026 and is currently awaiting notification of acceptance.

SiDiaC-v.2.0

- SiDiaC-v.2.0 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 245k words across 186 literary works that underwent thorough filtering, preprocessing, and copyright compliance checks, followed by extensive post-processing.
- Additionally, a subset of 60 documents totalling 71k words was annotated based on their written dates.
- SiDiaC-v.2.0 was developed using informed practices from other corpora, such as FarPaHC [8], SiDiaC-v.1.0 [7], and CCOHA [11]. It emphasises syntactic annotation and text normalisation, which reflect the low-resource status of Faroese [9] and the cleaning strategies of CCOHA.

[7] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.
[8] Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J., 2012, May. The Icelandic Parsed Historical Corpus (IcePaHC). In LREC (pp. 1977-1984).
[9] 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).
[11] 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).

Related Works

The Scraping-Bys
Category 1

- DACON

The Hopefuls
Category 2

- IcePaHC
- FarPaHC
- SiDiaC-v.1.0
- **SiDiaC-v.2.0**

The Rising Stars
Category 3

- LatinISE

The Underdogs
Category 4

- DIAKORP
- Swedish Culturomics Gigaword Corpus (SCGC)

The Winners
Category 5

- Chinese News Corpus (CNC)
- ARCHER
- DTA
- GerManC
- COHA
- PPCHE

Corpus	Language	Size
DACON [13]	Newar	~20k
IcePaHC [8]	Icelandic	~1 mil
FarPaHC [8]	Faroese	~53k
SiDiaC-v.1.0 [7]	Sinhala	~58k
SiDiaC-v.2.0	Sinhala	~245k
LatinISE [14]	Latin	~13 mil
DIAKORP [15]	Czech	~4 mil
SCGC [16]	Swedish	~1 bil
CNC [17]	Chinese	~541k*
ARCHER [18]	English	~3.3 mil
DTA [19]	German	~155 mil
GerManC [20]	German	~800k
COHA [21]	English	~410 mil
PPCHE [22]	English	~4.5 mil

- The language categorisations introduced by Joshi et al. (2020) [12] were further updated by Ranathunga and de Silva (2022) [9] based on data resources 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 [8] contains about 53,000 words, while the COHA [21] 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 [13].
- We identified 79 corpora with their respective sizes and the language classes during the literature review conducted for SiDiaC-v.2.0.

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

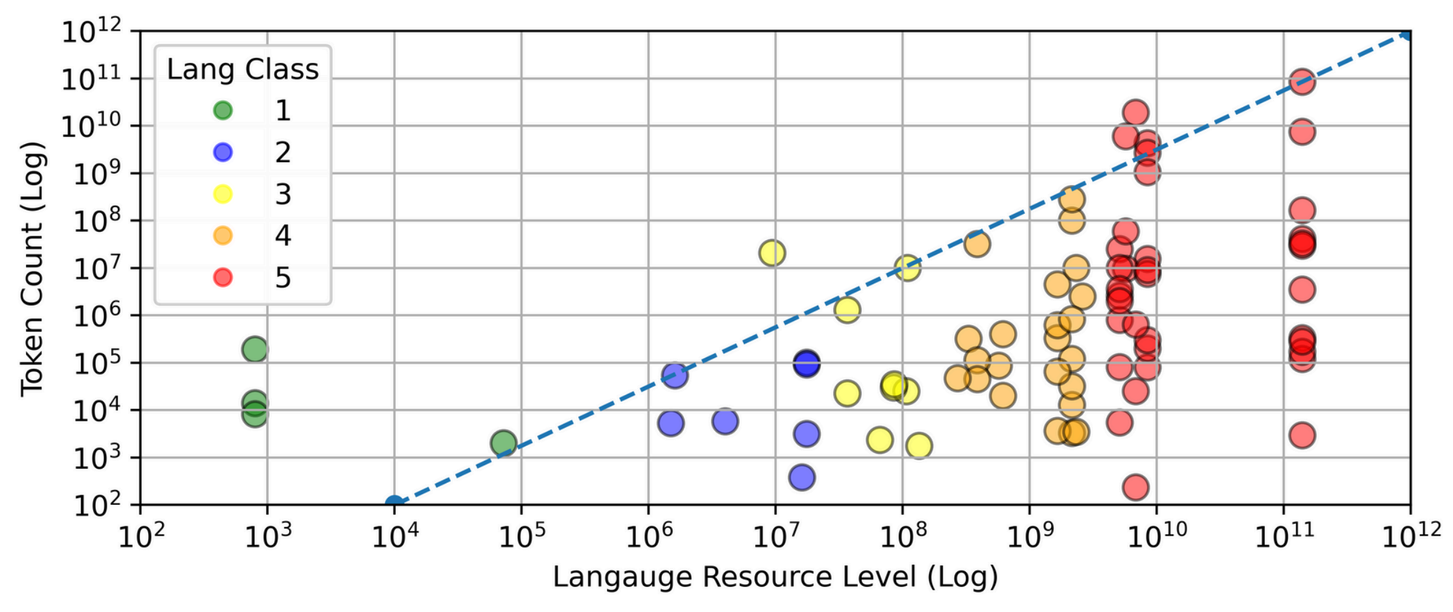
[9] Ranathunga, S. and De Silva, N., 2022, November. Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 823-848).

[12] 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.

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

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

Related Works Contd.



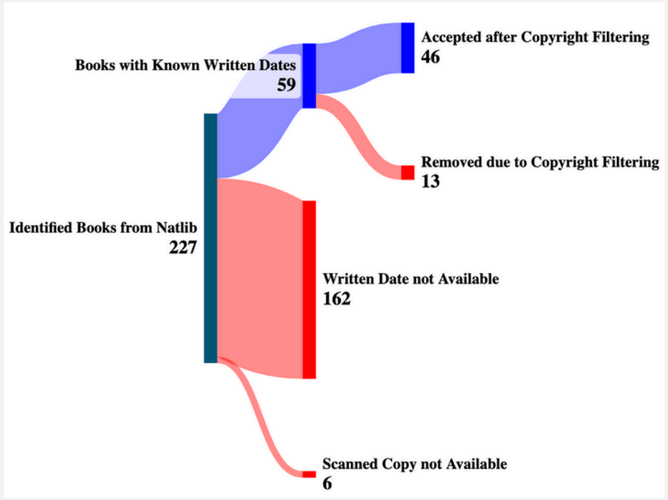
Log token counts of the existing Diachronic corpora against the log of language resource level calculated by Ranathunga and de Silva (2022) [9]

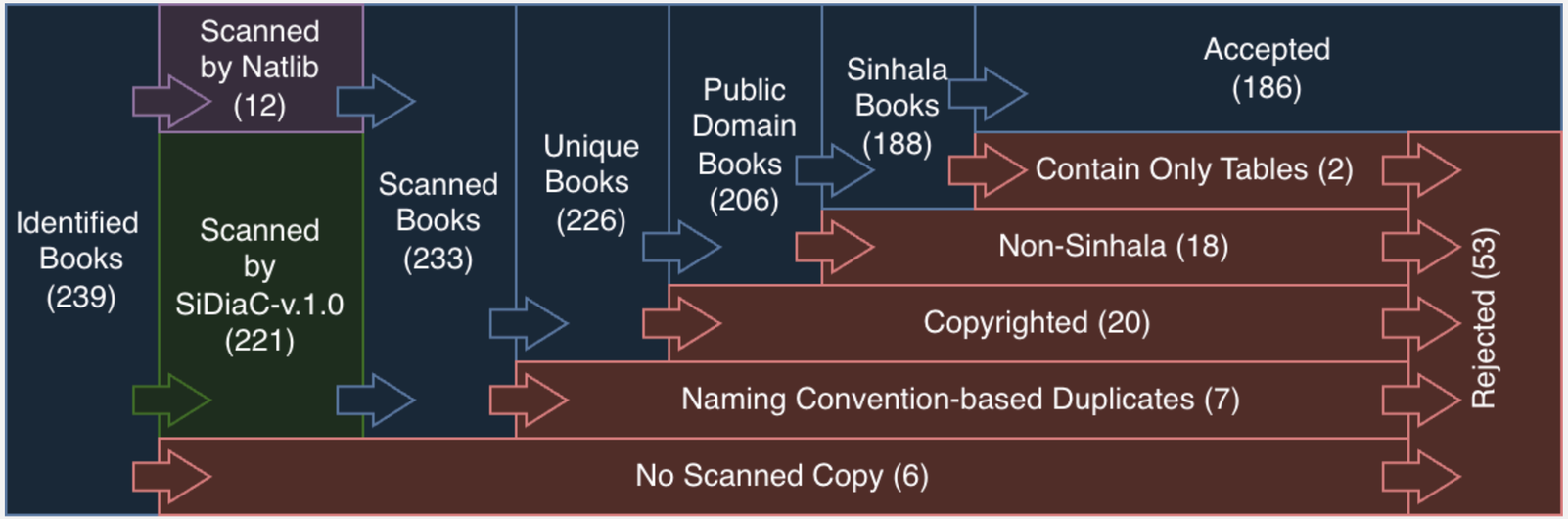
- We observe that 10^{-2} of the resource level seems to be the soft-cap for the size of the diachronic corpora for any language class, as shown by the dashed line.
- Note the extremely low-resourced language *Slavonic* punching way above its weight due to it being included in DIACU [23], PROIEL [24] and TOROT [25].

[9] 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).
[23] 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).
[24] 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.
[25] Eckhoff, H.M. and Berdicevskis, A., 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS treebank.

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 [24] and TOROT [25], annotate texts via manuscript dating, while others, such as COHA and RSC [26], mainly use publication years. COHA [21] also considers authors' lifespans when applicable.





[21] Davies, M., 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. Corpora, 7(2), pp.121-157.
[24] 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.
[25] Eckhoff, H.M. and Berdicevskis, A., 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS treebank.
[26] 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.

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 [10], we also utilised Sri Sumangala Shabdakoshaya 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.

Text Extraction using OCR

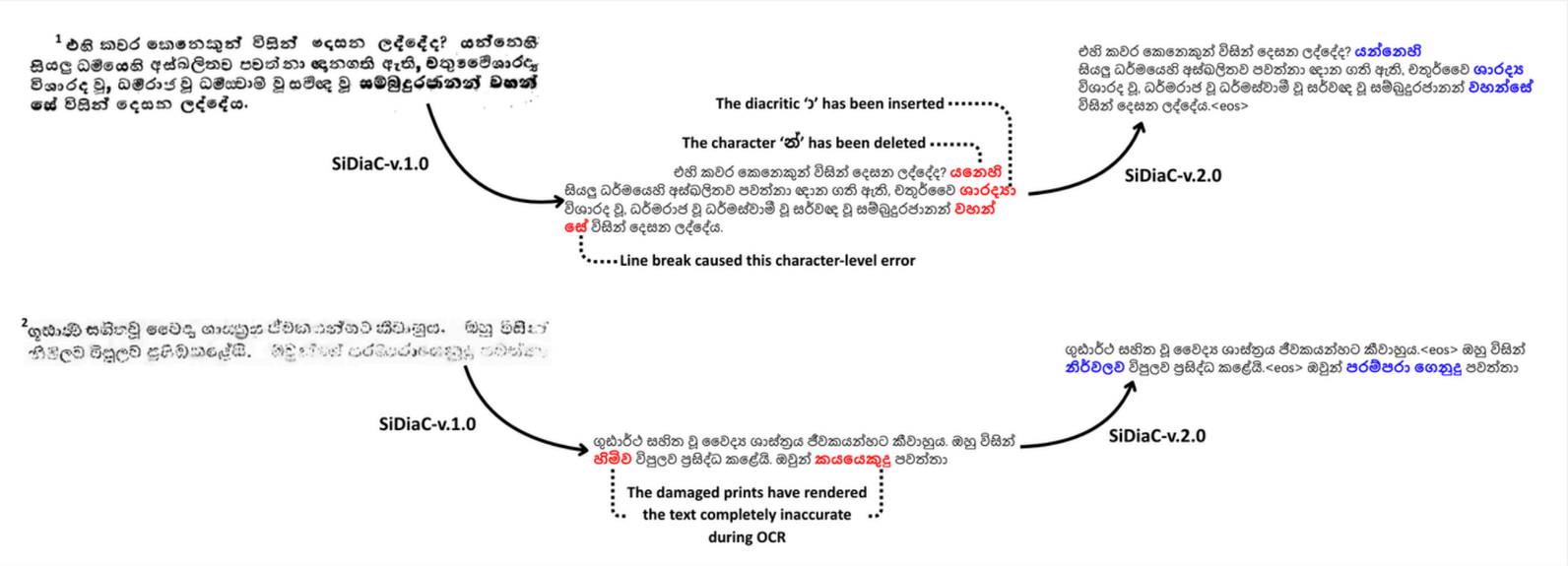
- We applied the same mechanism as SiDiaC-v.1.0 for OCR, utilising Document AI OCR. This ensures consistency in modern Sinhala syntax and facilitates morpheme segmentation of closed compounds without spaces.

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

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

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.

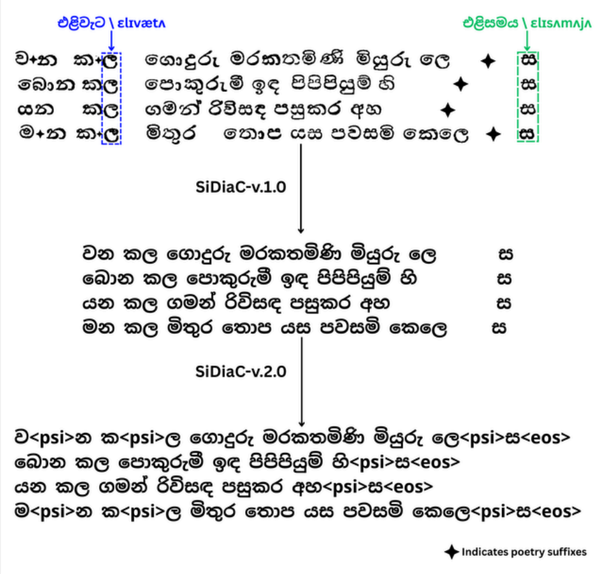
Sanskrit

[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 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.

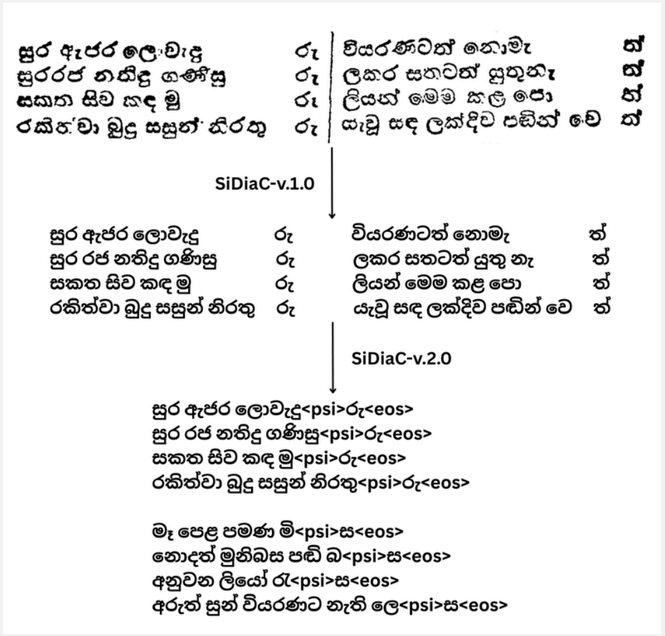
එසේ ආදියෙහි අවශ්‍ය වූ එම කරුණු විමසා බැලීමෙන් තොරව සිංහළ භාෂාවගේ නියම් ඉතිහාසාවබොධය අතිශයින් දුෂ්කරය.**<eos>** එහෙයින් යථොක්ත මාතෘකාවන් අනුව දැක්විය යුතු කරුණුද සැකෙවින් පළමුව ප්‍රකාශ කොට අනතුරුව ක්‍රමයෙන් මෙහි දක්වන මාතෘකාවන්ට ඇතුළත් වන කරුණු අනුසාරයෙන්ම දතයුතු සිංහළ භාෂා ඉතිහාසය විජයගේ අවතරණ ප්‍රවෘත්ති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

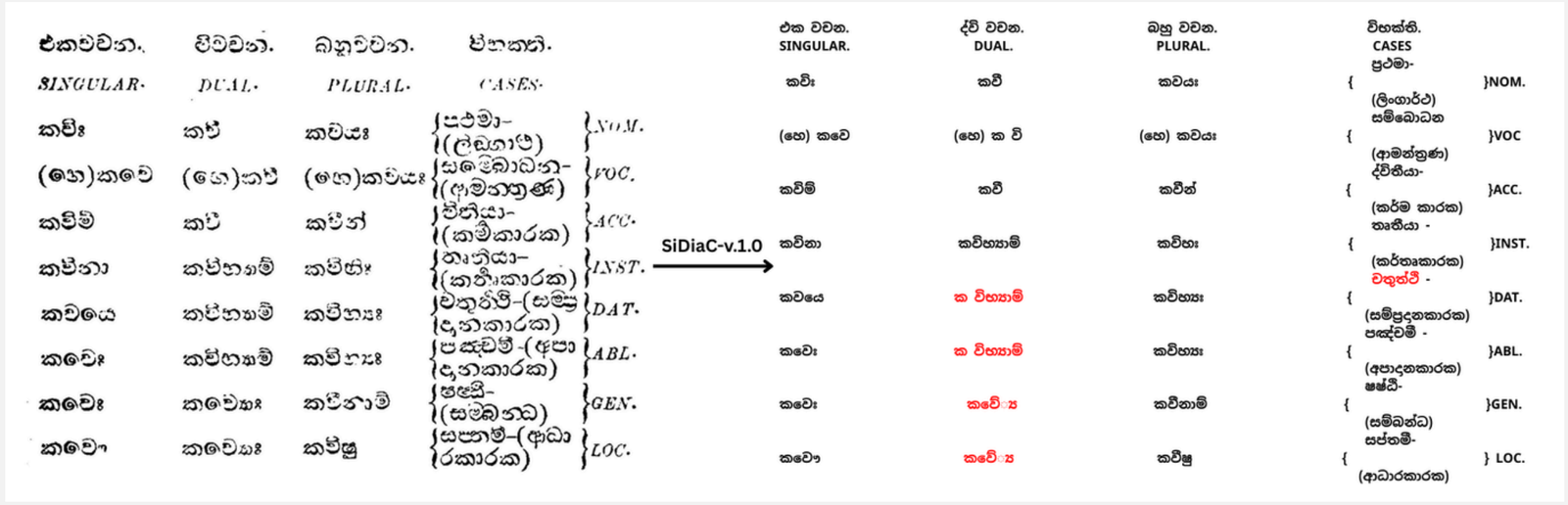
Inspired by CCOHA [11], we included an end-of-sentence-indicator <eos> to facilitate sentence-level diachronic analysis.

[11] 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 [27]	පානදුරා වාදය
title_en [28] ▲	**Paanadhuraa Waadhaya**
author [27] ◇	**පී.ඒ. පීරිස්; පී.ජේ. දියෙස්**
author_en [27, 28] ▲◇	**P.A. Peris; P.J. Dhiyes**
genre [8, 29]	**Non-Fiction; Religious**
issued_date [29] ✯	**1903**
written_date [30] †	**1873**
ocr_confidence [7]	**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 Sr 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.

[7] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.
[27] 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).
[28] 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).
[15] 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/.
[29] 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).
[30] 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 244k word tokens. After filtering for written dates, it contains 71k 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	Type	Documents	Total Words	Unique Words	Total Sentences
SiDiaC-v.1.0	Raw	46	58027	22837	-
	Filtered	41	47026	17147	1461
SiDiaC-v.2.0	Raw	186	244704	59962	6312
	Filtered	60	70654	23666	2115

- 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	*63
1900 - 1920	13	27	3	2	3	11	18	*40
1920 - 1940	15	35	6	4	1	17	21	*50
1940 - 1955	5	15	5	1	0	5	9	20
Total	50	136	18	17	5	54	87	186

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	13	1	3	0	1	9	14
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	44	5	10	2	15	27	60

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 [7, 31].
- However, due to corpus bias toward Buddhism and history, we manually audited the results to preserve domain-specific keywords before final stopword removal.

[7] Jayatilleke, N. and de Silva, N., 2025. SiDiaC: Sinhala Diachronic Corpus. arXiv preprint arXiv:2509.17912.
[31] 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 (2012) [21].
- We identified 78 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 few known words that have multiple meanings and conducted a manual inspection of their neighbouring words.
- The word සතර demonstrates polysemy across the corpus, representing the number four, skills, or a thief; notably, "thief" appears only in the 13th and 14th centuries, while education-related senses surge after the 16th 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 (13th and 15th centuries), given the term's high religious significance.

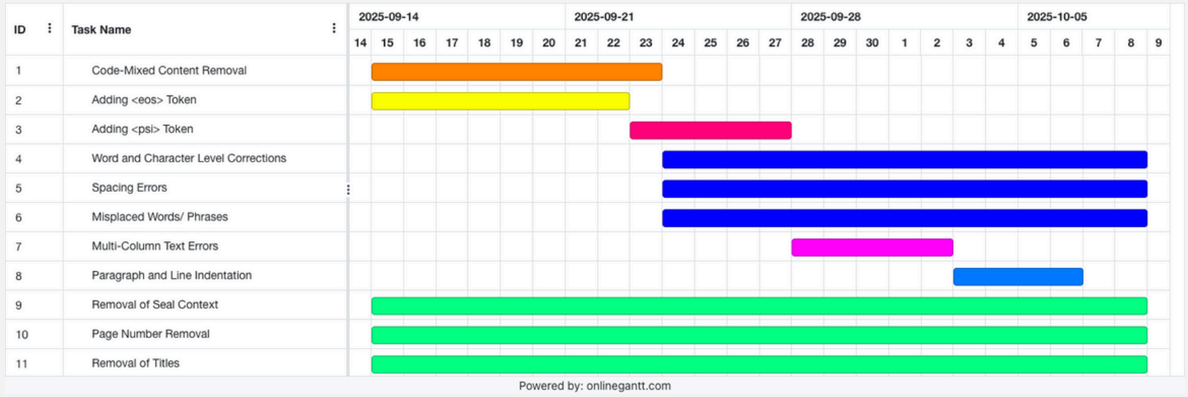
Neighbour	Related Meaning(s)	13th	14th	15th	16th	18th	19th	20th
උගණින් \ ʊgʌnɪθ	Intelligence						2	
උගතමනා \ ʊgʌθʌmʌnɑ:		1	1					
උගත් \ ʊgʌθ								1
උගතුන්යෙදූ \ ʊgʌθʊnjeðu:								1
උගන්වන \ ʊgʌnvʌnʌ	Learn, Educate						1	
උගනිවි \ ʊgʌnɪv							1	
උගනීද \ ʊgʌni:ðʌ					3			
ඉගනීමෙන් \ ɪgʌni:men								1
ඉගැන්වීම \ ɪgænvi:mʌ							1	
ඉගෙණ \ ɪgeɳʌ							1	
ඉගෙන \ ɪgenʌ		2			1			
සිප් \ sɪp							1	1
සිප්සතර \ sɪpsʌθʌrʌ								1
සිප්සතරා \ sɪpsʌθʌrɑ:								1
ගණන \ gʌnʌnʌ	Value, Math							1
ගණනක් \ gʌnʌnʌk							2	
ගණනකින්ද \ gʌnʌnʌkɪnðʌ								1
ගණන් \ gʌnʌn				3			2	2
ගණින \ gʌnɪθʌ								1
ගණිනයක් \ gʌnɪθʌjʌk								1
ගණිනයන් \ gʌnɪθʌjʌθ								1
ගණින \ gʌnɪnʌ								1
සොර \ sʊrʌ	Thief	1						
සොරාගත් \ sʊrɑ:gʌθ		1						
සොරුන් \ sʊrʊn		1	1					
සොරුන්හා \ sʊrʊnhɑ:			1					
අපාය \ ʌpɑ:jʌ	Hell	1		1				
ඥාණය \ ʤʰɑ:nʌjʌ	Wisdom, Knowledge		1					
ඥාන \ ʤʰɑ:nʌ		1	1			4		
ඥානය \ ʤʰɑ:nʌjʌ		1		1				
ඥානයක් \ ʤʰɑ:nʌjʌk							1	
ඥානයට \ ʤʰɑ:nʌjʌtʌ							1	
ඥානා \ ʤʰɑ:nɑ:								1
ඥානාදියෙන් \ ʤʰɑ:nɑ:ðɪjen								1
ඥාණ \ ʤʰɑ:nʌ								1
ඥාණයෙන් \ ʤʰɑ:nʌjen								1
ඥාණවන්ත \ ʤʰɑ:nʌvʌnθʌ								
දිශා \ ðɪʒɑ:	Direction	1						

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

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

Challenges and Deviations

- The post-processing phase faced significant challenges that required manual intervention due to mistakes made by the team.
  - A team of three undergraduates was assembled for the post-processing task.
  - Despite clear explanations of the process and transparent deadlines, fundamental mistakes were made repeatedly.
  - This situation necessitated a thorough manual review of the documents to identify and correct errors before submitting the camera-ready version following acceptance.



Planned Timeline of Manual Post-processing

- Although we initially planned to acquire newspapers during the review cycle, we decided not to pursue this approach.
  - We realised that improving the dataset synchronically would not be beneficial, as our focus is on the diachronic aspect.
  - Instead, we chose to concentrate on data acquisition related to the identified set of books and expedite the commencement of the computer science component of our research project.

Let's Discuss the Ongoing and Planned Work.

Ongoing Work on CoNLL Paper

- We are planning to submit a paper to CoNLL on February 20th, where we will unveil SiDiaC-v.2.5 and also set foot into studying diachronic semantic drift.
- SiDiaC-v.2.5 is an upgraded version of SiDiaC-v.2.0-filtered, featuring morphological analysis, lemmatisation, and Part-of-Speech (POS) tagging.
- Sinhala POS tagging is performed using the TnT POS tagger developed by Fernando and Ranathunga (2018) [32], while morphological analysis and lemmatisation are conducted with SinMorphy v1.2[2] [33](*Note that SinMorphy is also employed for guessing POS tags that the TnT POS tagger is unable to accurately assign.)
- We are developing methodologies to study diachronic semantic drift through word space alignment using a similarity matrix and two orthogonal Procrustes-based approaches on Word2Vec embeddings.
- Furthermore, we are developing an approach that incorporates semantic clustering using KMeans++ to examine diachronic semantic drift with contextualised embeddings (XLM-R and SinLlama).

[32] Fernando, S. and Ranathunga, S., 2018, May. Evaluation of different classifiers for sinhala pos tagging. In 2018 Moratuwa Engineering Research Conference (MERCon) (pp. 96-101). IEEE.
[33] Kumarasinghe, K., Dias, G. and Herath, I., 2021, July. Sinmorphy: A morphological analyzer for the sinhala language. In 2021 Moratuwa Engineering Research Conference (MERCon) (pp. 681-686). IEEE.
2. https://www.gov.lk/wordpress/wp-content/uploads/2015/03/IntellectualPropertyActNo.36of2003Sectionsr.pdf

Research Objectives

- Identifying Sinhala linguistic resources within a defined acceptable timeframe for dataset creation, and digitising them for research on diachronic semantic drift.
- Identifying semantics that exhibit significant diachronic drift and those that remain constant, serving as anchor or pivot words for further research.
- Assess the phenomenon of semantic emergence and disappearance in the evolution of the Sinhala language.
- Implementing a novel temporal dynamic embedding mechanism that captures the semantic drift of words over time by generating word vector representations for each time period.

Reaching the final objective will directly contribute to a system that can detect semantic drift.

In Progress....

Not Started Yet...

Planned Work: Leading to the Next Review

- We will provide a comprehensive update on SiDiaC-v.2.5 and share information about the diachronic semantic drift analysis we are currently conducting, targeting the CoNLL conference.

- A comprehensive update will be provided on SiDiaC-v.3.0, which is an expanded version of SiDiaC-v.2.5 that incorporates more pages from the existing 60 documents with written dates.
  - We have already collected more pages, and they have undergone OCR. Before filtration, we have approximately 250k tokens, which will provide us with sufficient data to train models and develop dynamic embeddings.
  - We are currently assembling a team of taggers (undergraduates and contract lecturers) for the task of post-processing and plan to begin the process soon.

- We also plan to start research on our final objective of building dynamic embeddings and will provide a brief update on this in the next review.

References

[1] Schlechtweg, D., Hätty, A., Del Tredici, M. and im Walde, S.S., 2019, July. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 732-746).

[2] Beinborn, L. and Choenni, R., 2020. Semantic Drift in Multilingual Representations. Computational Linguistics, 46(3), pp.571-603.

[3] Vikas Paruchuri, & Datalab Team. (2025). Surya: A lightweight document OCR and analysis toolkit.

[4] Google Codelabs. (2020). Optical Character Recognition (OCR) with Document AI (Python) | Google Codelabs. [online] Available at: https://codelabs.developers.google.com/codelabs/docai-ocr-python#0 [Accessed 17 Apr. 2025].

[5] Google Cloud. (2019). Detect Text (OCR) | Cloud Vision API Documentation | Google Cloud. [online] Available at: https://cloud.google.com/vision/docs/ocr.

[6] 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).

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

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

[9] 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).

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

[11] 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).

[12] 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.

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

[14] 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.

[15] 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/.

[16] 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).

[17] 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.

[18] 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.

[19] 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.

[20] 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).

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

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

[23] 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).

[24] 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.

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

[26] 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).

[27] 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).

[28] 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).

[29] 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).

[30] 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.

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

[32] Fernando, S. and Ranathunga, S., 2018, May. Evaluation of different classifiers for sinhala pos tagging. In 2018 Moratuwa Engineering Research Conference (MERCon) (pp. 96-101). IEEE.

[33] Kumarasinghe, K., Dias, G. and Herath, I., 2021, July. Sinmorphy: A morphological analyzer for the sinhala language. In 2021 Moratuwa Engineering Research Conference (MERCon) (pp. 681-686). IEEE.

Thank You!

Presented by Nevidu Jayatilleke