

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

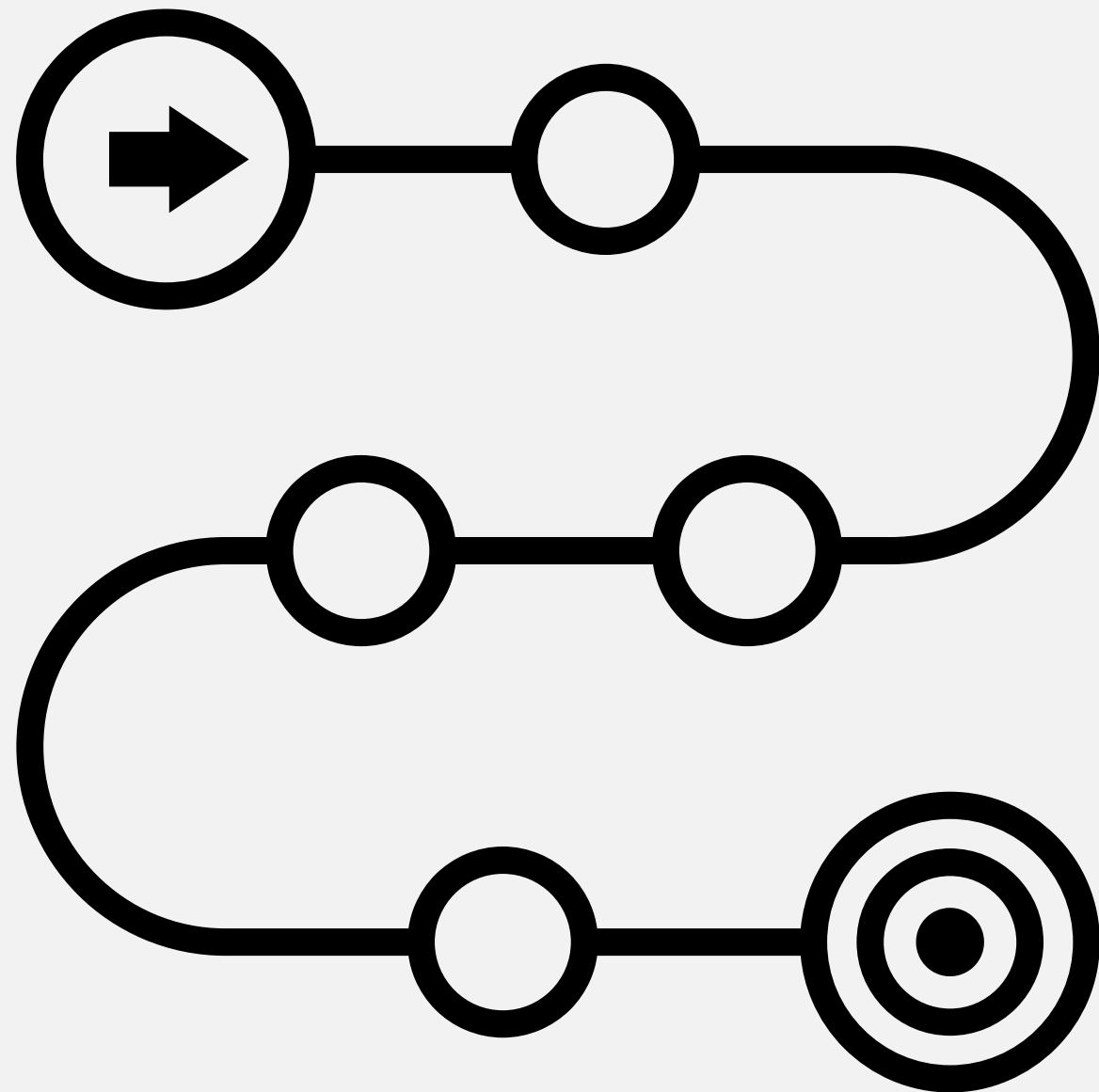
Progress Review 1

Presented By Nevidu Jayatilleke

Reg No: 258027C

Supervised By Dr. Nisansa de Silva

Agenda



- Introduction
- Research Objectives
- Initial Proposed Methodology
- Current Research Progress
 - Comparative Analysis of OCR Engines
 - Progress of the Corpus Creation Process
- Planned Work
- Papers and Grant Proposals
- References

Introduction

- Language is an evolving system, continuously reshaped by its speakers. [1].
- As we explore the fascinating process of language evolution, we find that words can shift in meaning for a variety of reasons [2].
- The Sinhala language 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 [3].
- This language has undergone significant evolution and transformation throughout its history, resulting in the form of modern Sinhala that we engage with today.



[1] Keidar, D., Opedal, A., Jin, Z. and Sachan, M., 2022, May. Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1422-1442).

[2] Hamilton, W.L., Leskovec, J. and Jurafsky, D., 2016, November. 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 (pp. 2116-2121).

[3] De Silva, N., 2019. Survey on publicly available sinhala natural language processing tools and research. arXiv preprint arXiv:1906.02358.

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.

Semantic Drift

```
graph TD; SD[Semantic Drift] --> S[Synchronic]; SD --> D[Diachronic]; S --- S_desc[semantic sense changes across domains [4]]; D --- D_desc[semantic sense changes across time [4]]
```

Synchronic

semantic sense changes across domains [4]

Diachronic

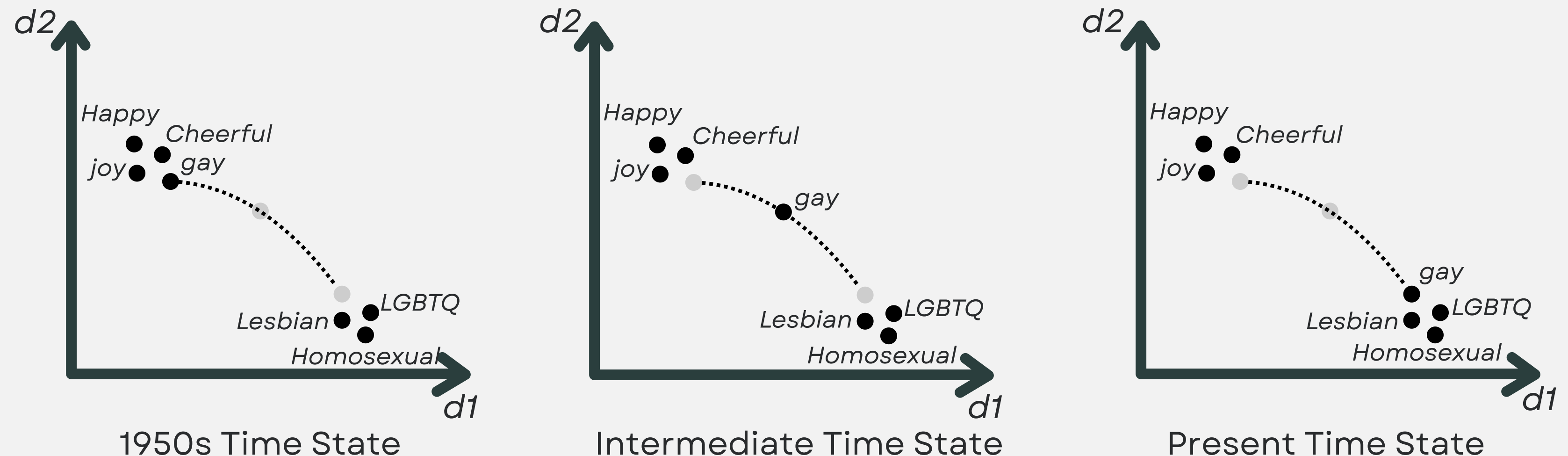
semantic sense changes across time [4]

[4] Schlechtweg, D., Häddy, 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 [5]



Diachronic Semantic Drift: Current Perspectives

- Diachronic semantic drift studies consider time as either a continuous dependency or a sequence of contiguous time intervals [6][7].
- The study area focuses on semantic drift detection [4][6], semantic appearance and disappearance [8], and dynamic embedding creation [9].
- Static word embeddings, such as those trained with algorithms like Word2Vec or FastText, are trained independently for each time slice of a corpus [10].
- Contextualized word embeddings, derived from Transformer-based language models like BERT, ELMo, and XLM-R, represent each word occurrence based on its specific context [10].
- However, research suggests that contextualized embeddings do not consistently outperform static embeddings for this specific task [10].
- There is a critical research gap in studying semantic drift in the Sinhala language, with no exploration by computational linguists and a lack of relevant datasets.

[4] Schlechtweg, D., Häddy, 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).

[6] Frermann, L. and Lapata, M., 2016. A Bayesian Model of Diachronic Meaning Change. Transactions of the Association for Computational Linguistics, 4, pp.31-45.

[7] Rosenfeld, A. and Erk, K., 2018, June. Deep neural models of semantic shift. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 474-484).

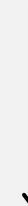
[8] Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J. and im Walde, S.S., 2021, August. Lexical Semantic Change Discovery. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 6985-6998).

[9] Hamilton, W.L., Leskovec, J. and Jurafsky, D., 2016, August. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1489-1501).



[10] Periti, F., Dubossarsky, H. and Tahmasebi, N., 2024, March. (Chat) GPT v BERT Dawn of Justice for Semantic Change Detection. In Findings of the Association for Computational Linguistics: EACL 2024 (pp. 420-436).

Research Objectives



- Identifying Sinhala linguistic resources within a defined acceptable timeframe for dataset creation, and digitizing 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...

Proposed Methodology

- We will create a data set for research using Sinhala linguistic resources, including:
 - Material from the Internet
 - Books
 - Articles
 - Newspapers
 - Physical resources will be digitized using existing OCR systems for the Sinhala language.
 - The research aims to:
 - Identify effective techniques for recognizing semantic changes over time.
 - Identify stable semantic changes within the latent space.
 - Evaluation mechanisms will be developed to assess:
 - Emergence of new semantics.
 - Disappearance of old semantics in the evolution of the Sinhala language.
 - We aim to establish the best methodology for:
 - Creating dynamic word embeddings that capture semantic drift.
 - Producing word representations that evolve over time.
-  **In Progress....**
 **Not Started Yet...**




Data Acquisition Plan

- We identified several sources to acquire linguistic resources beyond the Internet archives;
 - Department of Archives
 - National Library
 - Museum Libraries
 - University Libraries
 - Royal Asiatic Society of Sri Lanka
 - Buddhist Cultural Centre
 - Sinhala Dictionary Office
 - Postgraduate Institute of Archaeology
 - Department of Official Languages

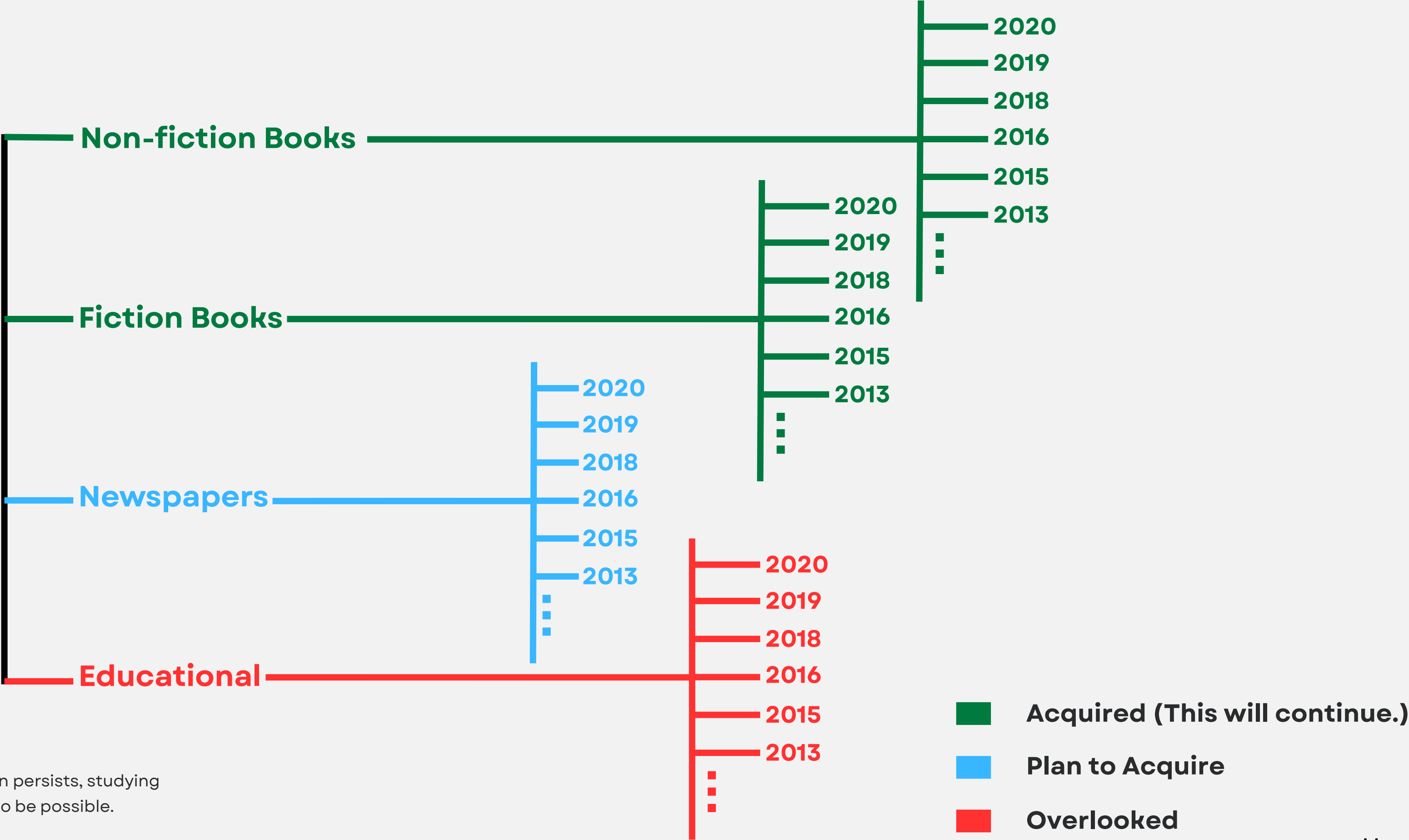
-  In Progress....
-  Plan to Reach in Phase 2...
-  Not Reached Yet...

Initial Dataset Assembly Strategy

- The dataset aimed to consist of **yearly data**. (Consists periods in which the literature selected were written [11])
- The aim was to collect literature published:
 - **Ranging from the latest to the earliest** (e.g., 2024, 2023, 2022, 2021, 2019, 2016, etc.).
 - Till 1737, which marks the beginning of the printing press in Sri Lanka. (They may have been written before 1737.)
- Initially, 100 sentences will be collected per year. (Approximately 1500 - 2000 word tokens)
- The goal was to categorize the dataset into distinct classes based on the type of literature.
- The identified categories are as follows;
 - Non-fiction books - high priority
 - Fiction books - high priority
 - **Newspapers - high priority**
 - **Educational - medium priority**
 - **Gov and Law - low priority**

-  **Followed**
-  **Plan to Follow**
-  **Diverged**

Dataset Overview



Note: If this data set creation plan persists, studying synchronic semantic drift will also be possible.
(Out of Scope for this study)

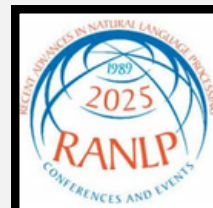
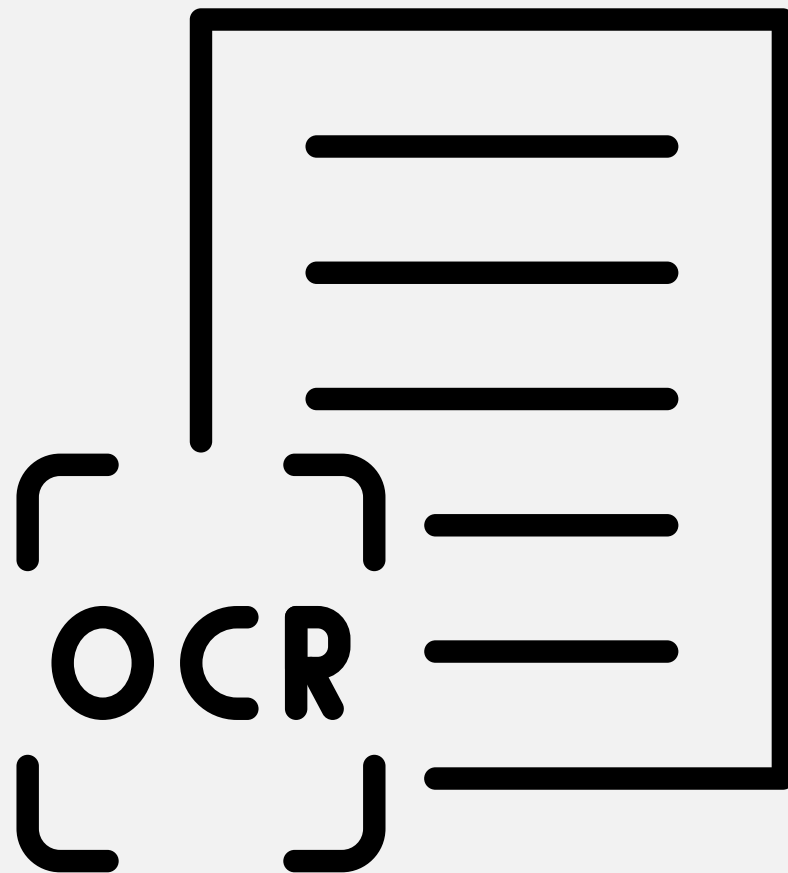
Let's Discuss Every Step since the Proposal Defense !!!



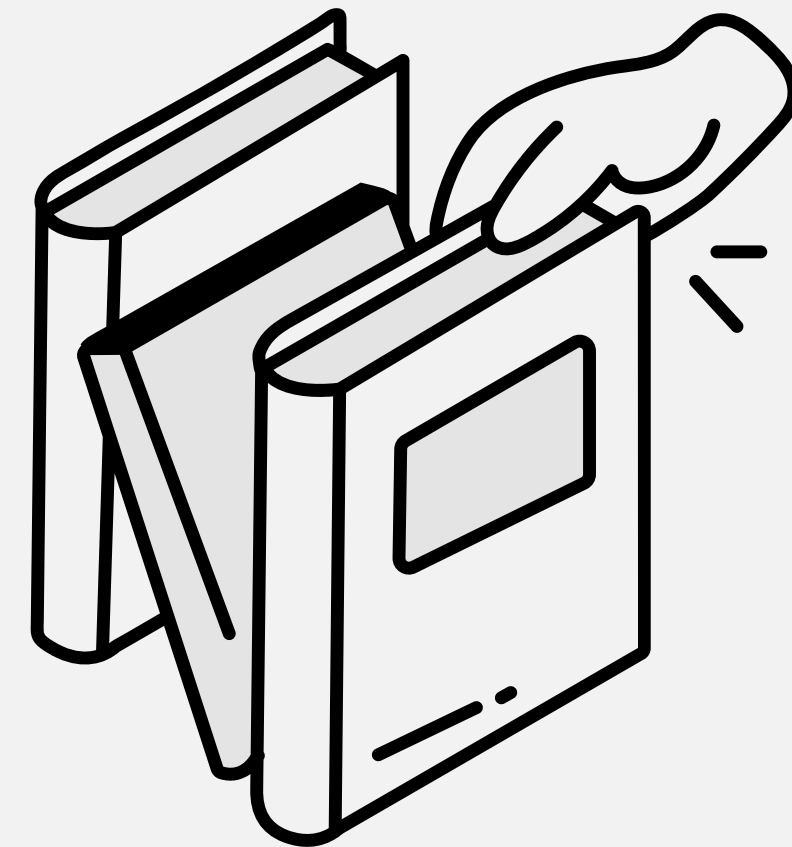
Current Research Progress

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

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



Acquiring literary works spanning multiple centuries.



Evaluation of OCR Engines for Sinhala and Tamil

- The majority of our literature resources will primarily be in physical format, as we are focusing on Sinhala texts that span several centuries. (Currently all selected resources were scanned copies of physical books)
- As we needed to identify the best OCR engine for Sinhala, we conducted a comparative analysis (Later included Tamil, as the results showed promise for a possible research paper).
- We evaluated six different OCR engines for both Sinhala and Tamil languages;
 - Subasa OCR [12]
 - Tesseract OCR [13]
 - OCR with Document AI [14]
 - Google Cloud Vision API [15]
 - Surya OCR [16, 17]
 - Easy OCR [18]



[12] LTRL UCSC, "Sinhala ocr," 2022. [Online]. Available: <https://ocr.subasa.lk>

[13] Tesseract OCR. (n.d.). Tesseract documentation. [online] Available at: <https://tesseract-ocr.github.io/>.

[14] 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].

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

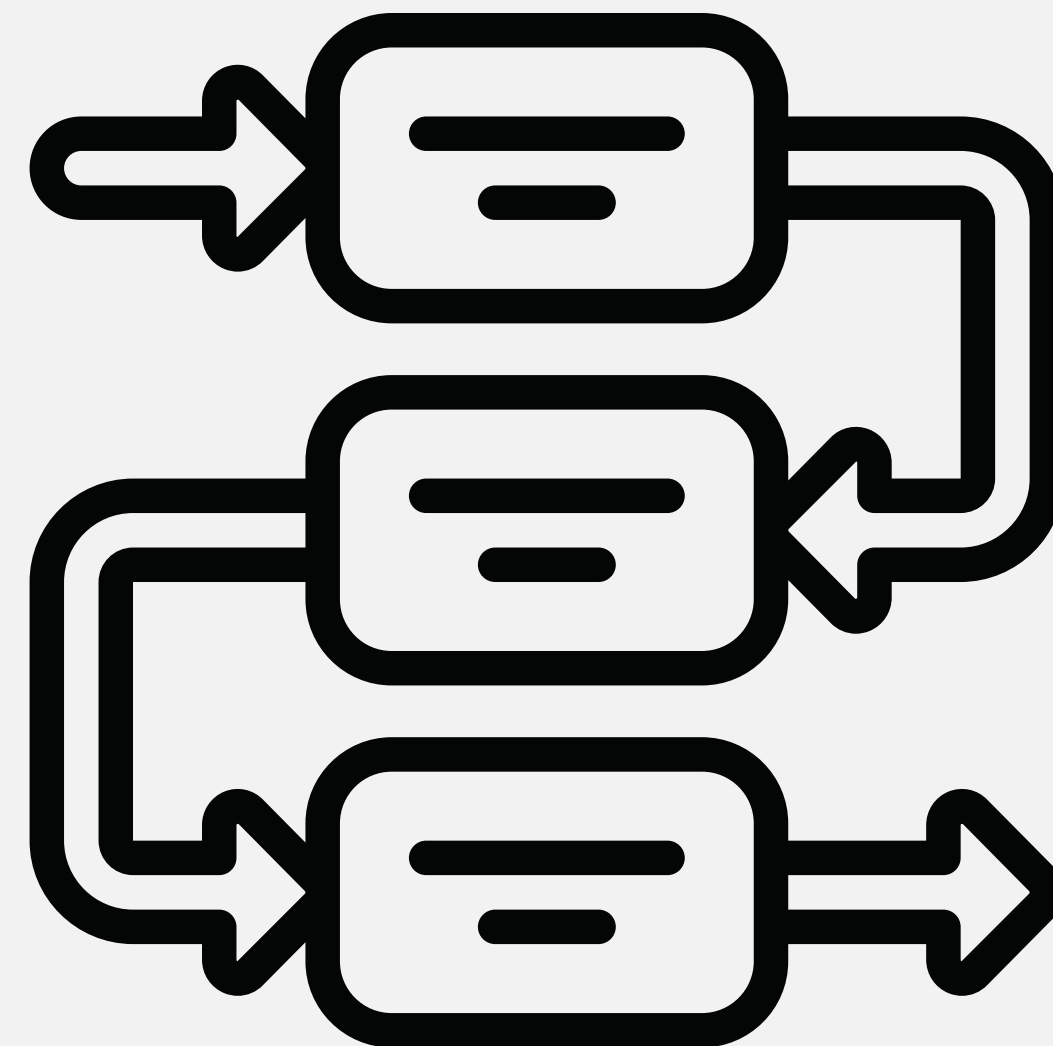
[16] Vikas Paruchuri, & Datalab Team. (2025). Surya: A lightweight document OCR and analysis toolkit.

[17] Purushoth. Velayuthan and Thanuja D. Ambegoda. 2025. Benchmarking OCR Models for Sinhala and Tamil Document Digitization. Technical report, Engineering Research Unit, University of Moratuwa.

[18] <https://github.com/JaidevAI/EasyOCR>

Methodology of the Comparative Analysis

- During the search for suitable datasets for both Sinhala and Tamil, we were unable to find two datasets with comparable characteristics and challenges.
- Therefore, we opted to use synthetically created datasets to ensure a controlled comparison.
- The Sinhala language was assessed using the `sinhala_synthetic_ocr-large` [19], while the Tamil language was evaluated with a synthetically created OCR dataset developed by us.



[19] https://huggingface.co/datasets/Ransaka/sinhala_synthetic_ocr-large

Evaluation Results of OCR for Sinhala and Tamil

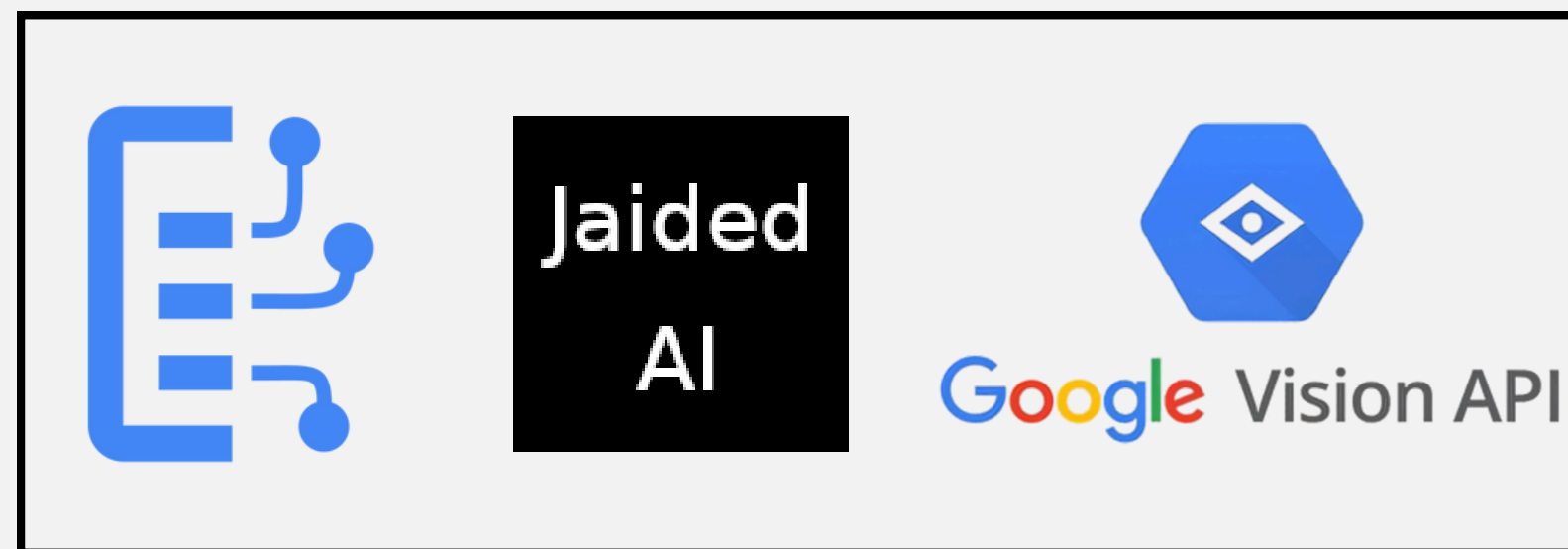
- To effectively evaluate the capabilities of the OCR engines for these two languages, we utilized five different metrics.

OCR System	Language	CER↓	WER↓	BLEU↑	ANLS↑	METEOR↑
Cloud Vision API	Sinhala	0.0619	0.0767	0.91934	0.9447	0.9269
	Tamil	0.0079	0.1204	0.5790	0.9922	0.8751
Surya	Sinhala	0.0076	0.0261	0.9396	0.9920	0.9723
	Tamil	0.1392	0.64999	0.1487	0.8672	0.3359
Document AI	Sinhala	0.0610	0.0758	0.9199	0.9455	0.9278
	Tamil	0.0078	0.1198	0.5803	0.9923	0.8762
Subasa OCR	Sinhala	0.0761	0.1799	0.6894	0.9259	0.8099
	Tamil	-	-	-	-	-
Tesseract	Sinhala	0.0702	0.1489	0.7553	0.9319	0.8436
	Tamil	0.0780	0.6145	0.0493	0.9264	0.3201
EasyOCR	Sinhala	-	-	-	-	-
	Tamil	0.1172	0.2876	0.3461	0.8828	0.6744

The evaluation of OCR systems for the Sinhala and Tamil languages

Conclusion of the Comparative Analysis

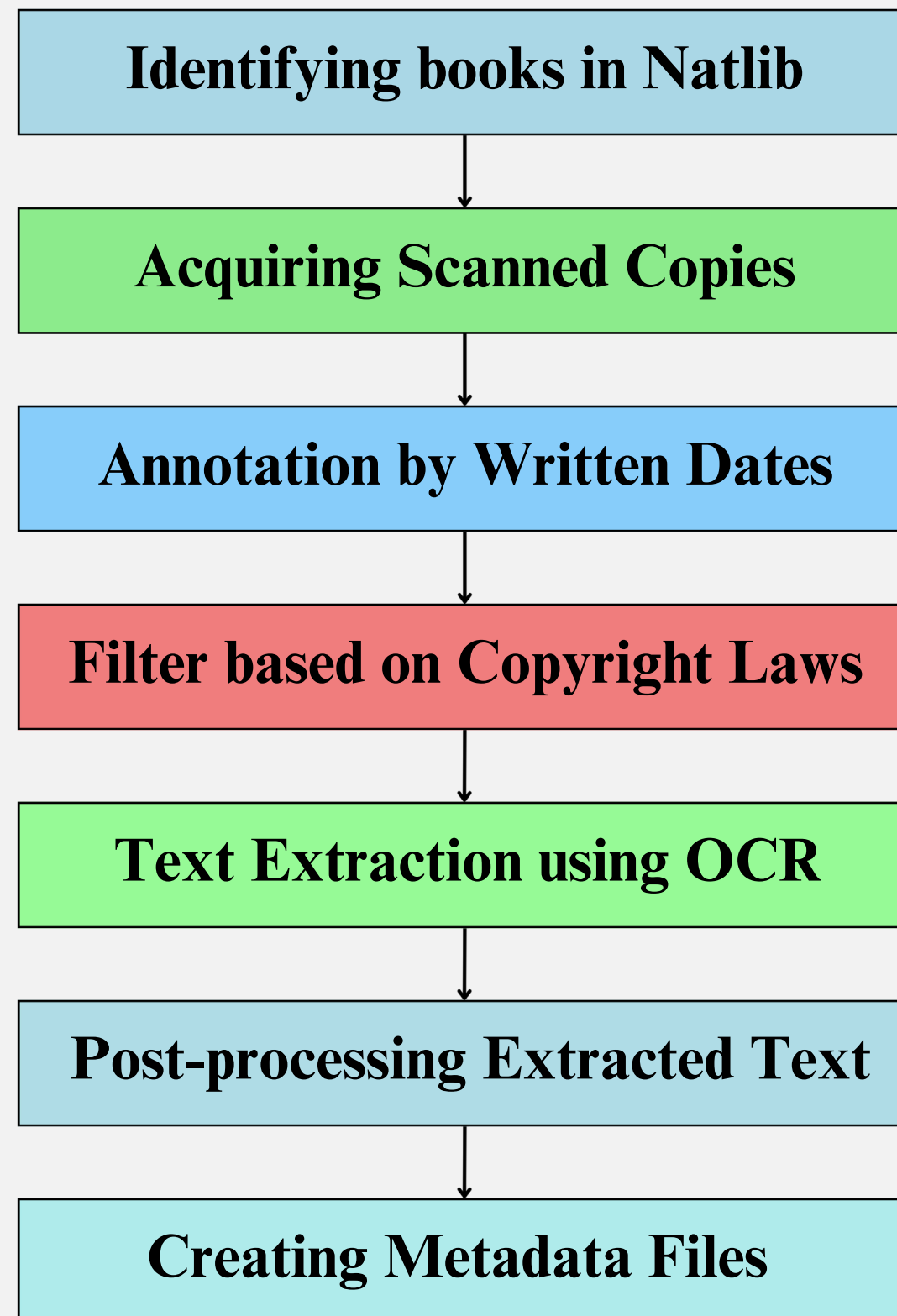
- We identified Surya as the best-performing OCR engine for Sinhala and Document AI as the best for Tamil.
- When evaluating the overall results for both languages, the Document AI and Cloud Vision API have achieved very similar outcomes, ranking first and second, respectively.



- This comparative study was compiled as a research paper and accepted at RANLP 2025 as a long paper [20].



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 comprises 58k words across 46 literary works which comfortably passes the 53,000 token count of FarPaHC for Faroese [21], which is in the same language resource category as Sinhala according to Ranathunga and De Silva (2022) [22].
- It was annotated based on the written date or period of the literature.
- The methodology involved identifying books, followed by several steps to create metadata files, as shown in the figure.

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

[22] 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).

Dataset Assembly

- We started acquiring Sinhala literature, both fiction and non-fiction, from the Internet Archives, but the amount of data was limited.
- As a result, we turned to the National Library (Natlib)¹ of Sri Lanka, the main institution for Sinhala literature, which has a digital repository²
- In the digital repository, we organized all content chronologically by issued date, showcasing publications dating back to 1800 CE.
- We selected the book title, author name, identifier number, and collection name for each book from that time onward.
- We identified 233 unique books printed between 1800 and 1955 in the Natlib digital repository, with only 12 available for open access.
- Our plan was to collect 100 sentences per year, estimating that five pages from each book would yield around 1500 to 2000 words (Assuming there are about 15 to 20 words per sentence)

1. <https://www.natlib.lk/>

2. <https://diglib.natlib.lk/>

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 [23].
- Therefore, a comprehensive analysis was conducted to ensure that the written year of each book was accurately represented.
- Following expert recommendations in Sinhala linguistics, we found a detailed book on Sinhala literature that covers information from Sinhala's inception up until 1994 CE [24].
- The date ranges identified correspond either to the period during which the book was authored or to the time period in which the author lived [25].
- Determining the written dates for the books became complicated due to some including commentaries and discourses on earlier works.
- In this dataset, these cases are linked to the original book's date since they contain information from both the original and its commentary, often quoting extensively from the original text.

[23] Gippert, J. and Tandashvili, M., 2015. Structuring a diachronic corpus. The Georgian National Corpus project. In Historical corpora. Challenges and perspectives (pp. 305-322). Narr.

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

[25] 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 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 2003.³
- 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.

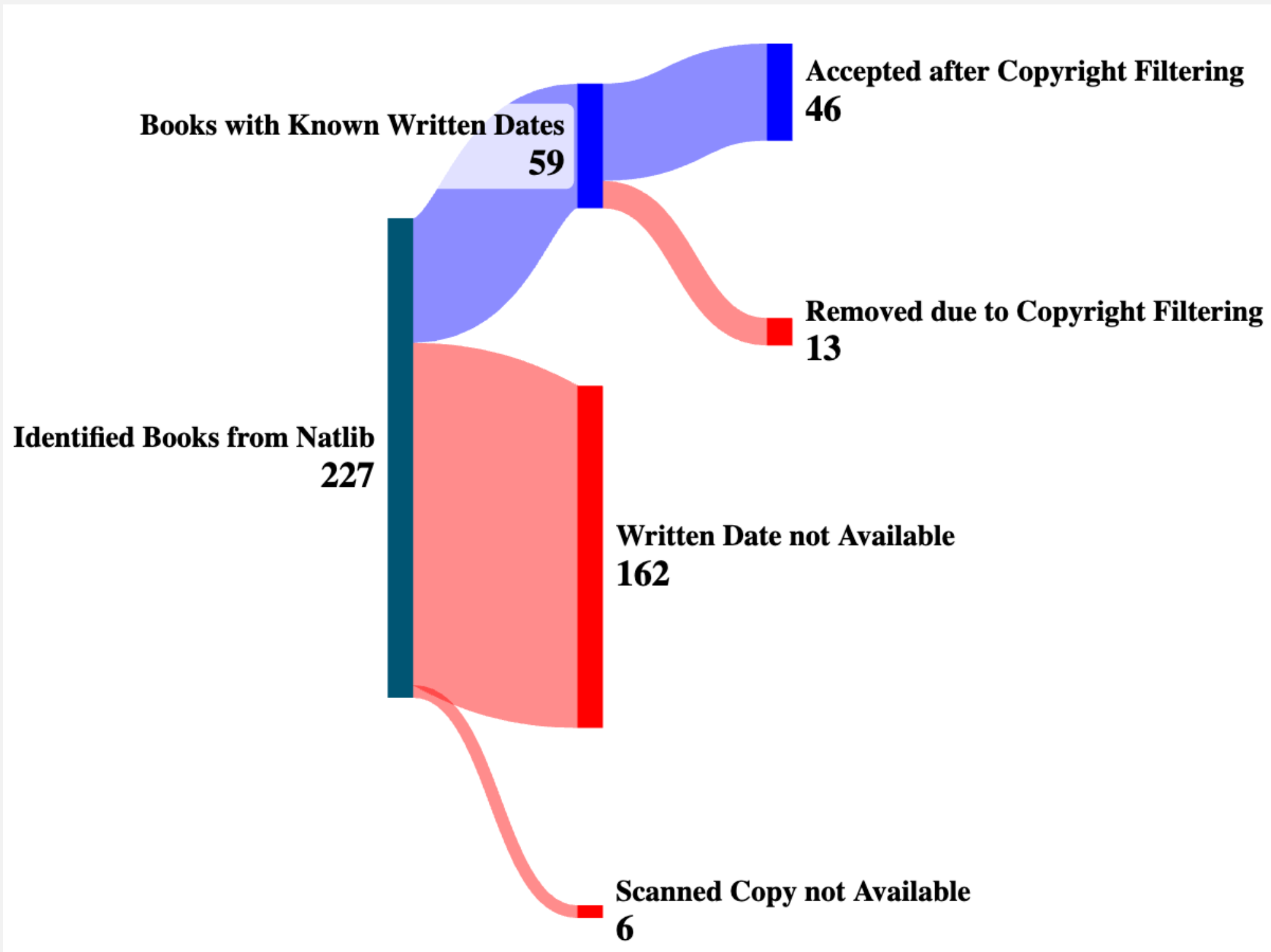
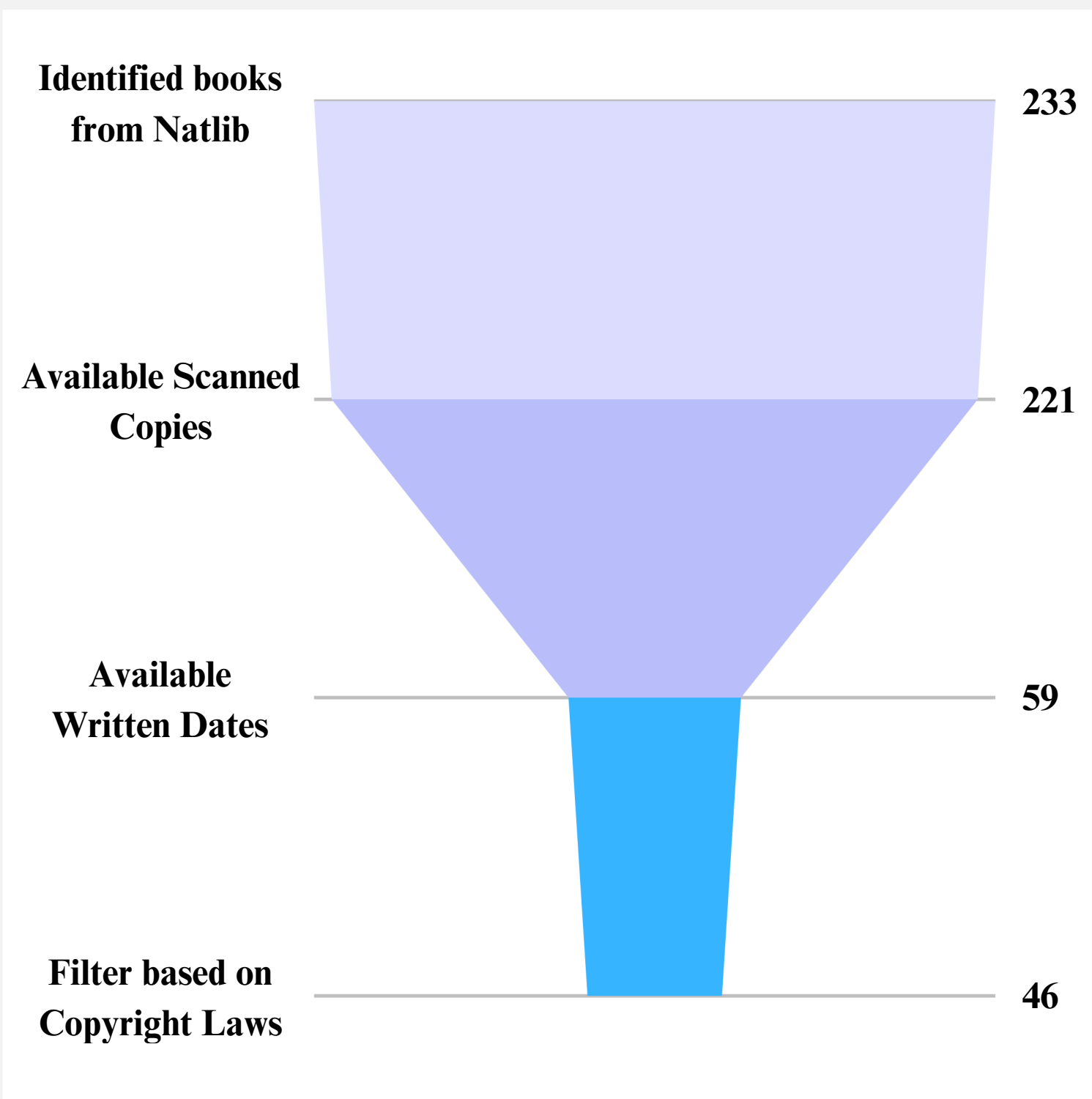


3. <https://www.gov.lk/wordpress/wp-content/uploads/2015/03/IntellectualPropertyActNo.36of2003Sectionsr.pdf>

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.
- 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

Data Filtration Contd.



Text Extraction using OCR

- Our comparative analysis identified Surya as the best OCR engine for Sinhala, with Document AI being the second best.
- However, during our text extraction process, we discovered that under realistic conditions, unlike the synthetic conditions used in the analysis, Google Document AI yielded more accurate results.
- We obtained model confidence for each page of the processed documents and calculated the average score, which is included in the metadata file of each book folder (the OCR accuracy averaged 96.84% across all documents).
- Document AI demonstrated that it can perform text recognition that goes beyond simple extraction.



Text Extraction using OCR Contd.

Text Modernisation

උපමාසීචාවක	→	උපමාර්ථචාවක
භෘතූචලිත	→	භාංගචලිත
සිංහලනාමය	→	සිංද්‍රාංගනාමය

Morpheme Segmentation

නොවන හෙයින්	→	නොවන හෙයින්
සපුමල් පොකුරක්	→	සපුමල් පොකුරක්
ලෙසිදඹුවා	→	ලෙසිදඹුවා

Text Modernisation & Morpheme Segmentation

සවකීයකර්තෘච්ඡනයයි	→	සවකීය කර්තෘ ච්ඡන යයි
පූර්වලොපසන්ධිවුනි	→	පූර්ව ස්වර ලොප සන්ධිවුනි
කලිකාගතිදපයභිවූ	→	කලිකා ගතිද පර්යාස වූ

- The evolution of the Sinhala language has led to different eras of syntax that, while linguistically equivalent, differ in grapheme representation [26].
- Document AI ensures consistency in modern Sinhala syntax across the dataset (as changes occur only at the grapheme level, this does not violate the syntactic or semantic properties).
- Additionally, in Historical Sinhala, words often form closed compounds without spaces [27].
- Document AI effectively identifies this phenomenon and performs morpheme segmentation in this context.

[26] Nandasara, S.T. and Mikami, Y., 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).

[27] Gaikwad, H. and Saini, J.R., 2024, May. Identification of Closed Compound Words in Devanagari Scripted and Non-Devanagari Scripted Corpora. In Doctoral Symposium on Computational Intelligence (pp. 411-418). Singapore: Springer Nature Singapore.

Post-Processing Extracted Text

- Although the OCR accuracy is very high across all documents, the formatting issues were significant enough to require manual adjustments.
- The Document AI's advanced performance streamlined manual post-processing tasks, greatly reducing the time needed for this work.
- 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
 - Page Number Removal
- 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 who are native Sinhala speakers.

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 follow the conventions set by the LatinISE corpus [28].
- 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 work by Sannasgala 2015 [24].

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

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

Creation of Metadata Files Contd.

- The genres of the books were selected based on the details provided by Sannasgala 2015 [24], 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 categorization was inspired by the methodologies followed in IcePaHC [29] and DIAKORP [30] corpora to ensure diverse genres.
- The first level of categorization was applied to all documents, while the second level applied only to selected categories of books.

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

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

[30] 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.

Evaluation of the SiDiaC-v.1.0 Corpus

- 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



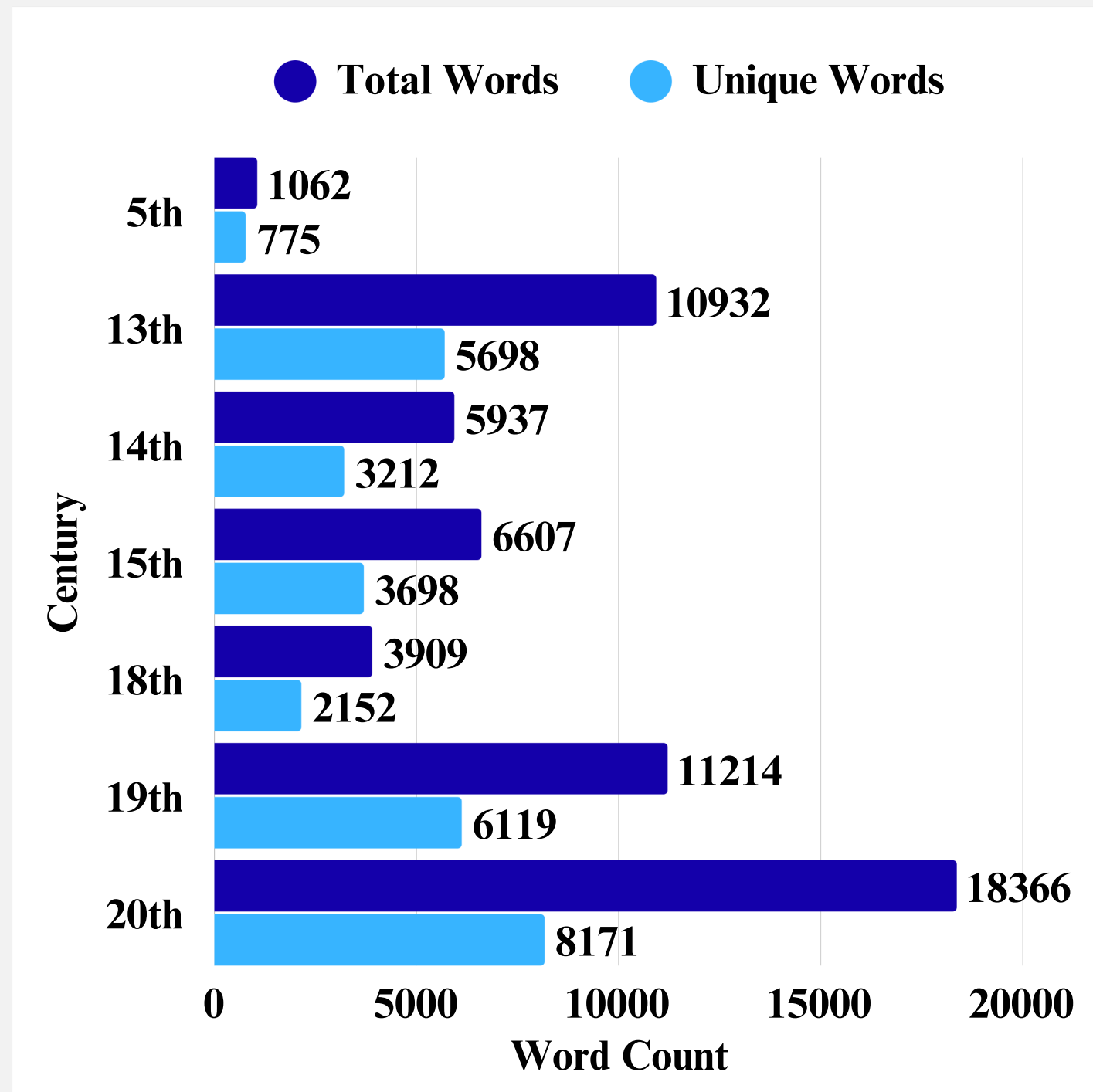
Evaluation of the SiDiaC-v.1.0 Corpus Contd.

	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

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.

Evaluation of the SiDiaC-v.1.0 Corpus Contd.



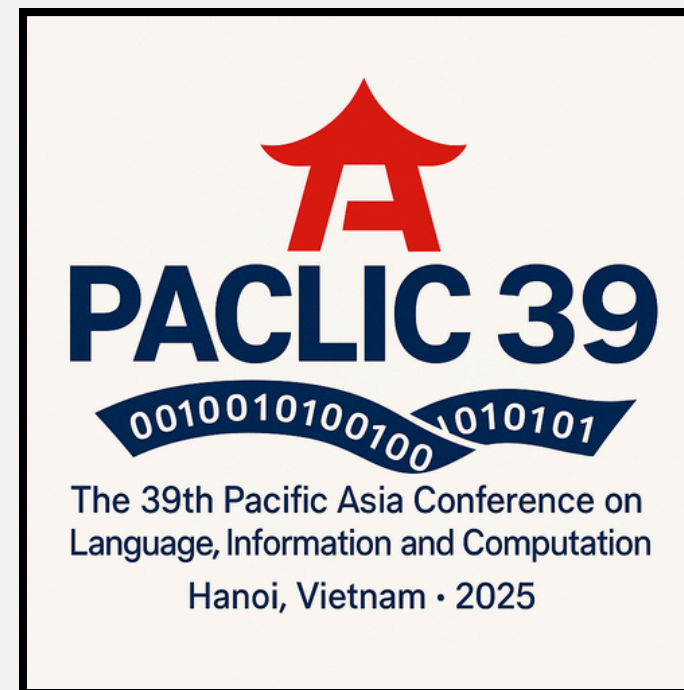
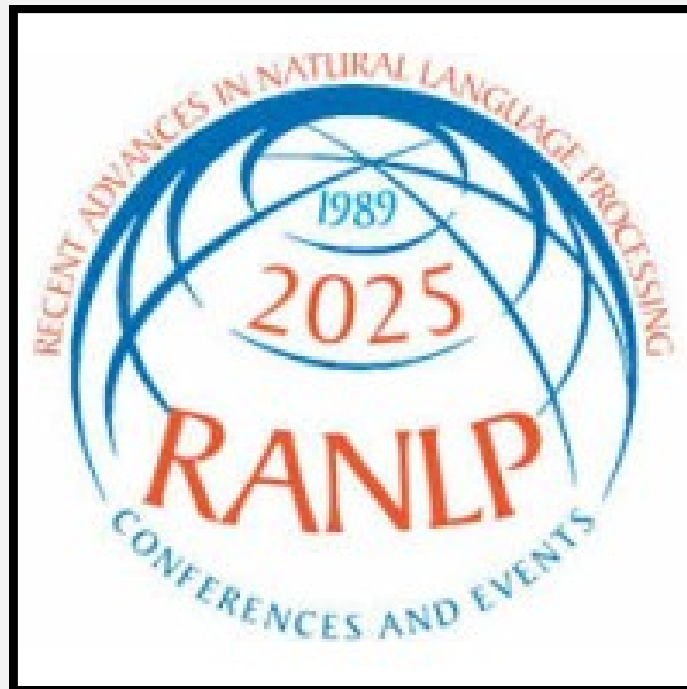
- SiDiaC consists of 58,027 whitespace word tokens.
- The corpus contains 833 words in Latin script, which accounts for just 1.42% of the entire dataset.
- Also, the complete dataset comprises 22,837 unique word tokens in Sinhala script, which accounts for 39.36% unique word coverage of all words.
- This total word token count 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) [21].

Planned Work: Leading to the Next Review

- We plan to continue progressing with the first objective of the study, which is to enhance the Sinhala Diachronic Corpus that has been created.
 - Aim to enhance the number of literary works utilised from 46 to more than 100.
 - Conduct deeper post-processing steps to improve the text quality;
 - Removal of Sanskrit, Pali and English.
 - Address character and word-level recognition errors.
 - More precise written date annotation (eg, books with commentaries)
 - Increase the number of tokens by incorporating additional pages from existing literature and by including new literature.
- We plan to submit the second version of the corpus to LREC 2026, with the submission deadline on 17th October 2025.

Papers and Grant Proposals

- Accepted paper at RANLP 2025: “Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis of Sinhala and Tamil.” <https://arxiv.org/abs/2507.18264>
- Submitted the first version of the corpus titled “SiDiaC: Sinhala Diachronic Corpus” at PACLIC 2025.
- Assisted the supervisor in writing a dataset proposal for the AI4Science workshop in collaboration with NeurIPS 2025.



References

- [1] Keidar, D., Opedal, A., Jin, Z. and Sachan, M., 2022, May. Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1422-1442).
- [2] Hamilton, W.L., Leskovec, J. and Jurafsky, D., 2016, November. 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 (pp. 2116-2121).
- [3] De Silva, N., 2019. Survey on publicly available Sinhala natural language processing tools and research. arXiv preprint arXiv:1906.02358.
- [4] 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).
- [5] Beinborn, L. and Choenni, R., 2020. Semantic Drift in Multilingual Representations. Computational Linguistics, 46(3), pp.571-603.
- [6] Frermann, L. and Lapata, M., 2016. A Bayesian Model of Diachronic Meaning Change. Transactions of the Association for Computational Linguistics, 4, pp.31-45.
- [7] Rosenfeld, A. and Erk, K., 2018, June. Deep neural models of semantic shift. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 474-484).
- [8] Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J. and im Walde, S.S., 2021, August. Lexical Semantic Change Discovery. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 6985-6998).
- [9] Hamilton, W.L., Leskovec, J. and Jurafsky, D., 2016, August. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1489-1501).
- [10] Periti, F., Dubossarsky, H. and Tahmasebi, N., 2024, March. (Chat) GPT v BERT Dawn of Justice for Semantic Change Detection. In Findings of the Association for Computational Linguistics: EACL 2024 (pp. 420-436).
- [11] McGillivray, B. and Kilgarrieff, A., 2013. Tools for historical corpus research, and a corpus of Latin. New methods in historical corpus linguistics, 1(3), pp.247-257.
- [12] LTRL UCSC, “Sinhala ocr,” 2022. [Online]. Available: <https://ocr.subasa.lk>
- [13] Tesseract OCR. (n.d.). Tesseract documentation. [online] Available at: <https://tesseract-ocr.github.io/>.
- [14] 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].
- [15] Google Cloud. (2019). Detect Text (OCR) | Cloud Vision API Documentation | Google Cloud. [online] Available at: <https://cloud.google.com/vision/docs/ocr>.
- [16] Vikas Paruchuri, & Datalab Team. (2025). Surya: A lightweight document OCR and analysis toolkit.
- [17] Purushoth. Velayuthan and Thanuja D. Ambegoda. 2025. Benchmarking OCR Models for Sinhala and Tamil Document Digitization. Technical report, Engineering Research Unit, University of Moratuwa.
- [18] <https://github.com/JaidedAI/EasyOCR>
- [19] https://huggingface.co/datasets/Ransaka/sinhala_synthetic_ocr-large
- [20] Jayatilleke, N. and de Silva, N., 2025. Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil. arXiv preprint arXiv:2507.18264.
- [21] Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F. and Wallenberg, J., 2012, May. The Icelandic Parsed Historical Corpus (IcePaHC). In LREC (pp. 1977-1984).
- [22] 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] Gippert, J. and Tandashvili, M., 2015. Structuring a diachronic corpus. The Georgian National Corpus project. In Historical corpora. Challenges and perspectives (pp. 305-322). Narr.
- [24] Sannasgala, P., 1961. Sinhala Sahithya Wanshaya. Colombo: Lake House Book.
- [25] Davies, M., 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. Corpora, 7(2), pp.121-157.
- [26] Nandasara, S.T. and Mikami, Y., 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).
- [27] Gaikwad, H. and Saini, J.R., 2024, May. Identification of Closed Compound Words in Devanagari Scripted and Non-Devanagari Scripted Corpora. In Doctoral Symposium on Computational Intelligence (pp. 411-418). Singapore: Springer Nature Singapore.
- [28] McGillivray, B. and Kilgarrieff, A., 2013. Tools for historical corpus research, and a corpus of Latin. New methods in historical corpus linguistics, 1(3), pp.247-257.
- [29] 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.
- [30] Wijeratne, Y. and de Silva, N., 2020. Sinhala language corpora and stopwords from a decade of sri lankan facebook. arXiv preprint arXiv:2007.07884.

Thank You!

Presented by Nevidu Jayatilleke