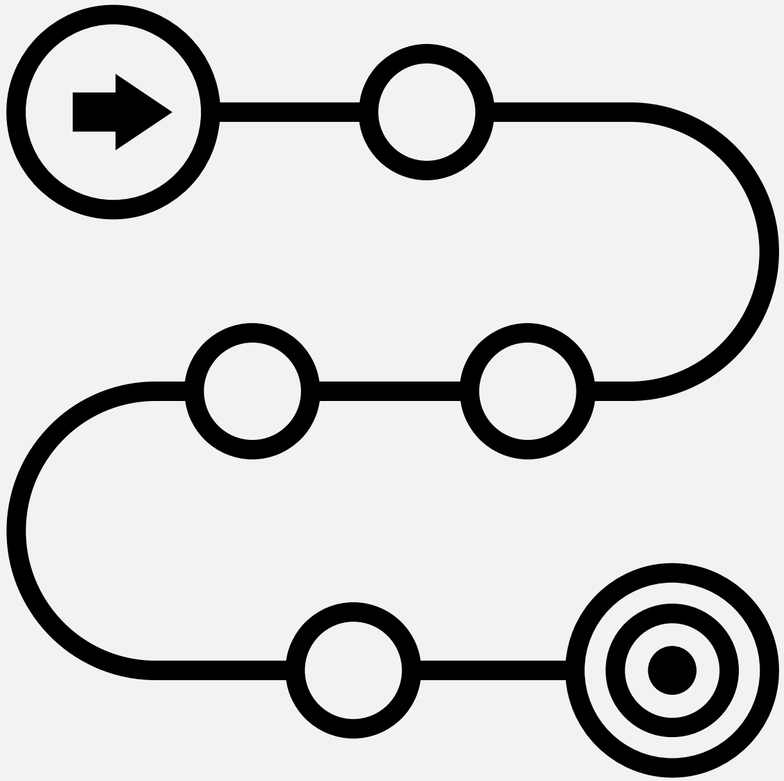
The Dynamics of Meaning: Towards the Evaluation of Diachronic Semantic Drift in Sinhala Language

Nevidu Jayatilleke

Supervised By Dr. Nisansa de Silva

Agenda



- Introduction
  - What is Semantic Drift?
  - Diachronic Semantic Drift
  - Diachronic Semantic Drift: Current Perspectives
- Research Objectives
- Proposed Methodology
  - Data Acquisition Plan
  - Dataset Assembly Strategy
  - Dataset Overview
  - Timeline of Dataset Publication
  - Digitization of Literature using OCR
- Current Findings
  - Evaluation of OCR Engines for Sinhala and Tamil
  - Evaluation Results of OCR for Sinhala and Tamil
- Challenges in the Studies of Lexical Semantic Change
- Resources
- 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.

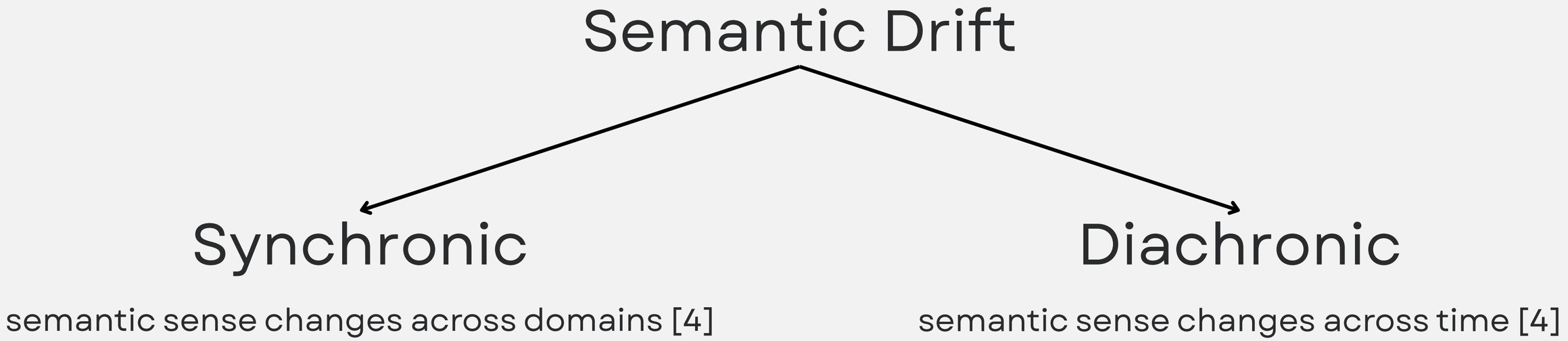
[1] Keidar, D., Opedal, A., Jin, Z. and Sachan, M., 2022, May. Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1422-1442).

[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.

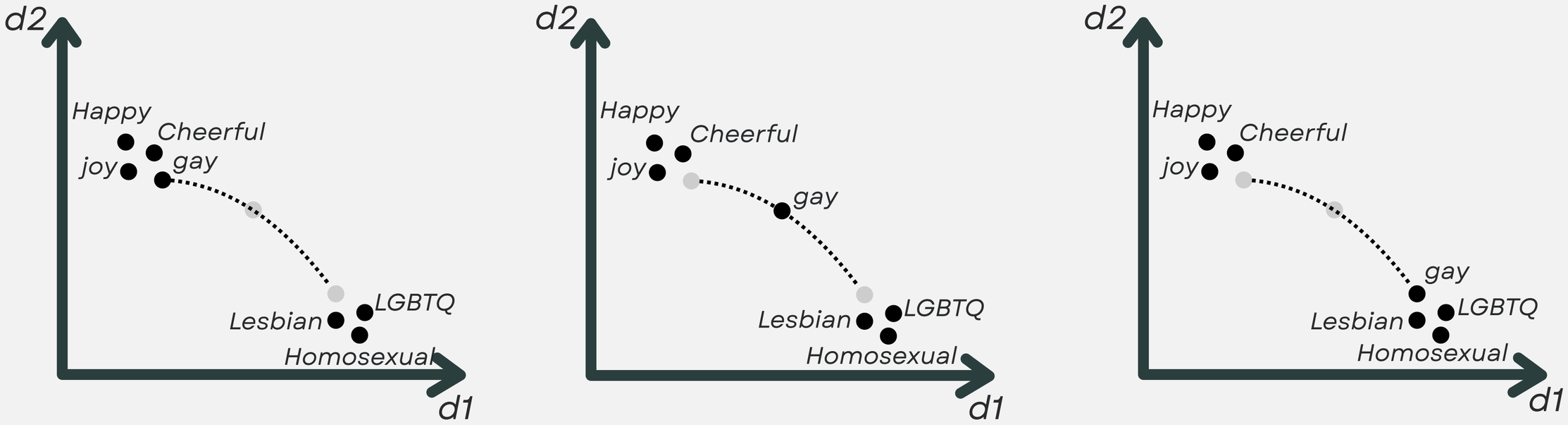


[4] Schlechtweg, D., Hätty, A., Del Tredici, M. and im Walde, S.S., 2019, July. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 732-746).

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]



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

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., 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).
[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.

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.

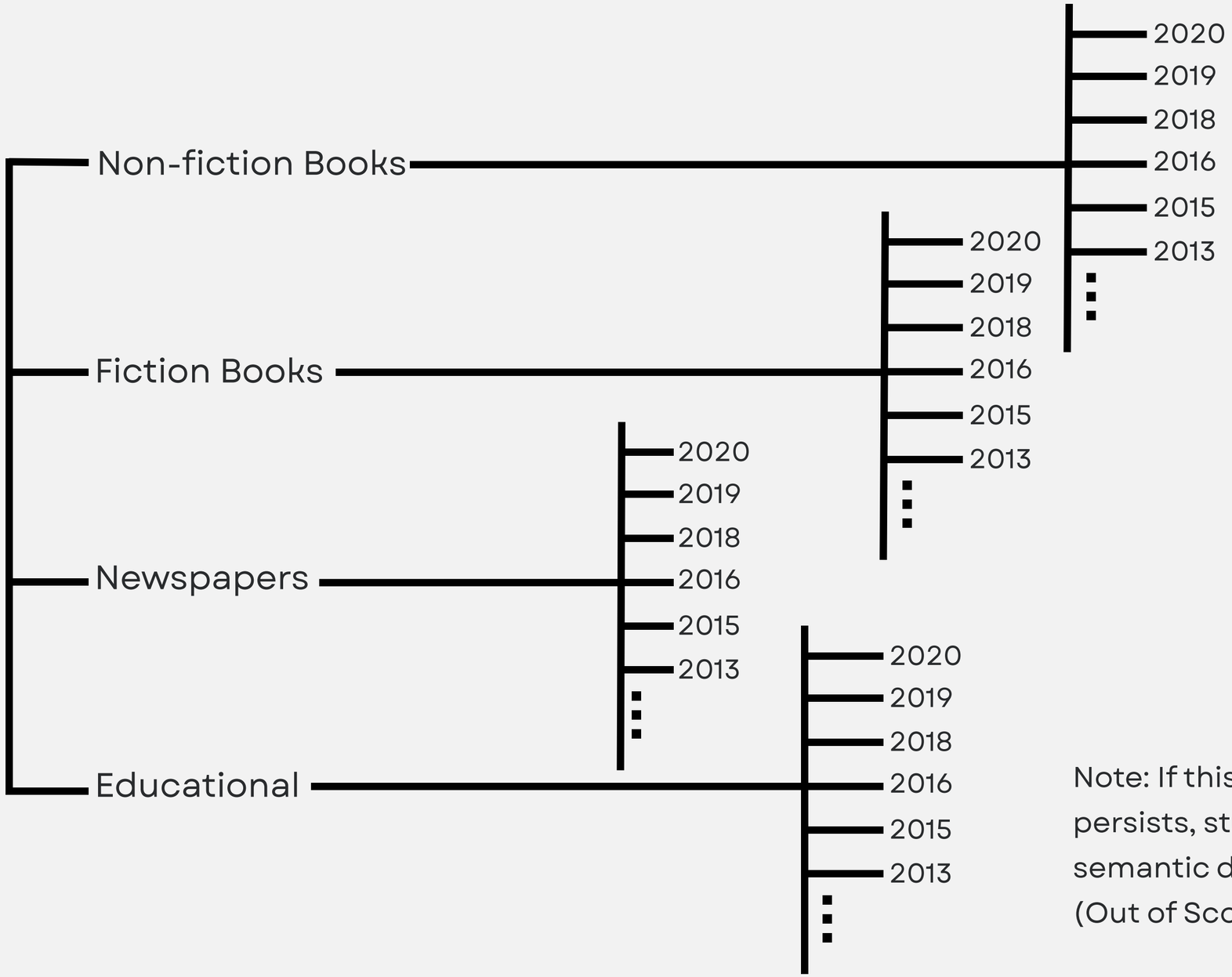
Data Acquisition Plan

- We have already 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

Dataset Assembly Strategy

- The dataset will consist of yearly data.
- The aim is 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 is 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

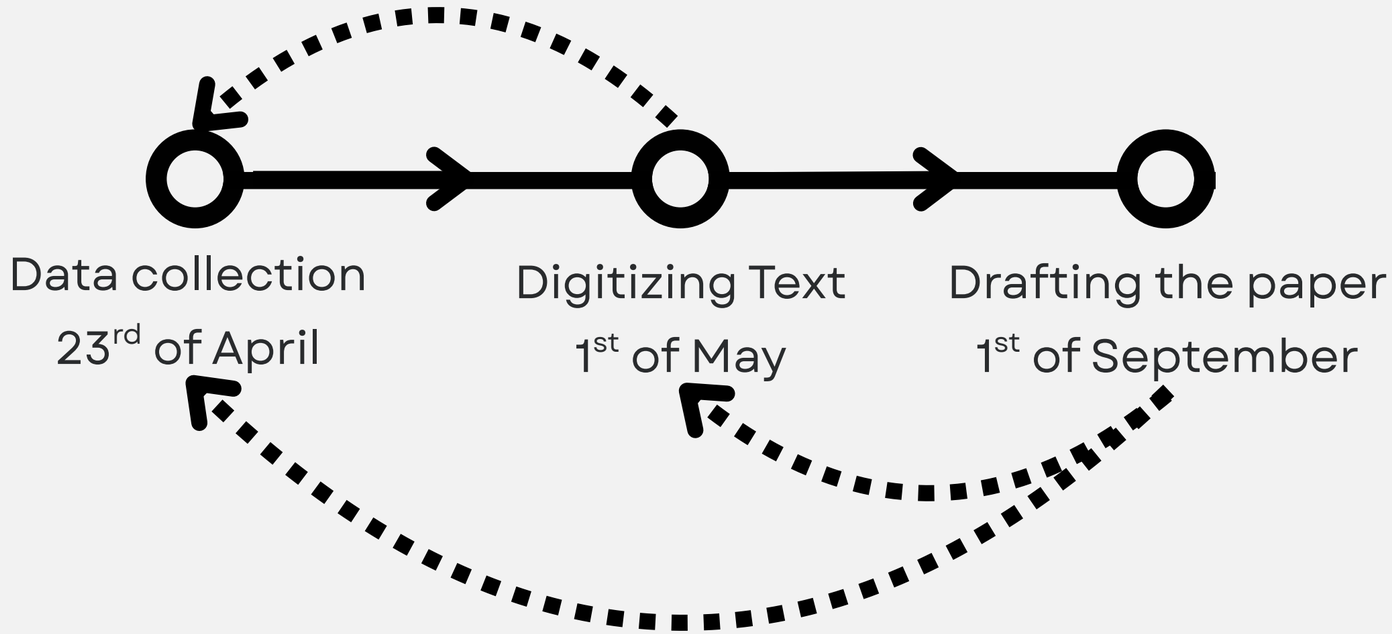
Dataset Overview



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

Timeline of Dataset Publication

- Our goal is to publish the base dataset for LREC 2026, where we expect the paper deadlines to be by the end of October 2025.
- Thus, the entire initial dataset creation process, including the accompanying paper, should take about six months.



Digitization of Literature using OCR

- The majority of our literature resources will primarily be in physical format, as we are focusing on Sinhala texts that span several centuries.
- Therefore, digitizing the identified sources will be a crucial task for dataset creation.
- We have identified several systems to conduct Optical Character Recognition (OCR) for Sinhala text;
  - Subasa OCR [11]
  - Tesseract OCR [12]
  - OCR with Document AI [13]
  - Google Cloud Vision API [14]
  - Surya OCR [15]
- The project's scope does not include developing OCR systems; however, we will use existing systems and modify them further if necessary to digitize the literature.

---

[11] LTRL UCSC, “Sinhala ocr,” 2022. [Online]. Available: https://ocr.subasa.lk
[12] Tesseract OCR. (n.d.). Tesseract documentation. [online] Available at: https://tesseract-ocr.github.io/.
[13] 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].
[14] Google Cloud. (2019). Detect Text (OCR) | Cloud Vision API Documentation | Google Cloud. [online] Available at: https://cloud.google.com/vision/docs/ocr.
[15] Vikas Paruchuri, & Datalab Team. (2025). Surya: A lightweight document OCR and analysis toolkit.

Evaluation of OCR Engines for Sinhala and Tamil

- We evaluated six different OCR engines for both Sinhala and Tamil languages;
  - Subasa OCR [11]
  - Tesseract OCR [12]
  - OCR with Document AI [13]
  - Google Cloud Vision API [14]
  - Surya OCR [15]
  - Easy OCR [16]
- The Sinhala language was assessed using the sinhala_synthetic_ocr_large [17], while the Tamil language was evaluated with a synthetically created OCR dataset developed by us.
- To effectively evaluate the capabilities of the OCR engines for these two languages, we utilized five different metrics.

[16] https://github.com/JaidedAI/EasyOCR
[17] https://huggingface.co/datasets/Ransaka/sinhala_synthetic_ocr-large

Evaluation Results of OCR for Sinhala and Tamil

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

Table 1: The evaluation of OCR systems for the Sinhala and Tamil languages

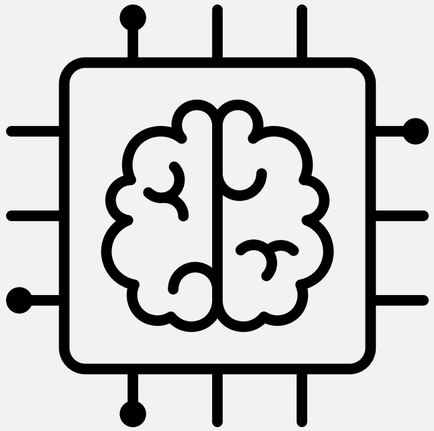
Challenges in the studies of Lexical Semantic Change

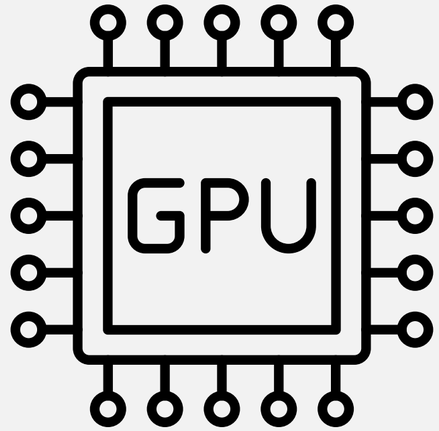
- Many existing semantic change detection methods using contextual embeddings are **unscalable in terms of memory and computation time** [17].
- Vector space alignment is a technical challenge in diachronic LSC, as embeddings from different time periods need to be compared in a common space [16].
- Distinguishing between **semantic change and other linguistic changes**, such as syntactic changes, can be challenging [15].
- Dataset creation will be extremely difficult, from acquiring literature to digitizing and formatting them.

[15] Periti, F. and Montanelli, S., 2024. Lexical semantic change through large language models: a survey. ACM Computing Surveys, 56(11), pp.1-38.
[16] Timmermans, M., Vanmassenhove, E. and Shterionov, D., 2022, May. “Vaderland”,“Volk” and “Natie”: Semantic Change Related to Nationalism in Dutch Literature Between 1700 and 1880 Captured with Dynamic Bernoulli Word Embeddings. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change (pp. 125-130).
[17] Montariol, S., Martinc, M. and Pivovarova, L., 2021, June. Scalable and interpretable semantic change detection. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4642-4652).

Resources

- **Models:** Pre-trained models will be sourced from HuggingFace to generate contextual embeddings.
- **Data:** The dataset will be assembled to conduct diachronic studies.
- **Computational Resources:** GPU cluster available in CSE department, UOM.





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] LTRL UCSC, “Sinhala ocr,” 2022. [Online]. Available: https://ocr.subasa.lk

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

[13] 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].

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

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

[16] https://github.com/JaidedAI/EasyOCR

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

[18] Periti, F. and Montanelli, S., 2024. Lexical semantic change through large language models: a survey. ACM Computing Surveys, 56(11), pp.1-38.

[19] Timmermans, M., Vanmassenhove, E. and Shterionov, D., 2022, May. “Vaderland”,“Volk” and “Natie”: Semantic Change Related to Nationalism in Dutch Literature Between 1700 and 1880 Captured with Dynamic Bernoulli Word Embeddings. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change (pp. 125-130).

[20] Montariol, S., Martinc, M. and Pivovarova, L., 2021, June. Scalable and interpretable semantic change detection. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4642-4652).

Thank You!

Presented by Nevidu Jayatilleke