# Sinhala Diachronic Corpus

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

## Abstract

We propose to build a comprehensive diachronic corpus for the Sinhala language. Sinhala is the native language of Sri Lanka, where it has official language status. This corpus is aimed at addressing the notable gap in available linguistic resources for the language, which has a rich literary heritage yet remains under-represented in research. This work would be beneficial mainly to the Sinhala-speaking community, which numbers approximately 16 million people, clearly denoting an underserved minority in the global language ecosystem. Further, researchers focusing on the historical evolution of other related languages, such as Dhivehi, Marathi, Sanskrit, and Hindi, may use this corpus as an auxiliary source in their work. We hope to compile a diverse dataset that spans various time periods and genres, drawing from a wide array of sources, including online content, books, articles, and newspapers. Given that Sinhala has gone through a significant number of evolutionary eras since its origins in the 3rd to 2nd centuries BCE, we intend to collaborate with Sri Lankan institutions and language scholars to annotate texts based on their original writing dates, as opposed to publication dates. Ultimately, following its open access release, this corpus will not only advance the understanding of the linguistic evolution of Sinhala but also contribute to the preservation of its literary legacy.

## 1 Introduction

Sinhala is a Low-Resource language [1] almost exclusively used in the island of Sri Lanka [2]. It is reported to be the mother tongue (L1) of approximately 16 million people [3]. The low-resourced status of the languages falls short at representing the rich and diverse literary heritage that has developed for this language over the course of several millennia, with its origins tracing back to between the 3rd and 2nd centuries BCE [3]. During this period, throughout its history, it has undergone significant evolution and transformation, resulting in the form of modern Sinhala that we engage with today.

Language evolution plays a vital role in enabling humans to adapt to societal changes across generations. As we explore the fascinating process of language evolution, we find that words can shift in meaning for a variety of reasons, including cultural influences such As technological advancements, as well as regular linguistic phenomena such as subjectification [4]. 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 the relative position of the word within the embedding space. 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" in the English language, of which the most frequently used meaning has shifted from *cheerful* to *homosexual* over the years [5].

We observe an obvious gap in diachronic resources for Sinhala. In fact, even contemporary Sinhala dictionary data sets are scarce [6, 7]. So are traditional Language Models (LMs) [8–10] and pre-trained Large Language Models (LLMs) [11, 12].

## 2 Relation to Prior Work in the Area

McGillivray and Kilgarriff [13] introduced LatinlSE, a 13-million-word historical Latin corpus developed for the Sketch Engine[1]. The relevance of this work to ours lies in the covered timespan, which goes beyond a millennium. Another relevant resource comes coupled with the *Icelandic Parsed Historical Corpus* (IcePaHC) [14], a one-million-word parsed historical corpus of Icelandic that spans from the late 12th century to the early 21st century. Here, more pertinent to our study is the *spin-off*, *Faroese Parsed Historical Corpus* (FarPaHC), a syntactically annotated collection of Faroese historical texts. The relevance here rises from the fact that according to Ranathunga and De Silva [1], Faroese belongs to the *language resource availability category 02*, similar to Sinhala. The work by Keersmaekers and Van Hal [15] presents a case study demonstrating how large-scale automated parsing of Greek papyri can produce richly annotated diachronic resources. Additionally, Chen and Liu [16] have developed a Chinese corpus that consists of news articles on land usage from the past 30 years. Even the corpora in higher-resourced languages, such as DIAKORP [17] (Czech), ARCHER [18], and COHA [19][2] (English), as well as DTA [20] and GerManC [21] (German), vary in terms of size, balance, annotation depth, and access models.

## 3 Dataset Objectives

We have outlined relevant objectives that need to be achieved to build a comprehensive diachronic corpus of the Sinhala language:

1. Systematically identify, acquire, and filter Sinhala literary works from the National Library of Sri Lanka (and other sources), ensuring compliance with copyright laws and historical relevance, on which extensive manual post-processing is performed.
2. Accurately establish the written dates of the selected texts (distinct from issue dates), with the help of language sources, anchoring each work in its correct historical period for a valid diachronic corpus.
3. Build a comprehensive diachronic corpus of Sinhala literature, providing a structured dataset that captures the linguistic evolution of Sinhala across different historical eras.
4. Identify semantics that show significant drift diachronically, as well as those that remain constant in the embedding space, which can be considered as anchor or pivot words in further investigations.

## 4 Proposed Methodology

***O1:*** Sinhala literary works will be systematically sourced from digital repositories. For physical resources, such as hard copies, identified materials will be digitised using existing OCR systems for Sinhala [22–24], with new models developed if necessary. Only works meeting copyright compliance [25] and historical relevance will be retained. The OCR processing will be followed by extensive manual post-editing by native speakers to ensure accuracy.

***O2:*** Since issue dates (after the introduction of the printing press in Sri Lanka in 1737 [26, 27]) may not reflect original authorship, written dates will be established using Sannasgala [28] and other linguistic references. Works containing commentaries will be anchored to the original texts rather than later interpretations, ensuring accurate historical placement.

***O3:*** The final corpus will contain at least 53,000 tokens, comparable to FarPaHC [14]. Metadata, including title, author, dates, genre, and OCR confidence, will be compiled, while texts will be categorised into Fiction/Non-Fiction and five subgenres [14, 17]. Orthography will be modernised [27] and morphemes segmented [29].

***O4:*** The corpus will be used to train diachronic word embeddings across historical periods [5]. Semantic change will be measured through vector distance comparisons across time slices. Words showing substantial drift will be identified, while stable words will serve as anchors or pivots for further investigations into lexical evolution, cultural influences, and semantic stability in Sinhala.

[1] https://www.sketchengine.eu/

[2] https://www.english-corpora.org/coha/

## References


[1] S. Ranathunga and N. De Silva, "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)*, 2022, pp. 823–848.

[2] J. Dunn and L. Edwards-Brown, "Geographically-Informed Language Identification," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 7672–7682.

[3] N. de Silva, "Survey on Publicly Available Sinhala Natural Language Processing Tools and Research," *arXiv preprint arXiv:1906.02358v25*, 2025.

[4] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Cultural shift or linguistic drift? comparing two computational measures of semantic change," in *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, vol. 2016. NIH Public Access, 2016, p. 2116.

[5] L. Beinborn and R. Choenni, "Semantic drift in multilingual representations," *Computational Linguistics*, vol. 46, no. 3, pp. 571–603, 2020.

[6] K. Wickramasinghe and N. De Silva, "Sinhala-English Parallel Word Dictionary Dataset," in *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2023, pp. 61–66.

[7] A. Wasala and R. Weerasinghe, "Ensitip: a tool to unlock the english web," in *11th international conference on humans and computers, Nagaoka University of Technology, Japan*, 2008, pp. 20–23.

[8] V. Dhananjaya, P. Demotte, S. Ranathunga, and S. Jayasena, "BERTifying Sinhala - A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification," in *Proceedings of the 13th language resources and evaluation conference*, 2022.

[9] D. Gurgurov, R. Kumar, and S. Ostermann, "LowREm: A Repository of Word Embeddings for 87 Low-Resource Languages Enhanced with Multilingual Graph Knowledge," *arXiv preprint arXiv:2409.18193*, 2024.

[10] B. Gamage, R. Pushpananda, and R. Weerasinghe, "The impact of using pre-trained word embeddings in sinhala chatbots," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2020, pp. 161–165.

[11] H. W. K. Aravinda, R. Sirajudeen, S. Karunathilake, N. de Silva, S. Ranathunga, and R. Kaur, "SinLlama-A Large Language Model for Sinhala," *arXiv preprint arXiv:2508.09115*, 2025.

[12] A. Üstün, V. Aryabumi, Z.-X. Yong, W.-Y. Ko, D. D'souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, F. Vargus, P. Blunsom, S. Longpre, N. Muennighoff, M. Fadaee, J. Kreutzer, and S. Hooker, "Aya model: An instruction finetuned open-access multilingual language model," *arXiv preprint arXiv:2402.07827*, 2024.

[13] B. McGillivray and A. Kilgarriff, "Tools for historical corpus research, and a corpus of latin," *New methods in historical corpus linguistics*, vol. 1, no. 3, pp. 247–257, 2013.

[14] E. Rögnvaldsson, A. K. Ingason, E. F. Sigurðsson, and J. Wallenberg, "The icelandic parsed historical corpus (icepahc)." in *LREC*, 2012, pp. 1977–1984.

[15] A. Keersmaekers and T. Van Hal, "Creating a large-scale diachronic corpus resource: Automated parsing in the Greek papyri (and beyond)," *Natural Language Engineering*, vol. 30, no. 5, pp. 1035–1064, 2024.

[16] C. Chen and R. Liu, "How administrative powers have impacted land-use development in China during the last 30 years: A diachronic corpus-based news values analysis," *Cities*, vol. 159, p. 105786, 2025.

[17] K. Kučera, A. Řehořková, and M. Stluka, "DIAKORP: diachronic corpus of Czech, version 6," *Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague*, 2015. [Online]. Available: http://www.korpus.cz/

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

[19] M. Davies, "Expanding horizons in historical linguistics with the 400-million word corpus of historical american english," *Corpora*, vol. 7, no. 2, pp. 121–157, 2012.

[20] A. Geyken, S. Haaf, B. Jurish, M. Schulz, J. Steinmann, C. Thomas, and F. Wiegand, "Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv," *Digitale Wissenschaft*, vol. 157, 2011.

[21] S. Scheible, R. J. Whitt, M. Durrell, and P. Bennett, "A gold standard corpus of early Modern German," in *Proceedings of the 5th Linguistic Annotation Workshop*, N. Ide, A. Meyers, S. Pradhan, and K. Tomanek, Eds. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 124–128. [Online]. Available: https://aclanthology.org/W11-0415/

[22] I. Anuradha, C. Liyanage, and R. Weerasinghe, "Estimating the Effects of Text Genre, Image Resolution and Algorithmic Complexity needed for Sinhala Optical Character Recognition," *International Journal on Advances in ICT for Emerging Regions (ICTer)*, vol. 14, no. 3, 2021.

[23] P. Velayuthan and T. D. Ambegoda, "Benchmarking OCR Models for Sinhala and Tamil Document Digitization," Engineering Research Unit, University of Moratuwa, Tech. Rep., 2025.

[24] N. Jayatilleke and N. de Silva, "Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil," *arXiv preprint arXiv:2507.18264*, 2025.

[25] Parliament of Democratic Socialist Republic of Sri Lanka, "Intellectual Property Act No. 36," November 2003, [Online; accessed 2025-08-25]. [Online]. Available: https://www.gov.lk/wordpress/wp-content/uploads/2015/03/IntellectualPropertyActNo.36of2003Sectionsr.pdf

[26] S. Wickremasuriya, "The beginnings of the sinhalese printing press," in *Senarat Paranavitana commemoration volume*. Brill, 1978, pp. 283–300.

[27] S. T. Nandasara and Y. Mikami, "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)*, vol. 8, no. 1, 2016.

[28] P. Sannasgala, *Sinhala Sahithya Wanshaya*. As. Godage saha Sahodarayo, 2015.

[29] H. Gaikwad and J. R. Saini, "Identification of closed compound words in devanagari scripted and non-devanagari scripted corpora," in *Proceedings of Fifth Doctoral Symposium on Computational Intelligence*, A. Swaroop, V. Kansal, G. Fortino, and A. E. Hassanien, Eds. Singapore: Springer Nature Singapore, 2024, pp. 411–418.

[30] J. B. Disanayake, *National Languages of Sri-Lanka: Sinhala*. Department of cultural affairs, 1976.

[31] T. G. Perera, *Sinhala Bhashawa*. M D Gunasena (Sri Lanka), 1985.

[32] L. Bauer, *Linguistics Student's Handbook*. Edinburgh University Press, 2007.

[33] Department of Census and Statistics Sri Lanka. Percentage of population aged 10 years and over in major ethnic groups by district and ability to speak sinhala, tamil and english languages. [Online]. Available: https://goo.gl/nnVZSd

[34] J. W. Gair and W. Karunatilaka, *Literary Sinhala.* ERIC, Cornell University. New York, 1974.

[35] Department of Census and Statistics, Sri Lanka. (2012) Census of Population and Housing of Sri Lanka. [Online]. Available: https://bit.ly/3bAgcXE

[36] Parliament of Democratic Socialist Republic of Sri Lanka, "The constitution of the democratic socialist republic of sri lanka," October 2022, [Online; accessed 2025-08-26]. [Online]. Available: https://www.parliament.lk/files/pdf/constitution.pdf

[37] H. Young. A language family tree - in pictures | education | the guardian. [Online]. Available: https://www.theguardian.com/education/gallery/2015/jan/23/a-language-family-tree-in-pictures

[38] A. B. Kanduboda, "The role of animacy in determining noun phrase cases in the sinhalese and japanese languages," *Science of words*, vol. 24, pp. 5–20, 2011.

[39] R. Arangala, "Location of the Sinhala in Regional Linguistic Historicity and the Identity of Sinhala Language," *Journal of Desk Research Review and Analysis*, vol. 2, no. 1, 2024.

[40] P. E. E. Fernando, "Palaeographical Development of the Brahmi Script in Ceylon from 3rd Century BC to 7th Century AD," *University of Ceylon Review*, vol. 7, no. 4, pp. 282–301, 1949.

[41] D. Bandara, N. Warnajith, A. Minato, and S. Ozawa, "Creation of precise alphabet fonts of early brahmi script from photographic data of ancient sri lankan inscriptions," *Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition*, vol. 3, no. 3, pp. 33–39, 2012.

[42] P. T. Daniels and W. Bright, *The world's writing systems*. Oxford University Press on Demand, 1996.

[43] M. H. Sirisoma, "Brahmi inscriptions of sri lanka from 3rd century bc to 65 ad," pp. 3–54, 1990.

[44] M. Dias, "Lakdiwa sellipiwalin heliwana sinhala bhashawe prathyartha namayange vikashanaya," *Department of Archaeology, Colombo Sri Lanka*, p. 1, 1996.

[45] A. S. Hettiarachchi, "Investigation of 2nd, 3rd and 4th century inscriptions," *Inscriptions: Volume Two, Archaeological Department Centenary (1890–1990), Commemorative Series. Colombo: Department of Archaeology*, pp. 57–104, 1990.

[46] S. Paranavitana and S. L. P. Depārtamēntuva, *Inscriptions of Ceylon*. Dept. of Archaeology, 1970.

[47] B. Hettige and A. S. Karunananda, "Computational model of grammar for english to sinhala machine translation," in *Advances in ICT for Emerging Regions (ICTer), 2011 International Conference on*. IEEE, 2011, pp. 26–31.

[48] A. M. Gunasekara, *A Comprehensive Grammar of the Sinhalese Language*. Asian Educational Services, New Delhi, Madras, India, 1986.

[49] H. P. Ray, *The archaeology of seafaring in ancient South Asia*. Cambridge University Press, 2003.

[50] A. Herath, Y. Hyodo, Y. Kawada, T. Ikeda, and S. Herath, "A practical machine translation system from japanese to modern sinhalese," *Gifu University*, pp. 153–162, 1994.

[51] N. DeVotta, "From ethnic outbidding to ethnic conflict: the institutional bases for sri lanka's separatist war 1," *Nations and Nationalism*, vol. 11, no. 1, pp. 141–159, 2005.

[52] N. Mallikadevi, "An analysis of the production of plural nouns in sinhala," *Strad*, vol. 10, p. 544 – 551, 2023.

[53] J. B. Disanayake, *Say it in Sinhala*. Lake House Investments Limited, 1985.

[54] S. Pallatthara and P. Weihene, *Sinhala Grammar in Linguistic Perspective*. Colombo, Sri Lanka: S Godage & Brothers, 1966.

[55] C. Liyanage, R. Pushpananda, D. L. Herath, and R. Weerasinghe, "A computational grammar of Sinhala," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2012, pp. 188–200.

[56] C. Jany, "The relationship between case marking and s, a, and o in spoken sinhala," *Santa Barbara Papers in Linguistics*, no. 17, pp. 68–84, 2006.

[57] J. Garland, "Morphological typology and the complexity of nominal morphology in sinhala," *Santa Barbara Papers in Linguistics*, no. 17, pp. 1–19, 2005.

[58] M. Henderson, "Between lexical and lexico-grammatical classification: nominal classification in sinhala," *Santa Barbara Papers in Linguistics*, p. 29, 2005.

[59] C. Kirov, C. Johny, A. Katanova, A. Gutkin, and B. Roark, "Context-aware transliteration of romanized south asian languages," *Computational Linguistics*, pp. 1–61, 2024.

[60] Y. De Mel, K. Wickramasinghe, N. de Silva, and S. Ranathunga, "Sinhala transliteration: A comparative analysis between rule-based and seq2seq approaches," *arXiv preprint arXiv:2501.00529*, 2024.

[61] M. Velayuthan and K. Sarveswaran, "Egalitarian Language Representation in Language Models: It All Begins with Tokenizers," *arXiv preprint arXiv:2409.11501*, 2024.

[62] A. Petrov, E. La Malfa, P. H. Torr, and A. Bibi, "Language model tokenizers introduce unfairness between languages," *arXiv preprint arXiv:2305.15425*, 2023.

[63] K. Kumarasinghe, G. Dias, and I. Herath, "SinMorphy: A Morphological Analyzer for the Sinhala Language," in *2021 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2021, pp. 681–686.

[64] Y. Ekanayaka, R. Pushpananda, V. Welgama, and C. Liyanage, "Applying Deep Learning for Morphological Analysis in the Sinhala Language," *The International Journal on Advances in ICT for Emerging Regions*, vol. 16, pp. 2–10, 2023.

[65] M. W. A. R. Sathsarani, T. P. A. B. Thalawaththa, N. K. Galappaththi, J. N. Danthanarayana, and A. Gamage, "Sinhala Part of Speech Tagger using Deep Learning Techniques," in *2022 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*. IEEE, 2022, pp. 1–6.

[66] D. Gunasekara, W. V. Welgama, and A. R. Weerasinghe, "Hybrid part of speech tagger for sinhala language," in *Advances in ICT for Emerging Regions (ICTer), 2016 Sixteenth International Conference on*. IEEE, 2016, pp. 41–48.

[67] A. P. B. Kanduboda, "On the usage of sinhalese differential object markers object marker */wa/* vs. object marker */ta/*," *Theory and Practice in Language Studies*, vol. 3, no. 7, p. 1081, 2013.

[68] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English," *arXiv preprint arXiv:1902.01382*, 2019.

[69] A. Taylor and A. S. Kroch, "The penn-helsinki parsed corpus of middle english," *MS. University of Pennsylvania*, p. 30, 1994.

[70] O. E. Haugen, T. M. Bruvik, M. Driscoll, K. G. Johansson, R. Kyrkjebø, and T. J. Wills, "The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources," 2008.

[71] C. Odebrecht, M. Belz, A. Zeldes, A. Lüdeling, and T. Krause, "RIDGES Herbology: designing a diachronic multi-layer corpus," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 695–725, 2017.

[72] C. Galves, "The tycho brahe corpus of historical portuguese: Methodology and results," *Linguistic Variation*, vol. 18, no. 1, pp. 49–73, 2018.

[73] M. Contreras Seitz, "Hacia la constitución de un corpus diacrónico del español de chile," *RLA. Revista de lingüística teórica y aplicada*, vol. 47, no. 2, pp. 11–134, 2009.

[74] J. Gippert and M. Tandashvili, "Structuring a diachronic corpus. The Georgian National Corpus project," in *Historical corpora. Challenges and perspectives*. Narr, 2015, pp. 305–322.

[75] L. Borin, M. Forsberg, and J. Roxendal, "Korp-the corpus infrastructure of spräkbanken." in *LREC*, vol. 2012, 2012, pp. 474–478.

[76] M. Cranston, *Lexique pour l'explication de texte*. JSTOR, 1983.

# A Introduction to the Sinhala Language

Sinhala is the native language of the Sinhalese, an ethnic group in Sri Lanka [30–32]. It is spoken as a mother tongue (L1) by 16 million people [33, 34] and used as a first (L1) or second (L2) language by the majority of the population on the island [35]. Sri Lanka is the only country where Sinhala is recognised as an official language [3, 36].

Belonging to the Indo-Aryan branch of the Indo-European family [37–39], Sinhala diverges from its Germanic relative English, while maintaining ties with South Asian linguistic traditions. Its script, a descendant of Brahmi [40–46], developed as an abugida system where consonant-vowel sequences are encoded as single graphemic units [47]. The evolution of the script reflects deep historical continuity, with borrowings from Tamil, Portuguese, Dutch, and English enriching its lexicon across centuries [48].

Sinhala has long-standing cultural and literary traditions, with recorded literature stretching back to the 3rd–2nd centuries BCE [49, 50]. This historic depth reinforces its significance in the socio-political fabric of Sri Lanka, where it functions as an official language alongside Tamil [36]. Politically, Sinhala has been both a unifying and contested marker of identity, with its legal status tied to shifts in governance and ethnic relations [51]. The diglossic nature of Sinhala, between a literary written form and a divergent spoken register, adds further complexity [52].

Linguistically, Sinhala exhibits Subject-Object-Verb (SOV) ordering and head-final structures. It enforces subject–verb agreement in its written form [53, 54], contrasting with greater flexibility in speech [55]. Grammatical distinctions based on gender, number, person, and animacy add further complexity [56–58]. These structural properties, combined with its unique orthography and socio-historical trajectory, makes Sinhala both a valuable linguistic heritage and a challenging target for computational modelling.

## B Current Status of Sinhala NLP

Although Sinhala has an extensive literary heritage, its status in natural language processing remains severely underdeveloped [3]. It is classified as a *category 02* low-resource language [1]. Existing digital resources for Sinhala are fragmented and sparse: even basic lexical resources, such as dictionary datasets, are limited in scale and scope [6, 7]. Similarly, traditional language models [8–10] and large pre-trained LLMs [11, 12] for Sinhala are still at an early stage of development.

Much of the research so far has focused on foundational tasks such as optical character recognition (OCR) [22–24], transliteration [59, 60], and basic tokenisation [61, 62], with models such as `Google Document AI`[6] and `Surya` [23, 24][7] offering some level of coverage. Yet, their accuracy is insufficient for corpus-scale applications without extensive manual correction [24]. Downstream NLP tools such as morphological analysers [63, 64], POS taggers [65, 66], parsers [55, 67], exist mostly as prototypes, with no large, publicly available implementations [3]. More recently, efforts such as SinLlama [11] and multilingual projects such as FLORES [68] and Aya [12] have begun to include Sinhala, but their contributions are still shallow compared to high-resource languages.

The most critical gap, however, lies in the absence of diachronic resources. Sinhala has evolved significantly over more than two millennia, but there is no structured corpus that allows researchers to trace changes in syntax, vocabulary, and semantics across historical periods. Without such a dataset, the study of semantic drift, a key method for identifying how words shift meaning over time, remains virtually impossible. The proposed Sinhala Diachronic Corpus is thus designed to fill this void by providing a curated, annotated, and open-access resource that enables semantic, historical, and computational investigations. In doing so, it will address one of the most urgent bottlenecks in Sinhala NLP.

## C Comparative Diachronic Corpora

The planned Sinhala Diachronic Corpus will be situated within a well-established tradition of historical corpus building. Comparative analysis with existing corpora in other languages demonstrates both feasibility and methodological grounding.

### C.1 LatinISE (Latin)

The `LatinISE` corpus [13] contains millions of words spanning more than a millennium, designed for integration into the Sketch Engine platform[1]. Like Sinhala, Latin represents a historically rich language with diachronic variation, making it a relevant benchmark for long-term textual evolution comparison.

### C.2 ARCHER, COHA, PPCHE (English)

The Corpus of Historical American English (`COHA`) [19] and the `ARCHER` [18] corpus represent large-scale English diachronic corpora covering multiple centuries. Although English is a high-resource language, these corpora provide methodological precedents in terms of balance, annotation, and access models that we may use for Sinhala. `COHA` demonstrates the importance of genre balance (fiction, non-fiction, newspapers, magazines) across decades to avoid skewing results by register shifts and `ARCHER` balances genres across 50-year periods to ensure representativeness. The Penn Parsed Corpora of Historical English (`PPCHE`) [69] illustrates the benefits of consistent syntactic annotation schemes for comparability across time.

### C.3 IcePaHC, FarPaHC, and Menota (Icelandic, Faroese and Nordic)

The Icelandic Parsed Historical Corpus (`IcePaHC`) is a syntactically annotated resource covering several centuries of Icelandic [14]. Its companion, the Faroese Parsed Historical Corpus (`FarPaHC`), provides a diachronic dataset for Faroese, a low-resource language cited as comparable to Sinhala in terms of data availability [1]. The `Menota` archive [70] shows how medieval texts can be represented at multiple levels (facsimile, diplomatic transcription, harmonised orthography), which is directly relevant for the potential analysis of ancient Sinhala texts.

### C.4 GerManC, DTA, ReM, RIDGES (German)

The GerManC corpus [21] shows how regional diversity and multi-genre sampling strengthen diachronic analysis. The Deutsches Textarchiv (DTA) [20] highlights the importance of large-scale coverage with TEI-based[3] metadata standards. The ReM klein2016handbuch corpus demonstrates diplomatic transcription alongside modernised annotation, while RIDGES [71] illustrates layered OCR-to-clean pipelines, which we may adapt for Sinhala OCR processing.

### C.5 DIAKORP (Czech)

The DIAKORP [17] project will serve as an important comparative model for us. It is a diachronic corpus of Czech. DIAKORP integrates texts from multiple historical periods, offering both diplomatic and normalised transcriptions, and provides TEI/XML-based metadata covering authorship, publication details, and genre. A key feature of DIAKORP is its layered annotation, including lemmatisation, morphological tagging, and syntactic parsing, which enables users to trace changes in Czech orthography, vocabulary, and grammar over centuries. Importantly, the project demonstrates how large-scale OCR digitisation efforts can be incrementally refined with manual correction to produce research-ready resources.

### C.6 Other Models (Portuguese, Spanish, Georgian, Swedish)

The Tycho Brahe Corpus of Historical Portuguese (TBCHP) [72] demonstrates long-span diachronic corpora with layered annotation. Spanish corpora like CORDE [73][4] and CdE[5] showcase scale and metadata richness. The Georgian National Corpus [74] demonstrates stratified historical layers with harmonised orthography. Swedish resources (e.g., Språkbanken's corpora [75]) show how very large OCR-based collections can be incrementally improved.

### C.7 Positioning Sinhala

The Sinhala Diachronic Corpus will follow these international best practices. It will incorporate:

- Balanced coverage across historical eras, comparable to ARCHER [18] and COHA [19].
- Layered OCR and post-correction pipelines, inspired by RIDGES [71] and DTA [20].
- Annotation schemes aligned with Penn-Helsinki practices [69], following IcePaHC and FarPaHC [14].
- A token count comparable to or exceeding that of FarPaHC, which is in the same language resource category as Sinhala.

Thus, our corpus will extend the tradition of diachronic corpus building to a South Asian, low-resource language, providing opportunities for comparative linguistic research on par with European and global efforts.

## D Planned Technical Methodology of Dataset Development

### D.1 Literature Acquisition and Selection

We will systematically identify Sinhala literary works from the digital collections of the National Library of Sri Lanka (Natlib) and other repositories such as the Department of Archives, university libraries, museum libraries, the Department of Official Languages, the Royal Asiatic Society of Sri Lanka, and the Postgraduate Institute of Archaeology. After filtering for the availability of scanned copies, historical relevance [28], and compliance with *Sri Lanka's Intellectual Property Act No. 36 of 2003* [25], only texts meeting these criteria will be retained. The dataset will aim to cover a wide historical span, ensuring diachronic representation across centuries.

[3] Text Encoding Initiative (TEI) guidelines

[4] http://www.rae.es/recursos/banco-de-datos/corde

[5] https://www.corpusdelespanol.org/hist-gen/

### D.2 Written Date Annotation

Since issue dates often fail to reflect original authorship, we will establish written dates using Sannasgala [28] and other linguistic references. Works consisting of commentaries [76] will be anchored to their original source texts rather than later interpretations, ensuring accurate placement within the diachronic timeline.

### D.3 OCR and Post-processing

Digitised works will be processed using existing Sinhala OCR systems such as `Google Document AI`[6] and `Surya`[7]. If necessary, we will extend or develop new OCR models to improve coverage [22, 23]. OCR outputs will undergo extensive manual post-processing by native Sinhala speakers, correcting spacing errors and misplaced words [27, 29], as well as removing page numbers and any other artefacts. This step will be critical for maintaining structural fidelity, particularly in poetry.

### D.4 Metadata Schema

Following the conventions established in corpora such as `LatinISE` [13] and `IcePaHC`/`FarPaHC` [14], each text will be paired with a metadata record including title, author (Sinhala and romanised), written and issue dates, genre classification, and OCR confidence. Genres will be categorised into broad groups (Fiction vs. Non-Fiction), with finer distinctions such as Religious, Historical, Poetic, Linguistic, and Medical texts.

### D.5 Corpus Composition

The corpus will be comparable in scale to established diachronic corpora in low-resource languages, such as `FarPaHC` [14] (Faroese). While broader chronological coverage will be sought, heavier representation is expected from later centuries following the introduction of the printing press in Sri Lanka in 1737 [26, 27]. Genres such as religious texts and poetry are likely to dominate, reflecting Theravāda Buddhist traditions and influence from Sanskrit literary culture [3].

### D.6 Research Potential

The resulting Sinhala Diachronic Corpus will be the first structured diachronic resource for Sinhala, enabling studies on lexical semantic change, neologism tracking, historical syntax, and corpus-based lexicography. Furthermore, it will provide the foundation for training diachronic embeddings to detect semantic drift and identify anchor words for future comparative linguistic and cultural investigations.

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

[7] `https://github.com/VikParuchuri/surya`