PROGRESS REVIEW 3

The Dynamics of Meaning:

Towards the Evaluation of Diachronic Semantic Drift in Sinhala Language

Presented By **Nevidu Jayatilleke**

Supervised By **Dr. Nisansa de Silva**

Reg No: 258027C

OVERVIEW

Agenda

1 Recap

- What is Semantic Drift?
- Previous Progress Review Work
- Papers Published, Presented & Accepted
- Related Work: The Low-Resource Gap

2 Progress Since Review 2

- SiDiaC-v.2.5: Lemmatised & POS-Tagged Corpus
- Diachronic Embeddings (Word2Vec, FastText, Contextualised)
- Computational Semantic Mapping (SMA & Orthogonal Procrustes)
- Evaluation of Semantic Drift & Bidirectional Impact Scoring
- AACL-IJCNLP 2026 (ARR Reviewed, Accepted to Main Conference)
- SiDiaC-v.3.0: Corpus Expansion (Ongoing)
- Synthetic LSC Evaluation Framework (Ongoing)

3 Looking Ahead

- Challenges & Deviations
- Research Objectives
- Planned Work
- References

RECAP

What is Semantic Drift?

Semantic drift refers to a change in the meaning of certain words for a variety of reasons and in different contexts. [1]

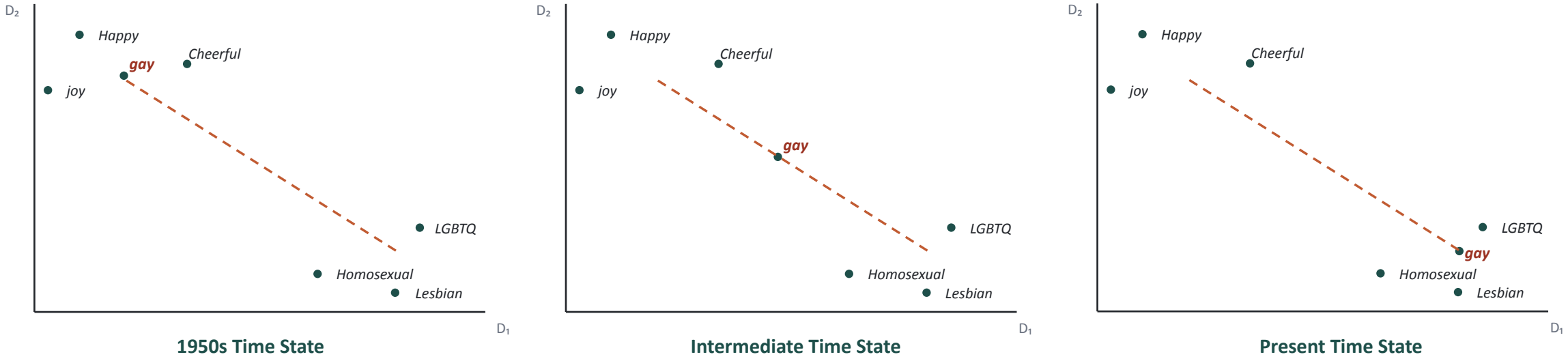
Synchronic

Semantic sense changes across domains

Diachronic

Semantic sense changes across time, the focus of this study

A well-known example: the word "gay" has shifted in meaning from cheerful ⟶ homosexual over the years. [2]



In technical terms, it involves a shift in a word's position within the latent space. [2]

[1] Schlechtweg, D., Hätty, A., Del Tredici, M. and Schulte im Walde, S., 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. ACL 2019, pp.732-746.

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

RECAP

The SiDiaC Family

SiDiaC-v.1.0

Published at PACLIC 2025

46 books · 58k words · 5th–20th c. CE

SiDiaC-v.2.0

Published at LREC 2026

186 books · 244k words · 1800–1955 CE (issued)

59 books · 67k words · 5th–20th c. CE

SiDiaC-v.2.5

Accepted at AACL-IJCNLP

Lemmatised + POS-tagged subset · 59 books · 67k words

SiDiaC-v.3.0

Ongoing

Targeting 62 books · 180k+ words · new tagging scheme

Previously (Reviews 1 & 2):

- SiDiaC-v.1.0 identified Surya as the best OCR engine for Sinhala; Document AI performed best in realistic conditions (RANLP 2025). [3]
- SiDiaC-v.2.0 became the largest Sinhala diachronic corpus to date: 244k words across 186 literary works. [4]

[3] Jayatilleke, N. and De Silva, N., 2025. Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil. RANLP 2025, pp.471-480.
[4] Jayatilleke, N. et al., 2026. SiDiaC-v.2.0: Sinhala Diachronic Corpus Version 2.0. LREC 2026, pp.6740-6763.

RECAP

Summary of Papers: Published, Presented & Accepted

Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil	**RANLP 2025**	**Presented & Published**
SiDiaC: Sinhala Diachronic Corpus (v.1.0)	**PACLIC 2025**	**Presented & Published**
Sinhala Diachronic Corpus (dataset proposal)*	**AI4Science @ NeurIPS 2025**	**Accepted (poster)**
SiDiaC-v.2.0: Sinhala Diachronic Corpus Version 2.0	**LREC 2026**	**Presented & Published**
Dynamics of Meaning: Towards the Evaluation of Diachronic Semantic Change in Sinhala†	**AACL-IJCNLP 2026**	**Accepted (Main Conference)**

** Dataset proposal authored by Dr. Nisansa de Silva (supervisor). Nevidu assisted in writing it and is not listed as an author.*

†This work also got accepted at GlobalSouthAI @ NeurIPS 2026 as non-archival short paper

The Semantic Drift Detection Landscape

Foundational Frameworks

- Hamilton et al. (2016) [5]: the “law of conformity” (frequent words change slowly) and “law of innovation” (polysemous words change rapidly); global embedding shifts capture regular drift, local-neighbourhood shifts capture irregular cultural change.
- Frermann & Lapata (2016) [6]: SCAN, a dynamic Bayesian model treating word senses as evolving over continuous time; outperforms temporally-isolated baselines on sparse historical data.
- Rudolph & Blei (2017) [7]: Dynamic Bernoulli Embeddings, modelling semantic evolution as a Gaussian random walk across time slices.
- Cuba Gyllensten et al. (2020) [8] / Kiyama et al. (2025) [9]: contextualised clustering (SenseCluster) and diachronic word-similarity matrices for continuous, unsupervised shift analysis.

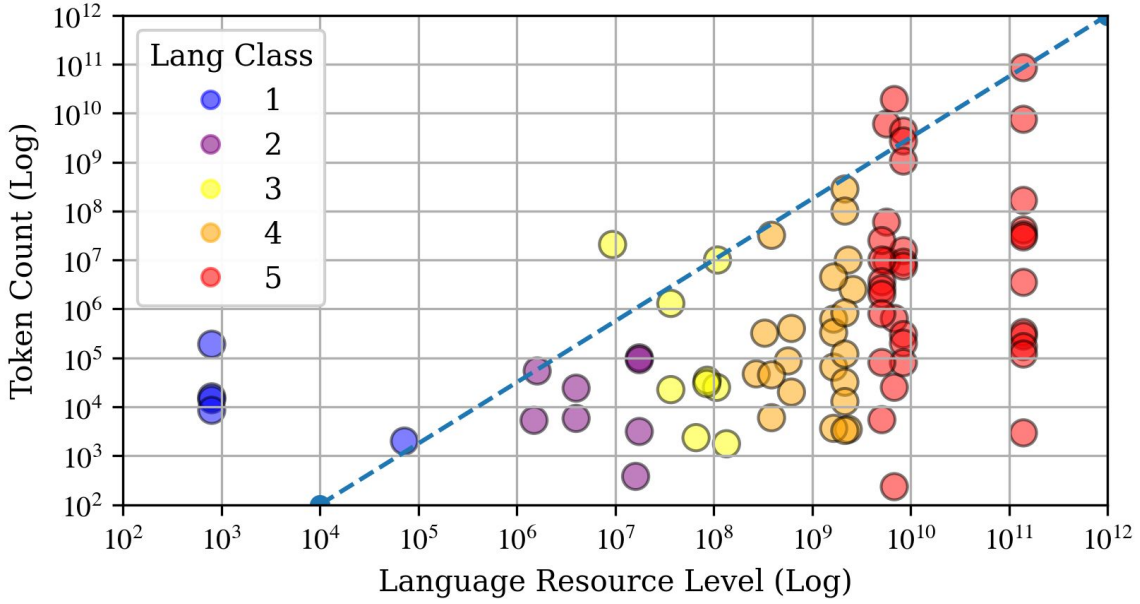
Low-Resource Specific Studies

- Ehrenworth & Keith (2023) [10]: downsampled SemEval-2020 data to ~150k tokens; found a 67% average performance drop across static and contextual models under data scarcity.
- Montariol & Allauzen (2019) [11]: tested dynamic embedding models under simulated scarcity (down to 1% of a corpus); a “backward external” initialisation from modern pretrained vectors significantly improved sparse historical performance.
- Across both studies: low-resource settings are treated as a constraint to work around, rarely as the primary object of study, a gap our work addresses directly for Sinhala.

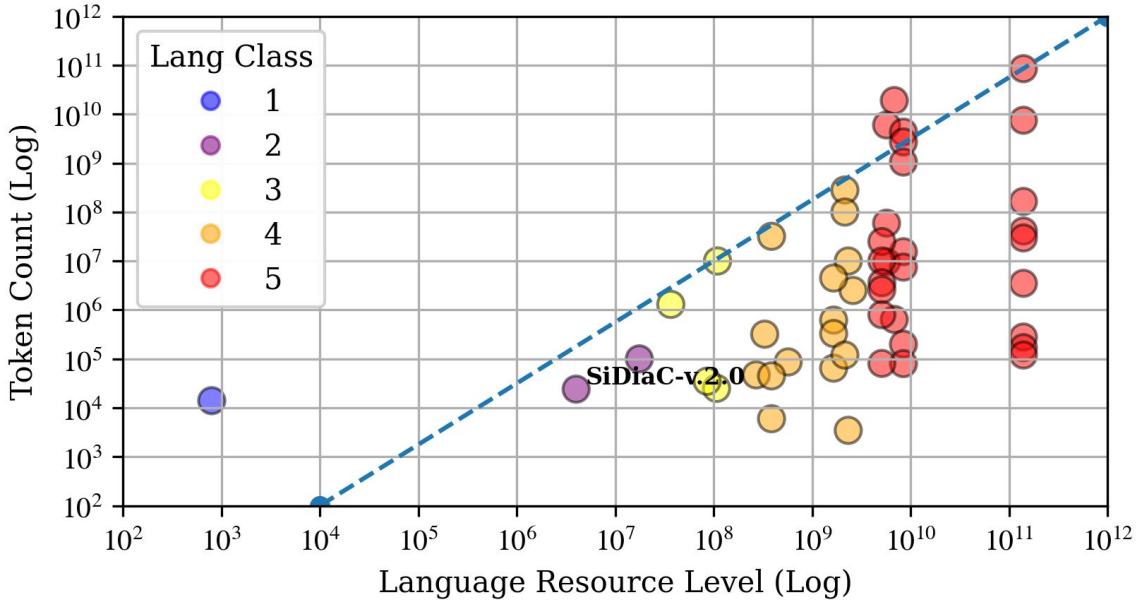
[5] Hamilton, W.L., Leskovec, J. and Jurafsky, D., 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL 2016, pp.1489-1501.
[6] Frermann, L. and Lapata, M., 2016. A Bayesian Model of Diachronic Meaning Change. TACL, 4, pp.31-45.
[7] Rudolph, M. and Blei, D., 2017. Dynamic Bernoulli Embeddings for Language Evolution. arXiv:1703.08052.
[8] Cuba Gyllensten, A., Gogoulou, E., Ekgren, A. and Sahlgren, M., 2020. SenseCluster at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. SemEval 2020, pp.112-118.
[9] Kiyama, H. et al., 2025. Analyzing Continuous Semantic Shifts with Diachronic Word Similarity Matrices. COLING 2025, pp.1613-1631.
[10] Ehrenworth, J. and Keith, K., 2023. Literary Intertextual Semantic Change Detection: Application and Motivation for Evaluating Models on Small Corpora. LChange 2023, pp.1-14.
[11] Montariol, S. and Allauzen, A., 2019. Empirical Study of Diachronic Word Embeddings for Scarce Data. RANLP 2019, pp.795-803.

The Low-Resource Gap in LSC Research

Plotting every diachronic corpus surveyed for SiDiaC-v.2.0 [4] by log token count against log language-resource level exposes how little semantic-drift research touches low-resource languages.



(a) 81 listed diachronic corpora.



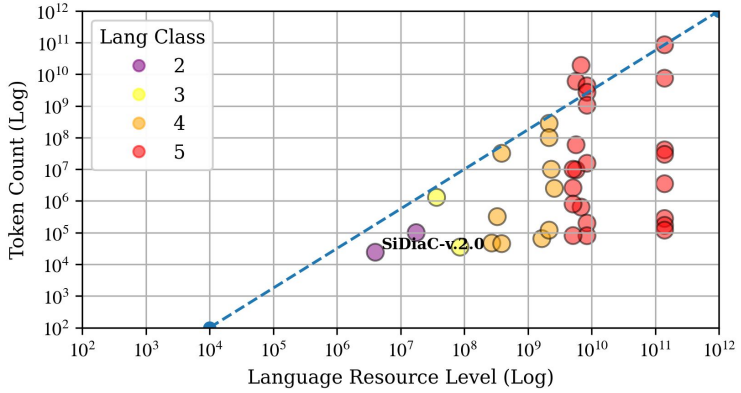
(b) 50 Corpora that have undergone semantic drift evaluation.

Left: all 81 listed diachronic corpora, coloured by language-resource class (0 = lowest, 5 = highest). Right: only the corpora that have actually undergone semantic-drift evaluation; SiDiaC-v.2.0 is one of very few Class-2 corpora represented at all.

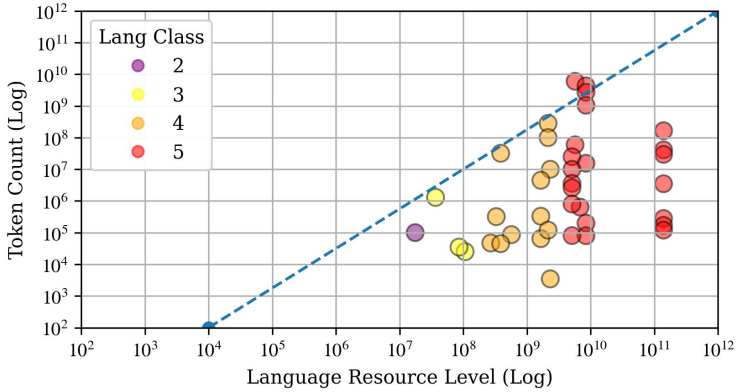
[4] Jayatilleke, N. et al., 2026. SiDiaC-v.2.0: Sinhala Diachronic Corpus Version 2.0. LREC 2026, pp.6740-6763.

The Gap Widens by Evaluation Approach

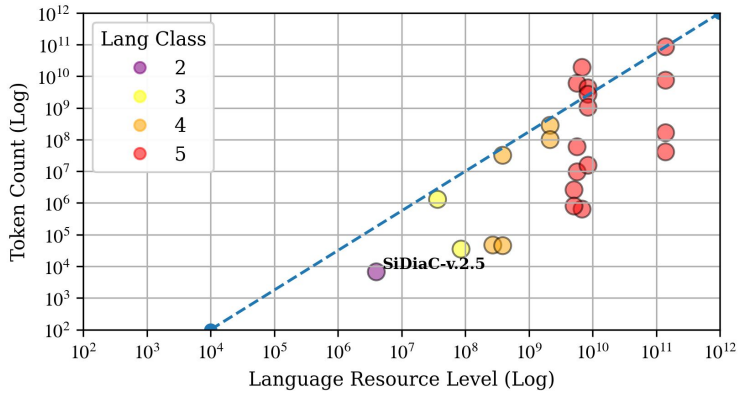
Splitting those same corpora by the type of semantic-drift evaluation they have undergone (statistical, qualitative, or embedding-based) shows the low-resource gap is worse than it first appears.



(a) 37 have undergone statistical semantic drift evaluation.



(b) 39 have undergone qualitative semantic drift evaluation.



(c) 22 have undergone embedding based semantic drift evaluation.

Embedding-based evaluation (the most powerful and most data-hungry approach) has the fewest low-resource corpora of the three (figure c). This is precisely the gap the Word2Vec/FastText/Llama-FT framework in this study targets for Sinhala.

[4] Jayatilleke, N. et al., 2026. SiDiaC-v.2.0: Sinhala Diachronic Corpus Version 2.0. LREC 2026, pp.6740-6763.

SECTION 2

Let's Discuss the Progress Since the Previous Review

NEW CORPUS VERSION

SiDiaC-v.2.5: Enhanced Sinhala Diachronic Corpus

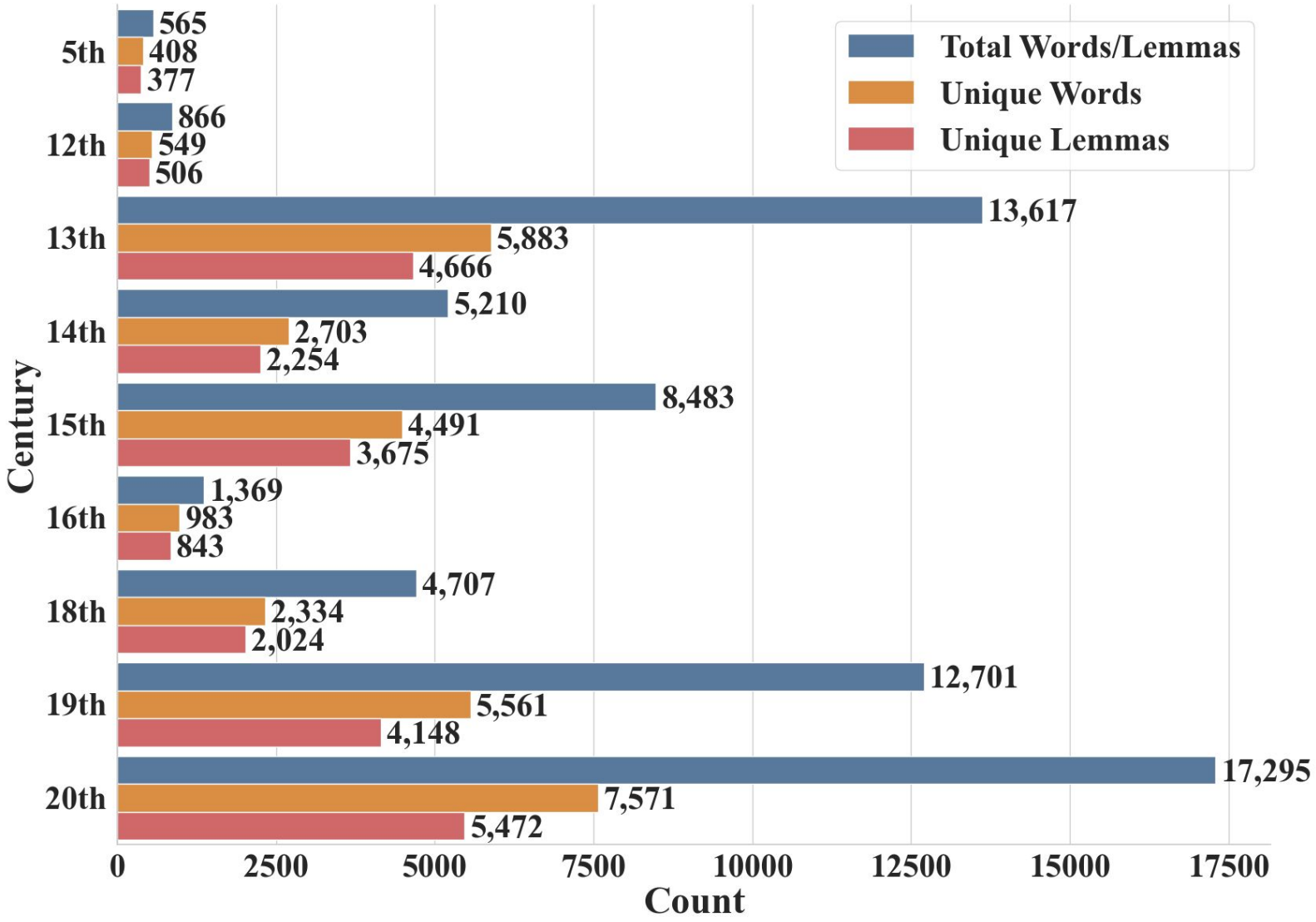
- SiDiaC-v.2.5 upgrades the written-date-filtered subset of SiDiaC-v.2.0 by adding two new linguistic features: lemmas and Part-of-Speech (POS) tags for every word.
- Built on 59 books totalling 67,005 words, spanning the 5th to the 20th century CE.
- Lemmatisation performed using SinMorphy-v.1.2 [12], a Finite State Transducer-based morphological analyser for Sinhala.
- POS tagging performed with a composite pipeline combining SinMorphy, an improved TnT POS Tagger for Sinhala [13], and the Sinling library.

Why enhance the corpus?

- Lemmas collapse morphological variance so that word forms can be compared reliably across centuries.
- POS tags offer an additional feature that researchers can utilise to study historical grammar in future research.
- Together, these features make SiDiaC-v.2.5 suitable for training and evaluating diachronic word embeddings.

[12] Kumarasinghe, K., Dias, G. and Herath, I., 2021. SinMorphy: A Morphological Analyzer for the Sinhala Language. MERCon 2021, pp.681-686.
[13] Fernando, S. and Ranathunga, S., 2018. Evaluation of different classifiers for Sinhala POS tagging. MERCon 2018, pp.96-101.

Word & Lemma Token Counts by Century



Key figures

- The 5th, 12th, and 16th centuries were excluded from the semantic-change study due to insufficient word counts.
- The 16th century's small size reduced consistent shared words across all centuries to 80; removing it raised this to 210.
- The final analysis spans six century-specific corpora: 13th, 14th, 15th, 18th, 19th, and 20th.

Composite Lemmatisation & POS Tagging Pipeline

1

SinMorphy v1.2 [12]

Rule-based FST morphological analyser. Assigns candidate POS tags & lemma strings; a "guesser" handles unlisted words.

2

TnT POS Tagger [13]

Statistical HMM tagger (improved with a beam-search Viterbi for efficiency). First consulted to disambiguate multi-tag cases.

3

Sinling (fallback)

Used when TnT is unknown or non-matching, to resolve remaining ambiguous POS candidates.

4

Final Tag + Lemma

Standardised to 7 main-level POS categories; lemma isolated by splitting the morphological string at the "+" delimiter.

Handling ambiguity:

A strict priority hierarchy maps raw tags from TnT and Sinling to seven main-level categories (NOUN, VERB, PART, PROPER NOUN, ADJECTIVE, PRONOUN, FOREIGN WORD). Of 23,180 words unresolved by SinMorphy's lexicon (labelled "UNK"), the composite pipeline resolves the correct tag by cross-referencing TnT and Sinling outputs, defaulting to the analyser's first candidate only when no match is found.

[12] Kumarasinghe, K., Dias, G. and Herath, I., 2021. SinMorphy: A Morphological Analyzer for the Sinhala Language. MERCon 2021, pp.681-686.
[13] Fernando, S. and Ranathunga, S., 2018. Evaluation of different classifiers for Sinhala POS tagging. MERCon 2018, pp.96-101.

Three Vectorisation Approaches

Word2Vec [14]

A separate Skip-gram model trained per century (vector size 50, window 5, negative sampling). Hyperparameters were set to suited small, sparse corpora; it learns useful representations even for infrequent terms.

FastText [15]

A pre-trained Facebook FastText model (cc.si.300.bin) trained on Sinhala Common Crawl, extended with century-specific vocabulary and fine-tuned per century, then pruned to that century's words.

Contextualised (SLMs)

Five Small Language Models evaluated: XLM-R (encoder) and four decoder models (Llama-3.1-8B base/instruct, SinLlama [16], Gemma-2-9b), continually pre-trained on SiDiaC-v.2.0.

Why period-specific retraining is bypassed for contextualised models:

Unlike Word2Vec and FastText, frozen pretrained SLM representations are leveraged directly per century. The hypothesis: high-dimensional embedding spaces contain inherent low-dimensional manifolds representing century-specific semantic shifts, bypassing the overfitting and alignment issues that come with retraining on small per-century datasets.

[14] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality. NeurIPS 2013.
[15] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2017. Enriching Word Vectors with Subword Information. TACL, 5, pp.135-146.
[16] Aravinda, H.W.K. et al., 2025. SinLlama: A Large Language Model for Sinhala. MERCon 2025, pp.617-622.

Selecting the Contextualised Embedding Model

Five Small Language Models were evaluated on SiDiaC-v.2.0 via Perplexity and Bits-Per-Byte (BPB) after Continual Pre-Training (CPTing):

Model	Tokeniser	Training	Perplexity	Bits-Per-Byte
Llama-3.1-8B	Base	✗	2.86	1.04
Gemma-2-9b			25.17	1.49
SinLlama			186.67	0.92
Llama-3.1-8B-it			3.42	1.22
XLM-R			–	0.80
Llama-3.1-8B	Base	✓	**2.18**	**0.77**
SinLlama			2754999.26	2.62
XLM-R			–	3.14
Llama-3.1-8B	Expanded	✓	623.76	1.22
SinLlama			203103.64	1.96
XLM-R			–	1.10

Evaluation of Small Language Models on SiDiaC-v.2.0, based on Perplexity and Bits-Per-Byte metrics.

Selected model: Llama-FT

- Llama-3.1-8B fine-tuned with its native base tokeniser ("Llama-FT") achieved the best Perplexity and BPB.
- Vocabulary expansion + CPT unexpectedly increased both metrics, likely due to sparse softmax distributions and insufficient data to stabilise new token embeddings.
- 4096-dim final-hidden-layer word vectors extracted via mean pooling over sub-tokens; a sliding window (510 tokens, step 300) handles long sentences.
- Following [17]'s approach of using instruct models for embeddings, the instruct variant of Llama-3.1-8B was also tested, but underperformed the base model.
- Llama-FT embeddings power the Bidirectional Semantic Impact Scoring analysis.

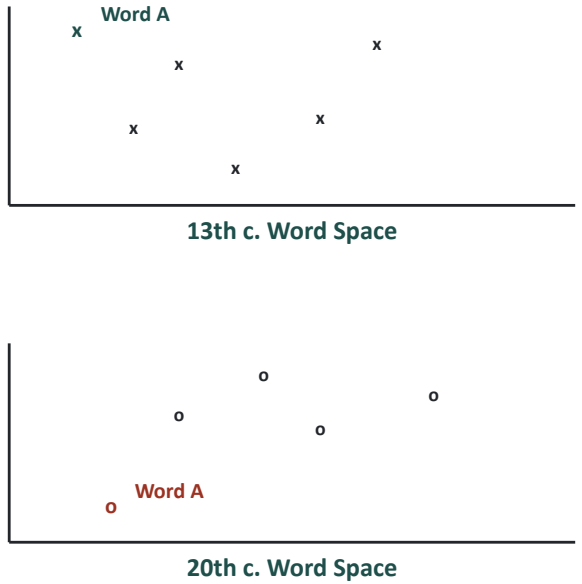
[17] Wang, L. et al., 2024. Improving Text Embeddings with Large Language Models. ACL 2024, pp.11897-11916.

Computational Semantic Mapping Framework

Word2Vec and FastText spaces are randomly initialised per century, so they are "arbitrarily rotated" relative to one another (non-identifiability) [5]. Two complementary alignment strategies address this:

Similarity Matrix Alignment (SMA)

For each lemma, computes a similarity profile: cosine distances to 261 common anchor lemmas. The Pearson correlation between profiles of the same lemma across two centuries measures semantic similarity, bypassing explicit rotation.



Orthogonal Procrustes (OP) Alignment

Rotates one century's space to another using a rotation matrix Q over 210 consistent anchor words. Two variants: Analytical (closed-form SVD) and Neural (learned, orthogonality-penalised linear layer).

Two OP variants explored:

- Analytical Procrustes: rotation matrix Q derived via SVD; the 13th century is the fixed global reference for all others.
- Neural Procrustes: a bias-free linear layer trained (MSE + orthogonality penalty) to map each target century into the 13th-century coordinate system.

[5] Hamilton, W.L., Leskovec, J. and Jurafsky, D., 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL 2016, pp.1489-1501.

Evaluating Orthogonal Procrustes Alignment

Model	Approach	Avg. Procrustes Disparity	Avg. Rotational Distance
Word2Vec	Analytical OP	174.16	9.93
Word2Vec	Neural OP	157.16	9.16
FastText	Analytical OP	57.57	22.82
FastText	**Neural OP**	**57.06**	**19.68**

Interpreting the trade-off:

FastText's higher Rotational Distance (≈19.7) vs. Word2Vec (≈9.3) reflects its rigid, pre-trained subword structure; more rotational effort is required, but internal language structure stays intact, yielding superior alignment precision. Word2Vec's flexible, from-scratch coordinate system rotates easily but distorts as vocabulary expands with more data.

Key finding

- Neural OP for FastText significantly outperforms all other combinations: best manifold preservation, lowest error margins.
- FastText improves with more data (lower disparity from 18th→20th c.); Word2Vec worsens, as a from-scratch model expands its vocabulary space unpredictably with more text.
- Conclusion: FastText is prioritised for the statistical neighbourhood analysis.

Statistical Neighbourhood Analysis

- For each pair of time periods, drift is flagged when a word/lemma's similarity falls below a dynamic threshold: $\mu - (k \times \sigma)$, empirically set at $k = 0.8$.
- Validated against a theoretical Gaussian expectation (21.19% flagged): the four alignment pipelines flagged 16.19%–20.57% (mean 18.36%), confirming $k = 0.8$ is a stable, near-canonical cutoff.
- 15 century-aligned semantic spaces were analysed, covering 261 consistent lemmas and 210 consistent words, yielding 2,116 similarity instances across all model/alignment combinations.
- For each flagged word, the top-5 nearest neighbours in source and target spaces are compared to qualitatively interpret the nature of the shift.

Mean similarity of consistent keys across 15 temporal alignments

	SMA	Analytical OP	Neural OP
Word2Vec	30.23%	61.31%	52.92%
FastText	55.69%	82.11%	81.04%

Note: SMA appears less robust than OP techniques. it considers all consistent lemmas globally, so relational similarity shifts are distorted by localised drifts elsewhere in the vocabulary.

Case Study: රැස් / ræs (FastText, Neural Procrustes)

~ Gampola Kingdom



~ Kotte Kingdom

14th Century

Gathering & Movement

Neighbouring words:

රැස්කරවා (causing to gather), පිළිගෙන (receiving), මධ්‍යයෙහි (in the middle), වේගයෙන් (with speed)

Linked with collecting or assembling, in physical or spatial terms.

15th Century

Symbolic & Doctrinal

Neighbouring words:

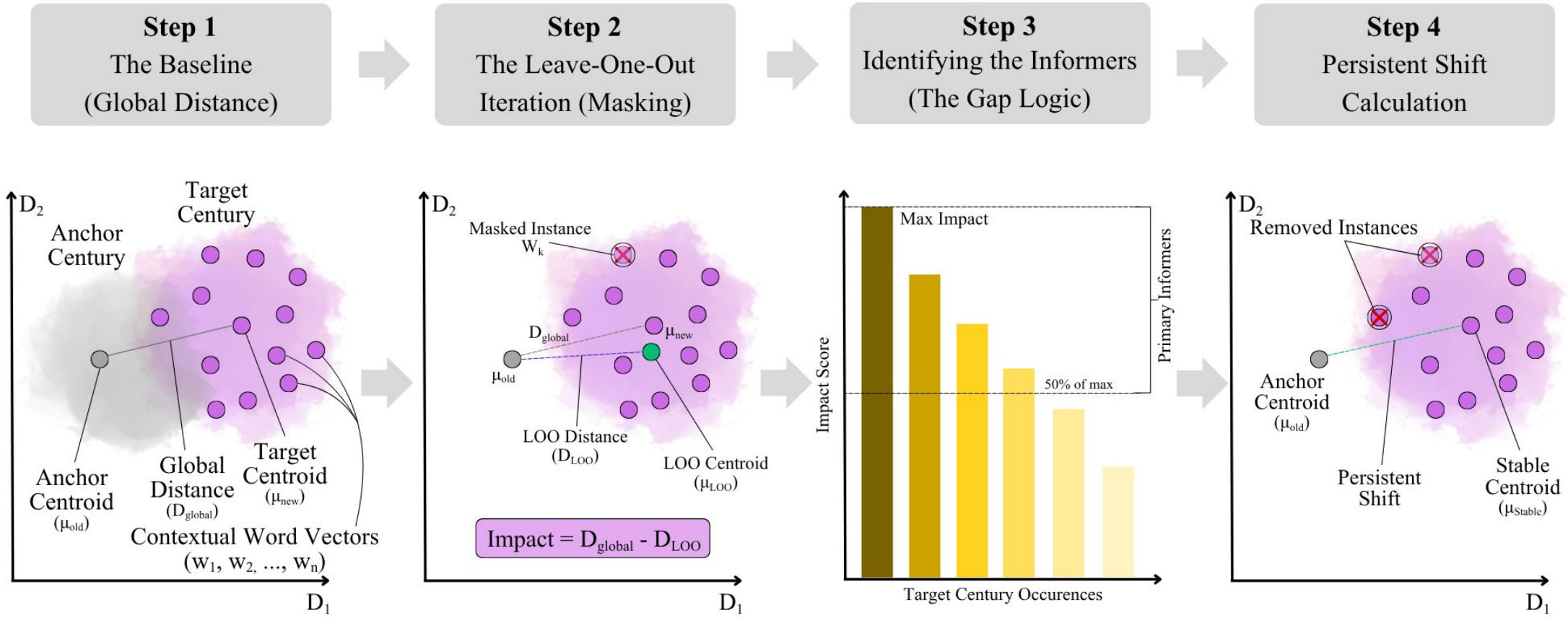
විනාස (destruction/end), ආයු (life span), රහස්‍ය (secret), රත්න (precious object), බුදුරැස් (Buddha's radiance)

Shifts to spiritual meanings: light, brilliance, sacred radiance.

This qualitative neighbourhood method is applied uniformly across all six model-to-mapping combinations to interpret quantitatively-flagged drift candidates.

Bidirectional Semantic Impact Scoring: LOO Analysis

Using Llama-FT contextual embeddings, a Leave-One-Out (LOO) diagnostic isolates which specific sentence usages drive a lemma's semantic drift between centuries.



Leave-One-Out (LOO) analysis for identifying contextual semantic divergence. Step 1: the baseline global distance is established. Step 2: individual contextual word vectors are iteratively masked to measure their impact on centroid displacement. Step 3: textual evidence of semantic divergence is extracted as "primary informers" based on this impact. Step 4: informers are permanently removed to isolate the word's persistent shift in core meaning.

LOO Results: Drift is Concentrated, Not Uniform

0.0295
⬇
0.0272

Avg. global centroid distance before / after removing high-impact instances

8%

Overall drift reduction from removing informer sentences

21,738 / 112,945

Instances identified as having drifted (≈19%)

18th & 19th

Century pair with the highest divergence (0.0340 avg. distance)

Conclusion

- Semantic drift in Llama-FT representations is low but consistent, driven by a smaller set of high-impact contextual instances rather than gradual, uniform change across all occurrences.
- The 18th-century sub-corpus stands out as a persistent outlier: all 10 permutations pairing it with another century show global distances at or above 0.0259, well above the cross-permutation mean (0.0238).
- Permutations targeting the 20th century show the largest instance pools (29,299 total, 5,564 drifted), yet several of the corresponding average distances remain among the lowest, showing greater lexical diversity does not necessarily mean greater semantic deviation per instance.

Case Study: ගුණ / guna, "virtue/quality" (14th → 15th c.)

~ Gampola Kingdom



~ Kotte Kingdom

14th Century

Royal & Sagely Praise

Praises kings, princes, and wise people as figures of high moral character. Related terms: clan/lineage (කුල), city (පුර), compassion (කරුණා), endowment with virtues (සමන්නාගත).

15th Century

Buddhist + Courtly Beauty

Broadens along two lines: intensified Buddhist compassion-praise, and courtly-poetic praise of feminine beauty (e.g. princess Ulakudaya Devi). Fame (කීර්ති) and triumph (ජයගන්) also emerge.

Compassion (කරුණා) remains a common thread across both centuries: the 11 informer sentences reduce drift by 27.31%, showing a stable abstract sense of "virtue/quality" whose objects of praise simply expand, rather than a complete change in meaning.

SUBMISSION STATUS

Accepted at AACL-IJCNLP 2026

“Dynamics of Meaning: Towards the Evaluation of Diachronic Semantic Change in Sinhala”

Reviewed through the ACL Rolling Review (ARR) process. The submission got accepted at AACL-IJCNLP 2026 main conference.

ARR scores received:

2.5	2.5	4.0	3.0
Reviewer 1	Reviewer 2	Reviewer 3	Meta-Review

Key contributions

- Introduces SiDiaC-v.2.5: a lemmatised, POS-tagged enhancement of SiDiaC-v.2.0.
- Aligns century-specific Word2Vec and FastText embeddings via SMA and two Orthogonal Procrustes variants.
- Introduces Bidirectional Semantic Impact Pruning using fine-tuned Llama-3.1-8B contextual embeddings and Leave-One-Out diagnostics.
- Shows diachronic drift in SiDiaC is driven by a smaller set of high-impact contextual instances, not gradual uniform change.
- Provides a preliminary computational framework for diachronic analysis in low-resource language settings.

Expanding the Corpus: A Density-Based Page Estimator

- SiDiaC-v.3.0 targetted a 200,000+ word corpus by extending each of the existing books, with each book targeted to contribute 4,000 words.
- Word density is calculated per book: the average number of words per page, based on pages already extracted in v2.
- This density is extrapolated linearly: dividing 4,000 by the density gives the total pages needed to reach 4,000 words for that book.
- Subtracting the pages already counted gives the additional pages still required for OCR and annotation.
- This assumes word density stays roughly constant across further pages, a projection rather than an exact count from inspecting the remaining pages directly.

The estimation formula

Word density

words OCR'd ÷ pages OCR'd

Pages needed (total)

4,000 ÷ word density

Pages still required

pages needed − pages already OCR'd

*Target: 4,000 words * books → 200,000+ words corpus-wide.*

Note: After post processing of the books we are at 180,000+ words which is below our initial target but still bigger than 150,000 words used by Ehrenworth & Keith (2023) [10].

[10] Ehrenworth, J. and Keith, K., 2023. Literary Intertextual Semantic Change Detection: Application and Motivation for Evaluating Models on Small Corpora. LChange 2023, pp.1-14.

Improving Data Quality: New Annotation Tags

Rather than discarding non-Sinhala and extraneous content as in earlier versions, SiDiaC-v.3.0 retains it and marks it inline, preserving information while keeping it clearly separable for downstream processing.

<sns> … <ens>

Non-Sinhala span

Wraps content that is not in Sinhala (e.g. Pali, Sanskrit, English) instead of removing it, so code-mixed passages remain in the corpus but are clearly delimited for filtering.

<smc> … <emc>

Miscellaneous span

Wraps footnotes and certain bracketed content, material that is present in the source text but sits outside the main narrative flow.

Why tag instead of strip: retaining and delimiting this content (rather than deleting it, as in SiDiaC-v.2.0) keeps the corpus closer to the original source while still letting downstream pipelines cleanly exclude or isolate it as needed.

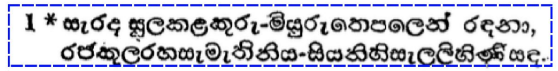
Expert-in-the-Loop

A linguistic expert has joined the SiDiaC-v.3.0 annotation process, both identifying additional eligible material and providing independent quality assurance over the corpus.

Expert-led workstreams

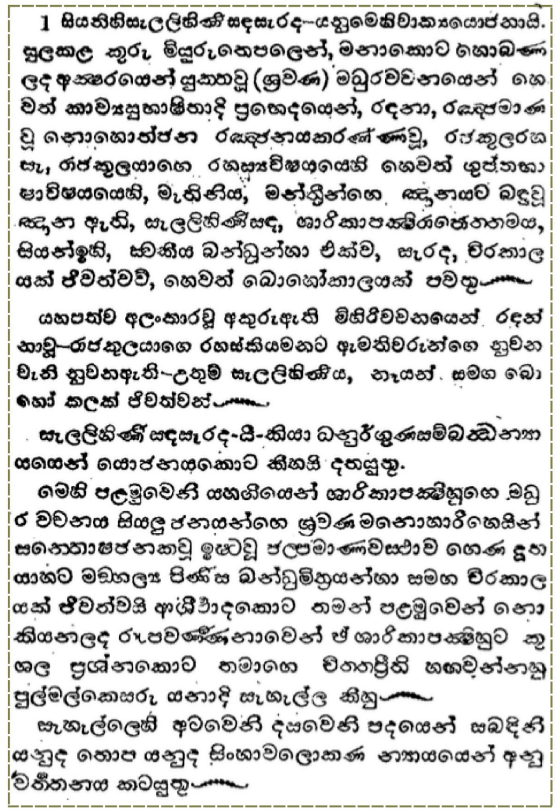
- Identifying additional eligible books: 3 new books added to the corpus this iteration, bringing the total from 59 to 62.
- Separating commentary from original content in “sanna” (commentary) books, and where possible, dating each portion separately rather than inheriting a single date for the whole work.
- Validating the post-processing work carried out by the native-speaker annotator team, providing an independent quality check on the manual corrections.

Original



1 * සැරද සූලනකුරු-මියුරුතොපලෙන් රඳනා,
රජකුලරහසැමැතිනිය-සියනිගිසැලලිඟිණි ගද.

Commentary
සන්නය \sʌnnʌjʌ



1 සියනිගිසැලලිඟිණි සඳසැරද–යනුමෙකිවාක්‍යයොජනායි.
සූලකළු කුරු මියුරුතෙපලෙන්, මනාකොට හොබණ
ලද අක්ෂරයෙන් යුක්තවූ (ශ්‍රව්‍ය) මධුරවචනයෙන් හෙ
වත් කාව්‍යසුභාෂිතාදි ප්‍රභෙදයෙන්, රඳනා, රූපමාණ
වූ නොහොත්ජනා රූපජනාගතාරම්මණවූ, රජකුලරහ
සැ, රාජකුලයාහෙ රහස්‍යවිෂයයෙහි හෙවත් ගුප්තභා
ෂාවිෂයයෙහි, මැතිනිය, මන්ත්‍රීන්හෙ ඤාණය බලවූ
ඤාණ ඇති, සැලලිඟිණිසඳ, ශාරිකාපක්ෂිරාජෝත්තමය,
සියන්ගිඟි, සඛනිය බන්ධූන්හා එක්ව, සැරද, චිරකාල
යක් ජීවත්වවි, හෙවත් බොහෝකාලයක් පවතු.

යහපත්ව අලංකාරවූ අතුරුඇති මිහිරිවචනයෙන් රඳන්
නාවූ–රාජකුලයාගෙ රහස්කියමනට ඇමතිවරුන්ගෙ භූවන
වැනි භූවනඇති–උතුම් සැලලිඟිණිය, නෑයන් සමඟ බො
හෝ කලක් ජීවත්වන්.

සැලලිඟිණි සඳසැරද-යී-නියා ධන්‍යර්ගුණසම්බන්ධීන්‍යා
යයෙන් යොජනායකොට නිහයි දතයුතු.

මෙහි පළමුවෙනි යහයිගෙන් ශාරිකාපක්ෂිහුගෙ මධු
ර වචනය සියලු ජනයන්හෙ ශ්‍රවණ මනොහාරිහෙයින්
සන්තොෂජනකවූ ඉෂ්ටවූ ප්‍රලපමාණවස්තුව හෙණ දූත
යාහට මඛ්‍යලය පිණිස බන්ධුම්හුයන්හා සමඟ චිරකාල
යක් ජීවත්වයි ආශීර්වාදකොට තමන් පළමුවෙන් නො
නියමලද රූපවණ්ණනාවෙන් ඒ ශාරිකාපක්ෂිහුට නු
ගල ප්‍රශ්නකොට තමාගෙ චිත්තප්‍රීති හඟවන්නාහු
පුල්මල්නොයරු යනාදි සැහැල්ල නිහු.

සැහැල්ලෙහි අවමෙනි දයමෙනි පදයෙන් සබඳිනි
යනුද නොප යනුද සිංහාවලොකණ නාහයයෙන් අනු
විනීතනාය කටයුතු.

Synthetic Ground-Truth for Semantic Drift Detection

- Building a gold-standard list of words that have genuinely undergone semantic drift is difficult: it demands deep historical-linguistic expertise and is inherently subjective, with no objective ground truth to check against.
- To sidestep this, we adapt the pseudoword-injection methodology of Shoemark et al. (2019) [18], “Room to Glo,” which was designed for a 5.5-year, monthly-resolution Twitter corpus.
- Real words in a neutral corpus are replaced by invented “pseudowords” following controlled probability schedules across time periods, producing an answer key of exactly which words were made to drift, and which were not.
- Completed as a domain-specific case study on SinhaLegal [19], a ~2 million-word benchmark corpus of Sinhala legislative texts spanning the 1980s–2010s. The same framework will next be adapted for SiDiaC-v.3.0.

Why this matters?

- Any LSC-detection method (static or contextual, aligned across time) can be scored against a known answer key rather than subjective human judgement.
- A discriminator validated this way can then be extrapolated with confidence to the real diachronic corpora: SiDiaC-v.3.0 and SinhaLegal.
- Because it requires no expert-curated gold standard and remains workable with limited data, the same framework could be adopted by other low-resource language studies facing this same evaluation gap.

[18] Shoemark, P. et al., 2019. Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. EMNLP-IJCNLP 2019, pp.66-76.
[19] Lasandi, M. and Jayatilleke, N., 2026. SinhaLegal: A Benchmark Corpus for Information Extraction and Analysis in Sinhala Legislative Texts. LoResLM 2026, pp.114-131.

Building the Artificial SinhaLegal Corpus

Documents are binned into four decades (1980s–2010s), standing in for the 66 monthly time-steps used in Shoemark et al.'s original methodology.

1. Shared Sample Pool

A reproducible 10% sample of the 2000s decade's paragraphs becomes the shared source for all four synthetic decades.

2. Replica-1 Baseline

Each decade is resampled with replacement from that same pool (70% of its original size); by construction, no real drift exists yet.

3. Frequency Threshold

A single elbow point on the rank-frequency curve (KneeLocator) defines which words are frequent and stable enough to serve as substitution targets.

4. Six Injection Schemas

Applied cumulatively, each adding 30 pseudowords on top of the last, totalling 180 with 60 each for genuine change vs not change instances.

Every substitution (decade, paragraph, token position, real word replaced, pseudoword, and schema) is logged, and eligibility (frequency, cross-decade presence, valid Sinhala script, never reused) is enforced globally across all six schemas.

Six Pseudoword-Injection Schemas

The C-schemas plant genuine semantic change; the D-schemas mimic surface-level effects (frequency shifts, competing noise) that a good detector should learn to ignore.

Schema	Should Register As	Mechanism
C1: Gradual Sense Acquisition	**Genuine change**	w1 replaced at constant probability; w2's probability starts low and rises over decades (6 timing/shape variants).
C2: Converging Sense Replacement	**Genuine change**	w1's probability falls 0.99→0.01 across decades; w2 is the complement, 1 − P(w1), so the sense fully flips.
C3: Multidimensional Sense Acquisition	**Genuine change**	4 words compete via a Dirichlet-multinomial redrawn per decade; a 5th, more stable sense rises separately.
D1: Increasing Frequency	**Not change**	Single word w1, rising replacement probability: a frequency shift only, no sense change.
D2: Spiking Frequency	**Not change**	w1 constant; w2 spikes up then back down: a “trending topic” effect, no genuine sense change.
D4: Multidimensional, No Stable Sense	**Not change**	4 words via Dirichlet-multinomial per decade, like C3's shifting part, but with no added stable sense.

Adapting Shoemark et al. (2019) [18] to Sinhala

- Decades, not months: dictated by SinhaLegal's decade-labelled documents rather than high-frequency social media data.
- D3 (periodic/seasonal spike) omitted entirely: it requires monthly granularity to model “same month every year,” impossible with 4 decade-level steps.
- A single elbow-point frequency threshold replaces Shoemark et al.'s 5 quantile bins: SinhaLegal's right-skewed frequency distribution would leave most quantile bins containing only unreliably rare words.
- No WordNet-based sense-exclusivity check: no comparable Sinhala WordNet resource exists; substituted with frequency, cross-decade presence, Sinhala-script validity, and strict global reuse exclusivity.
- Added change-timing variation (full-timeline / delayed-change / early-stabilising, crossed with linear/logarithmic shape) to recover probability-curve diversity otherwise lost to having only 4 timesteps instead of 66.
- C3 and D4 scaled from Shoemark et al.'s 8 words down to 5 (4-word Dirichlet-multinomial + 1 stable sense) due to the smaller corpus size.

[18] Shoemark, P. et al., 2019. Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. EMNLP-IJCNLP 2019, pp.66-76.

REFLECTION

Limitations & Challenges

Data Scarcity

Limited historical corpora weakened century-wise Word2Vec models, causing unreliable Procrustes alignment relative to fine-tuned FastText. Transformer perplexity/BPB did not improve with continual pretraining, likely due to insufficient data for expanded-vocabulary tokens.

Lemmatisation

SinMorphy may not have been trained on historical Sinhala orthography, reflecting the broader low-resource status of computational tools for the language [20].

POS Tagging

The absence of high-accuracy POS taggers for historical Sinhala [20] required a hybrid tagging system rooted in contemporary grammar, a modern linguistic bias that may affect accuracy on medieval and early-modern text.

These constraints collectively reflect Sinhala's Category 2 (low-resource) status per Ranathunga & de Silva (2022) [21], motivating the dynamic, data-scarcity-aware thresholds and dual embedding strategy used throughout this study.

[20] de Silva, N., 2025. Survey on Publicly Available Sinhala NLP Tools and Research. arXiv:1906.02358v25.
[21] Ranathunga, S. and De Silva, N., 2022. Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. AACL-IJCNLP 2022, pp.823-848.

LOOKING AHEAD

Research Objectives

Identify & digitise Sinhala linguistic resources

Identifying Sinhala linguistic resources within a defined acceptable timeframe for dataset creation, and digitizing them for research on diachronic semantic drift.

In Progress

95%

Implement a novel temporal dynamic embedding mechanism

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 .

In Progress

90%

Identify anchor words vs. words with significant drift

Identifying semantics that exhibit significant diachronic drift and those that remain constant, serving as anchor or pivot words for further research.

In Progress

50%

Assess semantic change across contiguous and continuous time periods*

Evaluating semantic change in a diachronic corpus across both discrete time periods and continuous time series.

In Progress

70%

** Objective revised from "assess semantic emergence & disappearance in the evolution of the Sinhala language" to "assess semantic change across contiguous and continuous time periods."*

LOOKING AHEAD

Planned Work: Leading to the Next Review

- Complete the SiDiaC-v.3.0 expansion: OCR and annotate the remaining estimated pages for all 62 books to reaching 180,000+ word target.
- Finalise expert-led work on sanna (commentary) books: separating commentary from original content and independently dating each portion where possible.
- Re-run lemmatisation, POS-tagging, and semantic drift evaluation on the expanded SiDiaC-v.3.0 corpus once annotation and expert validation are complete.
- Adapt the synthetic pseudoword-injection framework validated on the SinhaLegal legal-domain case study to SiDiaC-v.3.0, re-tuning frequency thresholds and pseudoword density for its smaller, sparser scale.
- Continue building out the semantic-change-between-time-periods objective (revised from the original emergence/disappearance framing) via the synthetic ground-truth discriminator.

Publication & Thesis Plan

- Write and submit a new paper targeting NAACL, aiming for the October ARR review cycle.
- Begin writing the MSc thesis, consolidating all completed work, and submit it.

APPENDIX

References

[1] Schlechtweg, D., Hätty, A., Del Tredici, M. and Schulte im Walde, S., 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. ACL 2019, pp.732-746.

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

[3] Jayatilleke, N. and De Silva, N., 2025. Zero-shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil. RANLP 2025, pp.471-480.

[4] Jayatilleke, N. et al., 2026. SiDiaC-v.2.0: Sinhala Diachronic Corpus Version 2.0. LREC 2026, pp.6740-6763.

[5] Hamilton, W.L., Leskovec, J. and Jurafsky, D., 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL 2016, pp.1489-1501.

[6] Frermann, L. and Lapata, M., 2016. A Bayesian Model of Diachronic Meaning Change. TACL, 4, pp.31-45.

[7] Rudolph, M. and Blei, D., 2017. Dynamic Bernoulli Embeddings for Language Evolution. arXiv:1703.08052.

[8] Cuba Gyllensten, A., Gogoulou, E., Ekgren, A. and Sahlgren, M., 2020. SenseCluster at SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. SemEval 2020, pp.112-118.

[9] Kiyama, H. et al., 2025. Analyzing Continuous Semantic Shifts with Diachronic Word Similarity Matrices. COLING 2025, pp.1613-1631.

[10] Ehrenworth, J. and Keith, K., 2023. Literary Intertextual Semantic Change Detection: Application and Motivation for Evaluating Models on Small Corpora. LChange 2023, pp.1-14.

[11] Montariol, S. and Allauzen, A., 2019. Empirical Study of Diachronic Word Embeddings for Scarce Data. RANLP 2019, pp.795-803.

[12] Kumarasinghe, K., Dias, G. and Herath, I., 2021. SinMorphy: A Morphological Analyzer for the Sinhala Language. MERCon 2021, pp.681-686.

[13] Fernando, S. and Ranathunga, S., 2018. Evaluation of different classifiers for Sinhala POS tagging. MERCon 2018, pp.96-101.

[14] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality. NeurIPS 2013.

[15] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2017. Enriching Word Vectors with Subword Information. TACL, 5, pp.135-146.

[16] Aravinda, H.W.K. et al., 2025. SinLlama: A Large Language Model for Sinhala. MERCon 2025, pp.617-622.

[17] Wang, L. et al., 2024. Improving Text Embeddings with Large Language Models. ACL 2024, pp.11897-11916.

[18] Shoemark, P. et al., 2019. Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. EMNLP-IJCNLP 2019, pp.66-76.

[19] Lasandi, M. and Jayatilleke, N., 2026. SinhaLegal: A Benchmark Corpus for Information Extraction and Analysis in Sinhala Legislative Texts. LoResLM 2026, pp.114-131.

[20] de Silva, N., 2025. Survey on Publicly Available Sinhala NLP Tools and Research. arXiv:1906.02358v25.

[21] Ranathunga, S. and De Silva, N., 2022. Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. AACL-IJCNLP 2022, pp.823-848.

Thank You!

Presented by Nevidu Jayatilleke

Next: From SinhaLegal to SiDiaC-v.3.0

SinhaLegal [19] (a dedicated Sinhala legal-text corpus, published as its own benchmark) is the domain-specific case study for this framework, rather than a subset carved out of SiDiaC itself.

~2M

Words in SinhaLegal, the legal-domain corpus where this framework has already been completed

180k+

Target words in SiDiaC-v.3.0: a much smaller, sparser corpus to adapt the method to next

Legal

Domain of the completed case study: Sinhala legislative texts, published as SinhaLegal

What may need to change

- SiDiaC-v.3.0 has substantially less data than SinhaLegal (~2 million words), so eligible substitution targets will be scarcer: the frequency threshold, pseudoword density, and eligibility rules used for SinhaLegal may all need re-tuning for SiDiaC's characteristics.
- SinhaLegal has already served as the domain-specific proof of concept for the synthetic ground-truth framework; the next step is validating the same framework directly on SiDiaC-v.3.0's mixed historical and literary text.
- Goal: use the artificial (ground-truth) datasets from both corpora to build a robust system that discriminates genuine semantic drift from non-drift, then extrapolate that discriminator to the real SiDiaC-v.3.0 and SinhaLegal corpora.

[19] Lasandi, M. and Jayatilleke, N., 2026. SinhaLegal: A Benchmark Corpus for Information Extraction and Analysis in Sinhala Legislative Texts. LoResLM 2026, pp.114-131.