Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# A Lemmatized Corpus of the Pali Canon with Critical Apparatus

**Dan Zigmond**

djz@shmonk.com

**Research Article**

**Keywords:** Pāli, corpus linguistics, lemmatization, morphological analysis, low-resource languages, language resources

**Posted Date:** October 5th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-8846980/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# A Lemmatized Corpus of the Pāli Canon with Critical Apparatus

**Dan Zigmond**

Received: date / Accepted: date

**Abstract** Pāli exemplifies a less-resourced language: despite a substantial textual corpus, computational infrastructure has lagged behind better-resourced classical languages like Sanskrit, Latin, and Ancient Greek. We present a lemmatized corpus of the Pāli Canon (Tipiṭaka), the canonical scripture of Theravāda Buddhism, a 1.6 million token corpus achieving 99.78% token-level and 97.8% unique-word lemmatization coverage through integration with the Digital Pāli Dictionary. Error analysis of unresolved forms reveals predictable categories: approximately half are potential dictionary additions, with sandhi compounds and metrical spelling variants comprising most of the remainder. The corpus enables previously infeasible computational tasks for Pāli: unsupervised text clustering and topic modeling on normalized vocabulary, training supervised morphological disambiguators, building syntactic treebanks following the Universal Dependencies framework, and conducting cross-lingual alignment with Sanskrit parallel texts. Unlike existing digital editions, this resource provides word-level lemmatization with sandhi decomposition for 42,449 compound forms, specialized handling for titles and metrical variants, and a critical apparatus recording textual variants across five independent witnesses representing the European critical, Burmese, Mahāsaṅgīti, Sinhalese, and Royal Thai editorial traditions. We compare coverage and design decisions against the Digital Corpus of Sanskrit, noting both similarities in scale and important differences in annotation depth and validation methodology. The corpus and reproducible pipeline are released under open licenses.

**Keywords** Pāli · corpus linguistics · lemmatization · morphological analysis · low-resource languages · language resources

D. Zigmond
Independent Scholar, Palo Alto, CA, USA
E-mail: djz@shmonk.com

# 1 Introduction

## 1.1 The Resource Gap for Classical Languages

Morphologically annotated corpora are foundational infrastructure for computational linguistics, yet coverage remains highly uneven across languages. While major modern languages benefit from million-token treebanks and pretrained models, classical and historical languages often lack even basic lemmatization, despite being essential for philological research, historical linguistics, and cultural heritage preservation.

Pāli exemplifies this resource gap. The Pāli Canon (Tipiṭaka), the foundational scripture of Theravāda Buddhism, is one of the oldest and most complete collections of early Buddhist texts. As a Middle Indo-Aryan language with approximately 2.7 million words of canonical text, Pāli presents a corpus comparable in size to Ancient Greek or Classical Latin. Yet while Sanskrit, Greek, and Latin have Universal Dependencies treebanks, morphological analyzers, and neural taggers, Pāli computational resources remain fragmented: digital editions exist but lack morphological annotation; dictionaries exist but lack corpus integration; morphological analyzers exist but lack corpus-scale coverage.

## 1.2 Computational Challenges of Pāli

Pāli presents specific challenges for NLP that distinguish it from both modern languages and some better-resourced classical languages:

**Rich morphology**: Nouns decline for eight cases and three numbers; verbs conjugate across multiple tenses, moods, voices, and persons. A single lemma may appear in dozens of surface forms. The noun *dhamma* (teaching) appears as *dhammo*, *dhammaṃ*, *dhammassa*, *dhammānāṃ*, *dhammesu*, and 15+ additional forms; the verb *karoti* (to do) has over 40 attested inflections.

**Pervasive sandhi**: Euphonic combination at word boundaries obscures lexical units. The surface form *etadavoca* combines pronoun *etad* + verb *avoca*; *bhaborūpaṃ* joins *bhava* + *arūpaṃ*. Without decomposition, such forms cannot be lemmatized.

**Orthographic variation**: Regional manuscript traditions (Sri Lankan, Burmese, Thai) exhibit systematic spelling differences, complicating cross-edition alignment. For example, the Burmese tradition writes *brahmā* where Sri Lankan manuscripts have *brāhmaṇa*; final vowels may be lengthened metrically (*bhikkhusatehī* vs. *bhikkhusatehi*); and anusvāra representation varies (*evaṃ* vs. *evam*).

**Limited training data**: Unlike Sanskrit, which benefits from the Digital Corpus of Sanskrit Hellwig (2010) and Sanskrit Heritage Engine Huet (2005); Goyal and Huet (2012), no corpus-scale annotated Pāli dataset has been available for machine learning.

### 1.3 Limitations of Existing Resources

Previous Pāli computational work has been constrained by resource availability. Alfter (2017) developed a finite-state morphological analyzer achieving high precision on in-vocabulary forms, but coverage of the full canon remained incomplete due to sandhi and out-of-vocabulary items. The Digital Pāli Reader Yuttadhammo (2013) provides dictionary lookup with sandhi analysis but does not offer downloadable corpus-wide annotation.

In earlier work, we demonstrated that frequency-based text mining could identify textual strata in the canon Zigmond (2021, 2023), but concluded that "the Pāli Canon in raw form is a poor foundation for this sort of textual analysis" due to the proliferation of inflected surface forms Zigmond (2023). A single lemma like *bhikkhu* (monk) appears in 270 distinct surface forms when counting declensions, compounds, and sandhi variants.

### 1.4 Contributions

This paper presents a lemmatized corpus that addresses these gaps. Our contributions provide new research capabilities for Pāli computational linguistics:

1. **Corpus-wide lemmatized analysis**: To our knowledge, the first corpus-wide morphological annotation of the Pāli Canon, enabling vocabulary studies, distributional semantics, and text classification at scale.
2. **Reproducible evaluation benchmark**: Systematic coverage statistics and error categorization that enable comparison with other classical language resources and tracking of future improvements.
3. **Adaptable pipeline**: A modular processing architecture applicable to other Middle Indo-Aryan languages (Prakrit, Buddhist Hybrid Sanskrit) sharing similar morphological features.
4. **NLP-ready critical apparatus**: Textual variant classification that supports annotation stability analysis and cross-tradition alignment experiments.

The primary purpose of this paper is to document and evaluate the shared corpus resource that enables these applications, rather than to exhaustively present each application domain.

## 2 Related Work

### 2.1 Sanskrit Computational Resources

The Digital Corpus of Sanskrit (DCS) Hellwig (2010) provides the closest comparison for our work. The DCS offers morphologically tagged text for over 650,000 sentences across 250+ texts, with lemmatization, POS tags, and dependency annotation. Recent extensions include the Sanskrit Sembank Hellwig

and Biagetti (2025), which adds lexical semantic annotation integrated with Princeton WordNet synsets.

Sanskrit word segmentation and morphological tagging have benefited from neural approaches. Hellwig (2016) applied sequence models to sentence boundary detection; Krishna et al. (2018) developed datasets and models for joint segmentation and morphological parsing. These methods achieve F1 scores above 90% on standard benchmarks, demonstrating that supervised learning is viable for morphologically rich classical languages given sufficient annotated data.

Our corpus aims to provide comparable infrastructure for Pāli, enabling similar supervised approaches.

### 2.2 Pāli Digital Editions

Five major digital editions of the Pāli Canon exist, each representing a distinct textual tradition:

**Pali Text Society (PTS/GRETIL)**: The Pali Text Society, founded in London in 1881, produced the first critical editions of the complete Canon based primarily on *Sri Lankan manuscripts*, with reference to Burmese and Thai sources where available. These editions, digitized by the Dhammakaya Foundation (1989–1996) and distributed through GRETIL, remain the standard scholarly citation reference but lack systematic annotation.

**Chaṭṭha Saṅgāyana (VRI/CST4)**: The Sixth Buddhist Council (1954–1956), convened in Rangoon under Burmese government sponsorship, produced a new recension based on the *Burmese manuscript tradition*. This edition, released digitally by the Vipassana Research Institute, is widely used in computational work due to early digital availability. It differs from PTS readings in hundreds of passages, reflecting the independent Burmese textual lineage.

**SuttaCentral (SC) Mahāsaṅgīti**: A digital edition of the Mahāsaṅgīti Tipiṭaka, a recension prepared in 1957 drawing on multiple manuscript traditions. This modern edition adds sentence-level segmentation with unique identifiers. While SuttaCentral has applied editorial normalization, its primary value for this project lies in providing sentence segmentation and an additional witness for cross-validation rather than representing an unmediated manuscript tradition.

**Buddha Jayanti Tipiṭaka (BJT)**: Published by the *Sri Lankan government* (1957–1989) to commemorate the 2,500th anniversary of the Buddha's passing, this edition represents a modern Sinhalese recension. The text was digitized by the Sri Lanka Tripitaka Project and is available in machine-readable form. As an independent Sinhalese editorial tradition distinct from the PTS critical editions, BJT provides an additional witness for collation.

**Thai Royal Edition (Syām Raṭṭha)**: Published under royal patronage beginning in 1893, the *Thai Royal Edition* represents an independent Southeast Asian recension. Digital versions are available through the E-Tipitaka

project. As a witness distinct from the SuttaCentral normalization of the Mahāsaṅgīti, the original Thai edition provides a fifth independent tradition for collation.

Table 1 explains the segment identifier format used throughout the corpus.

**Table 1** Segment Identifier Format

| Component | Meaning |
|---|---|
| `dn` | Collection (dn=Dīgha, mn=Majjhima, etc.) |
| `1` | Sutta number within collection |
| `:` | Separator |
| `1.1` | Section.sentence within sutta |
| *Examples:* | |
| `dn1:1.1` | DN sutta 1, section 1, sentence 1 |
| `mn22:13.4` | MN sutta 22, section 13, sentence 4 |
| `sn1.1:1.2` | SN saṃyutta 1, sutta 1, sect. 1, sent. 2 |
| `an1.1:1.1` | AN nipāta 1, sutta 1, sect. 1, sent. 1 |

The Digital Pāli Dictionary (DPD) Bodhirasa (2026) represents a breakthrough resource: 88,350 headwords with 1,275,089 inflected form mappings and sandhi decomposition for attested compounds. The DPD is actively maintained with monthly releases that expand coverage and correct errors; our corpus can be re-lemmatized against new DPD versions to benefit from ongoing improvements. Our corpus integrates the DPD to achieve corpus-wide annotation.

## 3 Corpus Construction

### 3.1 Source Integration

We collate the five editions described above, representing independent editorial traditions. All three piṭakas benefit from five-witness coverage:

- **PTS (base text)**: GRETIL transcriptions of the PTS critical editions serve as the authoritative reading, ensuring compatibility with standard scholarly citations.
- **VRI**: The Chaṭṭha Saṅgāyana provides an independent Burmese witness, enabling detection of tradition-specific readings and potential transcription errors.
- **SuttaCentral**: The Mahāsaṅgīti edition provides sentence-level segmentation and an additional witness for cross-validation.
- **BJT**: The Buddha Jayanti Tipiṭaka provides an independent Sinhalese witness.
- **Thai**: The Syām Raṭṭha Royal Thai Edition provides a fifth witness from the Southeast Asian manuscript tradition.

### 3.2 Preprocessing Pipeline

Source texts undergo normalization before lemmatization:

1. **Character normalization**: ṁ → ṃ; whitespace standardization
2. **Tokenization**: Regular expression matching Pāli orthography
3. **Case normalization**: Lowercase for dictionary lookup; original case preserved

### 3.3 Lemmatization Algorithm

For each token, we apply a cascaded lookup strategy:

1. **Direct lookup**: Query DPD inflected forms table (1,275,089 form mappings)
2. **Sandhi decomposition**: If DPD deconstructor provides analysis, recursively lemmatize components (42,449 compounds analyzed)
3. **Orthographic normalization**: Apply systematic patterns:
   - Final niggahīta: -n → -ṃ (1,555 forms)
   - Metrical lengthening: final ā→a, ī→i, ū→u (2,076 forms)
   - Internal metrical variants (78 forms)
4. **Particle splitting**: Separate trailing particles (ca, api, ti, va, tu) from sandhi compounds (96 forms)
5. **Pronoun-verb fusion**: Split patterns like *ahamanusāsissāmi* → *ahaṃ* + *anusāsissāmi* (28 forms)
6. **Compound decomposition**: Greedy longest-match with backtracking for long dvandva compounds (204 forms)
7. **Proper noun matching**: Query Dictionary of Pāli Proper Names Malalasekera (1938) with inflection handling (21 forms)
8. **Custom lemma database**: Fallback for words not in DPD, organized into potential DPD additions (73 entries), metrical variants (32), project-specific forms (42), and sandhi decompositions (29)
9. **Mark unresolved**: Flag remaining forms for manual review

The complete algorithmic specification, including regular expressions for orthographic normalization, backtracking conditions for compound decomposition, and the custom lemma database, is available in the documented source code at the repository listed in Section 6. Step ordering affects outcomes: earlier steps take precedence, so a form matching DPD directly will not undergo sandhi decomposition even if both analyses are valid.

### 3.4 Critical Apparatus Construction

Differences between editions are classified to support both philological analysis and annotation quality assessment:

- **Orthographic**: Predictable spelling differences (ṁ/ṃ, gemination)—normalized silently
- **Normalization**: Multiple witnesses (SC, VRI, BJT, Thai) agree against PTS where PTS form absent from DPD—normalized with original preserved in apparatus
- **Variant**: All forms are valid Pāli—recorded without normalization

Table 2 summarizes the collation results across the entire Tipiṭaka. Normalizations are applied only when multiple witnesses agree against PTS *and* the PTS reading is not attested in DPD. These represent editorial decisions rather than objective corrections, as some PTS readings may reflect legitimate regional or historical variants (see Section 7).

**Table 2** Five-Witness Collation Results across the Tipiṭaka

| Collection | Texts | Words | Normalizations | Variants |
|---|---|---|---|---|
| *Sutta Piṭaka* | | | | |
| DN | 34 | 144,000 | 4,355 | 9,135 |
| MN | 150 | 247,000 | 4,555 | 11,534 |
| SN | 1,564 | 265,000 | 15,617 | 10,411 |
| AN | 843 | 301,000 | 5,164 | 26,433 |
| KN | 17 | 640,000 | 8,532 | 10,683 |
| *Vinaya Piṭaka* | | | | |
| Vinaya | 3 | 421,000 | 4,580 | 5,000 |
| *Abhidhamma Piṭaka* | | | | |
| Abhidhamma | 3 | 820,000 | 5,145 | 5,409 |
| **Total** | | | **47,948** | **78,605** |

The normalizations fall into several categories, illustrated here with examples from the Dīgha Nikāya:

- **Anusvāra normalization**: *bhikkhūnam* → *bhikkhūnaṃ*; *evam* → *evaṃ*
- **Sandhi/particle additions**: *bhikkhusaṃghañ* → *bhikkhusaṃghañca*
- **Spelling corrections**: *icchānaṅkale* → *icchānaṅgale*
- **Retroflex consonants**: *khānumataṃ* → *khāṇumataṃ*
- **Vowel length**: *micchājivena* → *micchājīvena*

In addition to normalizations, the apparatus records **78,605 variant readings** across the Tipiṭaka where editions differ but all readings represent valid Pāli forms. A further 330,889 positions are flagged as uncertain, primarily in collections like Aṅguttara where heavy use of *peyyāla* (abbreviated repetitions) creates alignment challenges. This classification enables users to assess annotation stability: forms attested across all five traditions have higher confidence than forms appearing in only one witness.

The contribution of each witness to the apparatus—including tie-breaking behavior, error confirmation rates, and unique readings—is analyzed in detail in Section 4.

*Example: Five-Witness Apparatus*

To illustrate the apparatus in practice, Figure 1 shows a passage from the opening of the Brahmajālasutta (DN 1), the first sutta of the canon, in which the Buddha advises the monks not to feel elated when others praise the Triple Gem. Five loci in this brief passage illustrate both categories of our classification: three normalizations (where the PTS form is absent from DPD and corrected by other witnesses) and two variants (where all readings are valid Pāli forms reflecting independent textual traditions).

*Mamaṃ vā, bhikkhave, pare* **vāṇṇaṃ**[1] *bhāseyyuṃ, dhammassa vā vāṇṇaṃ bhāseyyuṃ, saṅghassa vā vāṇṇaṃ bhāseyyuṃ, tatra* **tumhe**[2] *na ānando na somanassaṃ na cetaso* **ubbillāvitattaṃ**[3] *karaṇīyaṃ.* [1.6.4:] *... tatra ce tumhe assatha ānandinō sumanā* **ubbillāvitā**[4] *tumhaṃ* **yev**[5] *assa tena antarāyo.*

---

[1] **vāṇṇaṃ**] *vaṇṇaṃ* SC VRI BJT Thai — NORMALIZATION
[2] **tumhe**] *tumhehi* SC VRI BJT; *tumehahi* Thai — VARIANT
[3] **ubbillāvitattaṃ**] *uppilāvitattaṃ* SC VRI; *ubbilāvitattaṃ* BJT Thai — VARIANT
[4] **ubbillāvitā**] *uppilāvitā* SC VRI; *ubbilāvino* BJT; *ubbilāvitattā* Thai — NORMALIZATION
[5] **yev**] *yevassa* SC VRI BJT — NORMALIZATION

**Fig. 1** Critical apparatus for DN 1:1.6.3–4 (D i 3). Translation: "But, monks, if others should speak in praise of me, of the Dhamma or of the Sangha, you should not on that account be pleased, happy or elated. If you were to be pleased, happy or elated at such praise, that would only be a hindrance to you" (Walshe, 1995, p. 68). Bold text follows PTS; superscript numbers key to apparatus entries below the rule. Each entry is classified as NORMALIZATION (PTS form absent from DPD) or VARIANT (all readings valid Pāli). Entry 4 shows a four-way split: PTS *ubbillāvitā*, SC/VRI *uppilāvitā*, BJT *ubbilāvino*, Thai *ubbilāvitattā*.

## 4 Evaluation

### 4.1 Coverage Statistics

Table 3 summarizes lemmatization results. We distinguish between *unique-word coverage* (97.8%)—the percentage of distinct word forms resolved—and *token-level coverage* (99.78%)—the percentage of all word occurrences in the corpus that are lemmatized. The difference arises from Zipfian vocabulary distribution: high-frequency words (particles, pronouns, common verbs) are well-covered by DPD, while unresolved forms are predominantly rare words (hapax legomena) from verse texts.

### 4.2 Error Analysis

The 2,756 unresolved unique forms (2.2% of vocabulary) represent only 3,578 token occurrences (0.22% of corpus), reflecting the Zipfian distribution where rare words contribute minimally to overall text. To understand failure modes,

**Table 3** Lemmatization Coverage

| Metric | Value |
|---|---|
| Total tokens | 1,606,474 |
| Unique word forms | 127,026 |
| Forms resolved | 124,270 (97.8%) |
| Forms unresolved | 2,756 (2.2%) |
| **Token-level coverage** | **99.78%** |
| Sandhi decompositions (DPD) | 42,449 |
| Orthographic normalizations | 1,555 |
| Metrical normalizations | 2,154 |
| Custom lemma entries | 176 |

we manually categorized a stratified sample of 496 unresolved forms (18% of unresolved vocabulary), selected to include high-frequency unresolved tokens and representatives from each nikāya. Table 4 presents this analysis; proportions in the full set of 2,756 forms may differ.

**Table 4** Unresolved Form Categories (n=496 sample)

| Category | Forms |
|---|---|
| Potential headwords (hapax legomena) | 249 |
| Sandhi endings (complex compounds) | 115 |
| Metrical variants (vowel changes) | 110 |
| Pronoun-verb fusions | 11 |
| Long compounds | 4 |
| Story titles (proper nouns) | 4 |
| Jhāna compounds | 3 |

The majority of unresolved forms are **potential headwords**—words that appear to be legitimate Pāli but are not in DPD, predominantly hapax legomena from verse sections of the Khuddaka Nikāya (Apadāna, Jātaka gāthās, Theragāthā/Therīgāthā). These texts exhibit greater lexical diversity due to poetic composition and include rare vocabulary (plant names, archaic terms, onomatopoeia) not found in prose suttas.

The remaining categories resist automated resolution for specific reasons: **sandhi endings** involve complex euphonic combinations that the DPD deconstructor does not handle (e.g., forms ending in truncated niggahīta like *pajan* for *pajaṃ*); **metrical variants** show vowel lengthening or shortening for poetic meter (e.g., *pavāressatī* for *pavāressati*); **pronoun-verb fusions** combine pronouns with verbs in ways not decomposed by DPD (e.g., *dubbalohamasmi* = *dubbalo* + *ahaṃ* + *asmi*); and **story titles** are compound proper nouns naming Jātaka or commentary narratives. Resolving these would require either manual lexicographic work (for legitimate words) or additional decomposition rules.

**Implications**: The token-level coverage (99.78%) indicates that dictionary lookup succeeds for the vast majority of tokens, though this metric does not

account for lemma ambiguity or potential lookup errors (see Section 7). The remaining gaps are concentrated in specialized vocabulary from verse texts. As DPD expands, coverage will improve; we have shared a list of 2,825 potential additions with the DPD maintainer.

### 4.3 Comparison with Sanskrit Resources

Table 5 compares our corpus with the Digital Corpus of Sanskrit.

**Table 5** Comparison with DCS

| | Pāli | DCS |
|---|---|---|
| Tokens | 1.6M | 4.5M |
| Unique lemmas | 39K | 90K+ |
| Coverage | 97.8% | 99%+ |
| Texts | 5,764 | 250+ |
| POS tags | No | Yes |
| Dependencies | No | Partial |

This comparison requires important caveats. First, the annotation types differ qualitatively: DCS provides human-verified POS tags and syntactic dependencies, while our corpus provides only lemmatization. Second, the coverage metrics are not directly comparable: DCS coverage reflects expert annotation accuracy, while ours measures dictionary lookup success. Third, DCS has benefited from 15+ years of iterative refinement with expert involvement. For the four main nikāyas (DN, MN, SN, AN), our lookup coverage approaches the DCS benchmark, with only 69 unresolved unique forms. Our corpus provides a foundation for similar iterative improvement for Pāli, but significant work remains to achieve comparable annotation depth.

### 4.4 Vocabulary Reduction

Lemmatization dramatically reduces vocabulary size, directly impacting downstream modeling. Table 6 shows the effect.

**Table 6** Vocabulary Reduction via Lemmatization

| | Surface | Lemmatized |
|---|---|---|
| Unique forms | 127,026 | 39,460 |
| Reduction ratio | – | 3.22× |

The 3.22× reduction means that statistical models trained on lemmatized text see over three times more examples per vocabulary item, improving gen-

eralization for rare words and making both unsupervised clustering and supervised neural approaches more feasible for this low-resource language. After lemmatization, the hundreds of inflected forms for common words (as noted in Section 1) consolidate under single dictionary headwords.

## 4.5 Witness Analysis

The five-witness collation annotated 2,880,314 word positions across the complete Tipiṭaka, of which 126,553 were classified as confident readings (47,948 normalizations and 78,605 textual variants).

### *4.5.1 The Three Core Witnesses (PTS, SC, VRI)*

Each witness makes distinct contributions to the apparatus. The PTS editions define the base text and provide the citation system (e.g., "D I 1") used throughout Western Buddhist studies; the 47,948 normalizations identified against PTS demonstrate that even long-established scholarly editions can benefit substantially from multi-witness collation. SuttaCentral contributes the segmentation framework—276,345 sentence-level segments with unique identifiers—that serves as the primary unit of analysis for the corpus. SC and VRI form the closest witness pair, both deriving from the Burmese Chaṭṭha Saṅgāyana tradition; their joint agreement against PTS was the initial basis for identifying normalizations, while their occasional divergences reflect SC's editorial corrections to the underlying Mahāsaṅgīti text.

### *4.5.2 Impact of the Fourth Witness (BJT)*

The addition of the Buddha Jayanti Tipiṭaka had a particularly dramatic effect on the Vinaya and Abhidhamma piṭakas, where it resolved a large number of previously uncertain readings. In the Vinaya, uncertain positions fell from 312,363 (three-witness) to 44,228 (four-witness)—a reduction of 86%. The Abhidhamma showed a similar pattern, with 15,055 uncertain readings resolved and 2,902 additional errors confidently identified. BJT is also the largest witness by word count (3,514,083 words), primarily because it fully expands *peyyāla* passages that other editions abbreviate; the size difference is most pronounced in the Saṃyutta Nikāya, where BJT contains 551,483 words compared to GRETIL's 279,549.

### *4.5.3 Impact of the Fifth Witness (Thai)*

The Thai Royal Edition was integrated using a GRETIL-anchored alignment algorithm, achieving 100% coverage across the complete Tipiṭaka. Table 7 shows Thai's agreement rates with the other witnesses at variant positions (themselves 2.7% of all word positions in the collation).

**Table 7** Thai Agreement Rates at Variant Positions

| Witness | Agreements | Rate |
|---|---|---|
| SC (Mahāsaṅgīti) | 53,827 | 57.7% |
| VRI (Chaṭṭha Saṅgāyana) | 41,328 | 44.3% |
| BJT (Buddha Jayanti) | 35,512 | 38.0% |
| GRETIL (PTS) | 16,758 | 18.0% |

Thai most frequently agrees with SC and VRI, both deriving from the Burmese Sixth Council tradition. The lower agreement with PTS is consistent with the Thai Royal Edition representing an independent Southeast Asian recension sharing more recent textual history with the Burmese tradition than with the European critical editions.

**Tie-breaking**: In 11,675 cases where the other four witnesses split into two equal camps (a 2–2 deadlock), Thai resolves the impasse 88.4% of the time (10,323 cases). Among resolved ties, Thai sides with PTS+BJT 57.4% of the time and with SC+VRI 42.6%. This asymmetry is noteworthy: although Thai agrees more often with SC overall, at positions where witnesses are evenly divided, Thai leans toward the PTS/Sri Lankan tradition—suggesting that the Thai edition may preserve more conservative readings at precisely the positions where the Burmese and European traditions have diverged.

**Error confirmation**: Thai confirms 23,298 normalizations to the PTS base text. When all four non-PTS witnesses agree that a PTS reading is not a valid Pāli form, the correction can be made with very high confidence. Across the canon, 81% of DN and MN normalizations are supported by all four non-PTS witnesses.

**Unique readings**: Thai provides a reading found in no other witness at 20,841 positions (22.3% of its participations). Many reflect orthographic differences, but some represent substantive textual variants. For example, at DN 1:1.6.4 (see Figure 1, entry 4), four witnesses give four distinct readings for the same word: PTS *ubbillāvitā* and SC/VRI *uppilāvitā* are past participles ("elated") with regional spelling differences, BJT *ubbilāvino* substitutes a derived adjective ("exultant ones"), and Thai *ubbilāvitattā* uses an abstract noun formation ("elatedness"). These are not mere orthographic variants but morphologically distinct formations of a single root, illustrating the organic variation that accumulates across centuries of manuscript transmission.

### *4.5.4 Witness Relationships*

The agreement data suggests a broad stemmatic structure. SC and VRI form the closest pair, both deriving from the Burmese Chaṭṭha Saṅgāyana tradition. GRETIL/PTS and BJT share some affinity as representatives of the Sri Lankan and European critical traditions. Thai occupies an intermediate position, agreeing with the Burmese-derived editions on most readings but siding with the Sri Lankan/European tradition on a significant fraction of contested

variants. These findings confirm the value of maintaining a five-witness apparatus: each edition makes a genuinely independent contribution, and no single witness is redundant.

## 5 Applications and Reuse

The corpus enables specific computational tasks previously infeasible for Pāli:

### 5.1 Unsupervised Text Analysis

Lemmatization dramatically improves unsupervised methods by reducing vocabulary sparsity. Earlier work applying hierarchical clustering to the raw canon was limited by the proliferation of surface forms: a single concept might be split across hundreds of inflected variants Zigmond (2021). With lemmatization, methods like topic modeling (LDA), document clustering, and distributional semantics operate on a vocabulary reduced by 3.22×, producing more interpretable and stable results. The vocabulary normalization also enables frequency-based identification of textual strata and formulaic patterns across the canon Zigmond (2023).

### 5.2 Training Morphological Disambiguators

The lemmatized corpus provides training data for supervised disambiguation. In NLP terminology, this is "silver-standard" data—automatically generated annotations that may contain errors, as opposed to "gold-standard" human-verified labels—but the 99.78% coverage rate suggests the error rate is low enough for practical use. A concrete application: the surface form *dhammassa* is ambiguous between genitive singular ("of the teaching") and dative singular ("for the teaching"). With 15,000+ occurrences in context, the corpus provides sufficient examples to train context-aware classifiers, following approaches successful for Sanskrit Krishna et al. (2018).

### 5.3 Cross-lingual Alignment with Sanskrit

Many Pāli texts have Sanskrit parallels (e.g., Pāli *Mahāparinibbāna Sutta* ↔ Sanskrit *Mahāparinirvāṇa Sūtra*). Lemmatization enables vocabulary-based alignment: the Pāli lemma *nibbāna* can be systematically mapped to Sanskrit *nirvāṇa*, supporting comparative studies of formulaic expressions and doctrinal terminology across Buddhist traditions.

### 5.4 Universal Dependencies Treebank Development

A treebank is a corpus annotated with syntactic structure, showing grammatical relationships between words (subject-verb, noun-modifier, etc.). The Universal Dependencies (UD) framework provides a cross-linguistically consistent annotation scheme used for over 100 languages.

The lemmatized corpus provides the morphological layer required for UD annotation. Pāli's relatively fixed word order and explicit case marking make it a tractable target for treebank development. The sentence segmentation from SuttaCentral provides natural annotation units averaging 12 tokens—comparable to UD treebank sentence lengths. A Pāli treebank would support both supervised parsing models and unsupervised grammar induction experiments.

## 6 Data Format and Availability

The corpus is released in JSON format with programmatic access via both Python and R.

### 6.1 Python Library

The `pali` Python library provides a high-level API:

```
from pali import Canon
canon = Canon()

# Extract raw Pali text
sutta = canon.get_sutta("dn1")
for seg in sutta.segments[:3]:
  print(seg.pali)
# "Evam me sutam-ekam samayam ..."

# Get lemmatized sutta
sutta = canon.get_sutta("dn1",
                        lemmatized=True)
for seg in sutta.segments:
  for tok in seg.tokens:
    print(f"{tok.word} -> {tok.lemma}")

# Search by lemma across corpus
results = canon.search_lemma("dhamma")

# Generate document-term matrix
dtm = canon.document_term_matrix("dn")
```

The API supports raw text extraction, corpus navigation, lemma search with full-text indexing, and vocabulary statistics. Sutta identifiers follow the format described in Table 1; a full list of valid identifiers is included in the package documentation. The processing pipeline comprises approximately 20,000 lines of Python across 49 modules.

### 6.2 R Package

For statistical analysis, the `tipitaka` R package (available on CRAN) provides the corpus in `tidy` data formats:

```
library(tipitaka)

# Lemma frequencies by nikaya
head(tipitaka_long_critical)
#  word      n    total   freq  book
#  dhamma 12847 304857 0.042    an

# Document-term matrix
tipitaka_wide_critical[1:5, 1:5]

# Pali-aware alphabetical sorting
pali_sort(c("nana", "nama", "n"))
```

The R package includes specialized functions for Pali alphabetical sorting (implemented in C++ for performance), stop word lists, and pre-computed frequency matrices at both nikāya and sutta granularity.

### 6.3 Print Edition Preparation

The Python library also includes LaTeX export for preparing traditional printed critical editions. Suttas can be typeset with critical apparatus, variant readings in footnotes, and PTS page references:

```
# Export sutta as LaTeX with apparatus
canon.export_latex("dn1",
  output="dn1_critical.tex",
  include_variants=True,
  include_refs=True)
```

This enables scholars to produce print-ready editions that combine the computational benefits of the lemmatized corpus with the traditional format expected by philological scholarship.

### 6.4 Availability

The corpus and Python library are available at `https://github.com/dangerzig/pali-canon` under CC-BY-SA 4.0. The R package is available at `https://cran.r-project.org/package=tipitaka`.

## 7 Limitations and Future Work

**Coverage metrics reflect dictionary lookup, not verified accuracy**: The 99.78% token-level coverage measures successful DPD lookup, not independently verified lemmatization accuracy. Since we validate against the same

resource used for lemmatization, this is inherently circular. A gold-standard evaluation against expert human annotation would provide a more rigorous accuracy measure; we identify this as essential future work. Users should interpret coverage figures as upper bounds on accuracy.

**Significant lemma ambiguity**: An estimated 15–20% of successfully resolved tokens have alternative lemmatizations not recorded in our output. When multiple lemmas match a surface form, we select the first DPD entry—an arbitrary choice based on DPD's internal ordering rather than corpus frequency or contextual disambiguation. This affects downstream analyses: topic models, frequency counts, and vocabulary statistics all inherit this ambiguity. Context-aware disambiguation using distributional methods or neural sequence labeling remains future work.

**Editorial normalizations, not objective corrections**: The 47,948 normalizations where PTS readings were changed based on agreement among other witnesses represent *editorial decisions*, not objective corrections. The criterion "PTS form absent from DPD" conflates dictionary coverage with textual authenticity; a form may be absent from DPD because it represents a legitimate rare variant, regional orthographic convention, or scribal tradition with historical validity. Users conducting philological research may prefer to consult the original PTS readings preserved in our apparatus rather than the normalized forms; the Python API supports extracting text with or without normalizations applied.

**No POS tags or syntactic annotation**: Unlike the Digital Corpus of Sanskrit, which provides human-verified POS tags and partial dependency annotation, our corpus includes only lemmatization. The DCS comparison in Table 5 should be interpreted accordingly: while token counts and coverage percentages are comparable, the qualitative depth of annotation differs substantially. Building comparable syntactic annotation for Pāli remains future work.

**Sandhi decomposition unverified**: The 42,449 sandhi decompositions are taken directly from DPD without independent verification. Sandhi analysis is notoriously difficult, and some decompositions may be linguistically questionable.

**Version dependency**: The corpus depends on a specific DPD release (version as of February 2026). As DPD updates monthly, users should note the version used; re-lemmatization against newer DPD versions may yield different results. We intend to rebuild the corpus as DPD improves.

## 8 Conclusion

We have presented a lemmatized corpus of the Pāli Canon that provides a foundation for computational work on this less-resourced classical language. Dictionary lookup succeeds for 97.8% of unique words (99.78% of tokens), though users should note the significant caveats detailed in Section 7: coverage metrics reflect lookup success rather than verified accuracy, 15–20% of tokens

have unrecorded alternative lemmatizations, and editorial normalizations represent interpretive decisions rather than objective corrections. POS tagging and dependency annotation remain future work, as does gold-standard evaluation against expert human annotation. Despite these limitations, the corpus enables vocabulary normalization for unsupervised text analysis and provides training signal for supervised approaches. By releasing the corpus and pipeline under open licenses, we aim to support iterative community improvement following the model established by the Digital Corpus of Sanskrit.

## Declarations

Funding

No funding was received for conducting this study.

Competing Interests

The author has no competing interests to declare that are relevant to the content of this article.

Data Availability

The lemmatized corpus and processing pipeline are openly available at `https://github.com/dangerzig/pali-canon` under the CC-BY-SA 4.0 license. The `tipitaka` R package is available on CRAN at `https://cran.r-project.org/package=tipitaka`. The source editions used (PTS/GRETIL, VRI Chaṭṭha Saṅgāyana, SuttaCentral Mahāsaṅgīti, Buddha Jayanti Tipiṭaka, Thai Royal Edition) are publicly available from their respective distributors. The Digital Pāli Dictionary is available at `https://digitalpalidictionary.github.io/`.

**Acknowledgements** This work builds on the contributions of multiple digital Pāli initiatives. The Pali Text Society editions, digitized by the Dhammakaya Foundation and distributed through GRETIL, provided the base text. The Vipassana Research Institute (VRI) released the Chaṭṭha Saṅgāyana Tipiṭaka following the Sixth Buddhist Council, providing an independent Burmese textual witness. SuttaCentral contributed the Mahāsaṅgīti edition with sentence-level segmentation and reference IDs. The Buddha Jayanti Tipiṭaka, digitized by the Sri Lanka Tripitaka Project, provided a Sinhalese witness, and the Thai Royal Edition (Syām Raṭṭha) provided a fifth independent tradition. The Digital Pāli Dictionary (DPD), developed by Bhikkhu Bodhirasa, supplied the morphological database including 88,000+ headwords and 1.2 million inflected form mappings that made corpus-wide lemmatization possible. The Python library (∼20,000 lines across 49 modules) and the `tipitaka` R package (available on CRAN) were developed with coding assistance from Claude Anthropic (2024).

## References

Alfter, D. (2017). Morphological analyser and generator for Pāli. Master's thesis, University of Gothenburg.

Anthropic (2024). Claude.

Bodhirasa (2026). Digital Pāli dictionary. Version 0.3.20260202.

Goyal, P. and Huet, G. (2012). Completeness analysis of a Sanskrit reader. In *Proceedings of the 5th International Sanskrit Computational Linguistics Symposium*, pages 130–171, Mumbai.

Hellwig, O. (2010). Performance of a lexical and POS tagger for Sanskrit. *Lecture Notes in Computer Science*, 6465:162–172.

Hellwig, O. (2016). Detecting sentence boundaries in Sanskrit texts. In *Proceedings of COLING 2016*, pages 288–297, Osaka.

Hellwig, O. and Biagetti, E. (2025). The Sanskrit sembank. *Language Resources and Evaluation*, 59:3635–3658.

Huet, G. (2005). A functional toolkit for morphological and phonological processing, application to a Sanskrit tagger. *Journal of Functional Programming*, 15(4):573–614.

Krishna, A., Satuluri, P., and Goyal, P. (2018). A dataset for Sanskrit word segmentation. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, pages 105–114, Santa Fe.

Malalasekera, G. (1937–1938). *Dictionary of Pāli Proper Names*. Pali Text Society, London. 2 vols.

Walshe, M. (1995). *The Long Discourses of the Buddha: A Translation of the Dīgha Nikāya*. Wisdom Publications, Boston. Originally published 1987.

Yuttadhammo, B. (2013). Digital Pāli reader. Version 4.0.

Zigmond, D. (2021). Toward a computational analysis of the pali canon. *Journal of the Oxford Centre for Buddhist Studies*, 20:133–152.

Zigmond, D. (2023). Distinguishing commentary from canon in early buddhist texts using computational linguistics. In *Proceedings of the 18th World Sanskrit Conference*, Canberra.