भारत सरकार | Government of India | शिक्षा मंत्रालय | Ministry of Education

शिक्षा मंत्रालय
MINISTRY OF EDUCATION

# CENTRAL INSTITUTE OF CLASSICAL TAMIL

## செம்மொழித் தமிழாய்வு மத்திய நிறுவனம்

AN AUTONOMOUS INSTITUTION UNDER THE MINISTRY OF EDUCATION, GOVERNMENT OF INDIA
CHEMMOZHI SALAI, PERUMBAKKAM, CHENNAI - 600100

# DIGITAL ARCHIVES
# எண்ணிம ஆவணகம்

DIGITAL ARCHIVES OF CLASSICAL TAMIL
https://www.digitalarchives.cict.in/

SCAN ME

அனைத்து வளங்களும் இணையத்தளத்தின் மூலம் திறந்த அணுகலுடன் வழங்கப்படுகின்றன. திருக்குறள் Ground Truth Corpus சரிபார்க்கப்பட்ட எழுத்துப்பெயர்ப்புகள் நிறைவுபெறும் போதெல்லாம் படிப்படியாக வெளியிடப்படுகிறது (ஒவ்வொரு ஓலை ஏட்டிற்கும் தனித்தனி Zenodo DOI-களுடன், CC BY 4.0 உரிமத்தின் கீழ்). புதியதாகச் சரிபார்க்கப்படும் சுவடி எழுத்துப்பெயர்ப்புகள் மற்றும் தொடர்புடைய ஆய்வுவளங்கள் தாமதமின்றி அறிஞர்களுக்கும் பொதுமக்களுக்கும் சென்றடைவதை உறுதி செய்யும் வகையில், இம்மின்னணுக் காப்பகம் தினந்தோறும் துரித மற்றும் இயக்கநிலை (Mission Mode) அடிப்படையில் தொடர்ந்து புதுப்பிக்கப்பட்டு வருகிறது.

| 13 | 80K+ | 6.28M |
|---|---|---|
| வளங்கள் | தலைச்சொற்கள் | உரை எழுத்துகள் |

| # | எண்ணிமக் கருவி | விவரம் | அளவு |
|---|---|---|---|
| 1 | **திருக்குறள் மூலத் தொகுப்பு**<br>Tirukkural Ground Truth Corpus<br>**மூலபாடச் சேகரம் · தமிழ் *HTR* மாதிரிகள்** | தமிழ்க் கையெழுத்து உரை அறிதலுக்கான பயிற்சி மாதிரிகளாக, சரிபார்க்கப்பட்ட திருக்குறள் ஓலை எழுத்தாக்கங்கள் படிப்படியாக வெளியீடு. ஒவ்வோர் ஓலையும் எழுத்தர் வாசிப்பைச் செம்பதிப்புப் பாடத்துடன் இணைத்து, எழுத்தியல் குறிப்புகளில் வேறுபாடுகளைப் பதிவு செய்து, 30 மொழிகளில் மொழிபெயர்ப்பும் ஒலியும், IIIF ஆழ விரிவாக்கமும், HTR-தயார் பதிவிறக்கமும் வழங்குகிறது. | 64 ஓலைகள் · 640 / 1,330 குறள்கள் (48.1%) · Zenodo DOI |
| 2 | **சங்க இலக்கியச் சொல்லடைவு**<br>**Index of the Sangam Classic Corpus**<br>**சொல்லடைவு** | தொல்காப்பியம், எட்டுத்தொகை, பத்துப்பாட்டின் ஒவ்வொரு சொல்லுக்கும் தேடக்கூடிய அடைவு; எந்தவொரு சங்கச் சொல்லிற்கும் மூலத்துடன் சொற்பயன்பாட்டு எண்ணிக்கையையும் தருகிறது. | தொல்காப்பியம், எட்டுத்தொகை + பத்துப்பாட்டு |
| 3 | **செந்தமிழ்ச் சொற்பொருட் களஞ்சியம்**<br>Comprehensive Digital Encyclopaedia of Classical Tamil **எண்ணிமப் பேரகராதி** | சொற்பிறப்பு, பொருள் பரப்பு, சங்க மற்றும் சங்கம் மருவிய இலக்கியங்களில் ஆட்சி, மேற்கோள் ஆய்வு ஆகியவற்றைத் தரும் முதன்மை அகராதிக் குறிப்பு நூல். | 6,758 சொற்பதிவுகள் |
| 4 | **தமிழ்க் களஞ்சியம் (ஒலிவடிவம்)**<br>Tamil Audio Encyclopaedia (MTL, 1924–1939)<br>ஒலி அகராதி | சென்னைப் பல்கலைக்கழகத் தமிழ்ப் பேரகராதியின் தேடக்கூடிய இணைய ஒலி வடிவிலானப் பதிப்பு; 5,547 வடசொற்கள் சொற்பிறப்பியல் குறிப்புகளுடன். | 59,170 சொற்பதிவுகள் · 55,831 தலைச்சொற்கள் |
| 5 | **பிங்கல நிகண்டு**<br>Piṅkala Nigaṇṭu · Digital Edition<br>தொன்மையான நிகண்டு | பிங்கல முனிவர் இயற்றிய செவ்வியல் நிகண்டின் இணையப் பதிப்பு — கருத்து-வரிசை அமைப்பு; அரிய ஒத்த சொற்களையும் கவித்துவச் சொல்லாட்சிகளையும் பேணுகிறது. | கருத்து-வரிசை அமைப்பு |
| 6 | **சேந்தன் திவாகரம்**<br>Sēndan Tivākaram<br>பழமையான நிகண்டு *(கி.பி. 8–9 நூ.)* | திவாகர முனிவர் இயற்றிய திவாகர நிகண்டின் இணையப் பதிப்பு — தமிழில் தோன்றிய பழமையான, தலைசிறந்த நிகண்டு; நிகண்டு மரபின் ஊற்றுக்கண். | தமிழின் முதல் நிகண்டு |
| 7 | **தமிழ் இலக்கணப் பேரகராதி**<br>Tamil Grammatical Lexicon<br>இலக்கணப் பேரகராதி | தி. வே. கோபாலையரின் பதினேழு தொகுதி ஆய்வின் இணையப் பதிப்பு — தொல்காப்பியம், நன்னூல், வீரசோழியம், உரை மரபு முழுவதும். | 14,252 பதிவுகள் · 8,613 தலைச்சொற்கள் · 6.28M எழுத்துகள் |
| 8 | **திருக்குறள் தொடரடைவு**<br>Digital Concordance of the Tirukkural<br>**தொடரடைவு** | 1,330 குறள்களுக்கும் முழுமையான சொல்வாரி அடைவு — சொல், தொடரியல், நடை ஆய்வுகளுக்கு KWIC மரபில். | 4,898 தலைச்சொற்கள் · 11,467 சொற்பயன்பாடுகள் |
| 9 | **தண்டியலங்காரம்**<br>Tantiyalaṅkāram · Grammar of Poetical Ornaments<br>**அணியிலக்கணம் *(கு. 12-ஆம் நூ.)*** | அணிகள் பற்றிய அடிப்படை நூலின் இணையப் பதிப்பு, முழுமையான தொடரடைவுடன். | 126 சூத்திரங்கள் · 3 படலங்கள் · 98 பிரிவுகள் |
| 10 | **நன்னூல்**<br>Naṉṉūl<br>**செவ்வியல் இலக்கணம் *(12–13 நூ.)*** | மூலத்துடன் மயிலைநாதர் உரையும் விரிவான விருத்தியுரையும் — பிற்காலத் தமிழின் செந்தர-இலக்கணம் - பவணந்தியாரால் எழுதப்பட்டது. | மூலம் + 2 உரைகள் |
| 11 | **இலக்கணச் சுருக்கம்**<br>Ilakkaṇac Curukkam<br>**இலக்கணத் தொகுப்பு *(19-ஆம் நூ.)*** | ஆறுமுக நாவலரின் சுருக்கமான இலக்கணம் — எழுத்து, சொல், தொடரியல்; தமிழ் உரைநடை-அச்சு மரபின் முன்னோடியின் ஆக்கம். | எழுத்து, சொல், தொடரியல் |
| 12 | **41 செவ்வியல் தமிழ் நூல்களின் முதற்பதிப்புகள் First Editions of Forty-One Classical Tamil Texts**<br>**செவ்வியல் மூலப்பனுவல் கருவூலம் · எண்ணிமக் காப்பகம்** | மூலப்பனுவல் தொகுப்பு — தொல்காப்பியம், பதினெண்மேல்கணக்கு, பதினெண்கீழ்க்கணக்கு, இரட்டைக் காப்பியங்கள் — ஒரே எண்ணிமக் காப்பகமாக. | 41 செவ்வியல் தமிழ் நூல்களின் மூலப்பனுவல் தொகுப்பு |
| 13 | **பாட்டும் தொகையும்**<br>Pāṭṭum Tokaiyum<br>**செவ்விலக்கியக் காப்பகம்** | சங்கச் செவ்விலக்கியப் பாடல்களின் எண்ணிமக் காப்பகம் — அடிப்படைச் சேகரத்தைப் பாடலும் உரையுமாக வழங்குகிறது. | சங்கச் சேகரம் · பாடல் + உரை |

**உருவாக்கம்:**
**செம்மொழித் தமிழாய்வு மத்திய நிறுவனம்**

भारत सरकार | Government of India | शिक्षा मंत्रालय | Ministry of Education

# CENTRAL INSTITUTE OF CLASSICAL TAMIL

## செம்மொழித் தமிழாய்வு மத்திய நிறுவனம்

AN AUTONOMOUS INSTITUTION UNDER THE MINISTRY OF EDUCATION, GOVERNMENT OF INDIA
CHEMMOZHI SALAI, PERUMBAKKAM, CHENNAI - 600100

शिक्षा मंत्रालय
MINISTRY OF
EDUCATION

# DIGITAL ARCHIVES

# எண்ணிம ஆவணகம்

DIGITAL ARCHIVES OF CLASSICAL TAMIL
https://www.digitalarchives.cict.in/

SCAN ME

"All resources are openly accessible through the portal. The Tirukkural Ground Truth Corpus is published progressively as verified transcriptions are completed (per-leaf Zenodo DOIs, CC BY 4.0). The digital archive is updated daily on an accelerated and mission-mode basis, ensuring the timely dissemination of newly verified manuscript transcriptions and related scholarly resources."

| 13 RESOURCES | 80K+ HEADWORDS | 6.28M CHARACTERS OF EXEGESIS |
|---|---|---|

| # | Tool | Description | Scale |
|---|---|---|---|
| 1 | திருக்குறள் மூலத் தொகுப்பு<br>Tirukkural Ground Truth Corpus<br>*Ground Truth Corpus · Tamil HTR Specimens* | Verified palm-leaf transcriptions of the Tirukkural, released progressively as ground-truth training specimens for Tamil handwritten-text recognition. Each folio pairs the verbatim scribal reading with the critical-edition text, documents variants in paleographic notes, and carries translations and audio across 30 languages, with IIIF deep zoom and HTR-ready export. | 64 leaves · 640/ 1,330 kurals (48.1%) · Zenodo DOIs |
| 2 | சங்க இலக்கியச் சொல்லடைவு<br>Index of the Sangam Classic Corpus<br>*Concordance* | Searchable concordance of every word across the Eight Anthologies and Ten Idylls; traces any Sangam-era term to its source verse, with corpus frequency and semantic neighbourhoods. | Tolkappiyam + Eight Anthologies + Ten Idylls |
| 3 | செந்தமிழ்ச் சொற்பொருட் களஞ்சியம்<br>Comprehensive Digital Encyclopaedia of Classical Tamil *Digital Encyclopaedia* | Definitive lexicographical reference giving etymology, semantic range, attested usage across the Sangam and post-Sangam corpora, and citational scholarship. | 6,758 entries |
| 4 | தமிழ்க் களஞ்சியம் (ஒலிவடிவம்)<br>Tamil Lexicon: Audio Format (MTL, 1924–1939)<br>*Audio Dictionary* | Searchable digital audio edition of the Madras University Tamil Lexicon; 5,547 entries of Sanskrit origin (vaṭacol) are annotated for etymological provenance. | 59,170 entries · 55,831 headwords |
| 5 | பிங்கல நிகண்டு<br>Piṅkala Nigaṇṭu · Digital Edition<br>*Classical Thesaurus* | Digital edition of the classical thesaurus by Piṅkalar, organized thematically; preserves rare synonyms, regional vocabulary, and poetic registers for reading Sangam literature. | Thematically organized |
| 6 | சேந்தன் திவாகரம்<br>Sēndan Tivākaram<br>*Earliest Nikaṇṭu (8th–9th c. CE)* | Digital edition of the Tivākara Nikaṇṭu — the earliest and most distinguished thesaurus-lexicon to appear in Tamil, composed by Divākara Munivar; the fountainhead of the nikaṇṭu tradition. | Earliest Tamil nikaṇṭu |
| 7 | தமிழ் இலக்கணப் பேரகராதி<br>Tamil Grammatical Lexicon<br>*Grammatical Lexicon* | Digital edition of T. V. Gopal Iyer's seventeen-volume survey, traversing the Tolkāppiyam, Naṉṉūl, Vīracōḻiyam, and the full commentarial tradition. | 14,252 entries · 8,613 headwords · 6.28M chars |
| 8 | திருக்குறள் தொடரடைவு<br>Digital Concordance of the Tirukkural<br>*Concordance* | Complete word-by-word index to the 1,330 kurals of Tiruvaḷḷuvar, aligned in the keyword-in-context tradition for lexical, syntactic, and stylistic analysis. | 4,898 headwords · 11,467 occurrences |
| 9 | தண்டியலங்காரம்<br>Tantiyalaṅkāram · Grammar of Poetical Ornaments<br>*Alaṅkāra Treatise (c. 12th c.)* | Digital edition of the foundational twelfth-century treatise on poetical ornaments (alaṅkāra), with a complete concordance. | 126 sūtras · 3 paṭalams · 98 sections |
| 10 | நன்னூல்<br>Naṉṉūl<br>*Classical Grammar (12th–13th c.)* | The mūlam presented with the celebrated Mayilainātar commentary and the fuller Viruttiyurai — the standard grammar of later Tamil, composed by the Pavaṇanti Munivar. | mūlam + 2 commentaries |
| 11 | இலக்கணச் சுருக்கம்<br>Ilakkaṇac Curukkam<br>*A concise Tamil grammar (19th c.)* | Ārumuka Nāvalar's concise grammar — orthography, morphology, and syntax — by the Jaffna Śaiva scholar and founding figure of Tamil prose and printing. | Orthography · Morphology · Syntax |
| 12 | 41 செவ்வியல் தமிழ் நூல்களின் முதற்பதிப்புகள்<br>First Editions of Forty-One Classical Tamil Texts *Digital Archive* | The foundational corpus — Tolkāppiyam, the Caṅkam anthologies and idylls (Patiṉeṉmēlkaṇakku), the eighteen lesser texts (Patiṉeṉkīḻkaṇakku), and the twin epics — as one digital archive. | First Editions of Classical Tamil Literary Corpus (41 Texts) |
| 13 | பாட்டும் தொகையும்<br>Pāṭṭum Tokaiyum<br>*Classical Poetry Archive* | Digital archive of Classical Tamil (Caṅkam) poetry — 'the Songs and the Anthologies' — presenting the foundational corpus in verse with commentary. | Sangam corpus · verse + commentary |

Developed by
CENTRAL INSTITUTE OF CLASSICAL TAMIL

भारत सरकार | Government of India | शिक्षा मंत्रालय | Ministry of Education

CENTRAL INSTITUTE OF CLASSICAL TAMIL

செம்மொழித் தமிழாய்வு மத்திய நிறுவனம்

AN AUTONOMOUS INSTITUTION UNDER THE MINISTRY OF EDUCATION, GOVERNMENT OF INDIA

CHEMMOZHI SALAI, PERUMBAKKAM, CHENNAI - 600100

शिक्षा मंत्रालय MINISTRY OF EDUCATION

# STATUS REPORT
# நிலை அறிக்கை

DIGITAL ARCHIVES OF CLASSICAL TAMIL
https://www.digitalarchives.cict.in/

SCAN ME

All resources are provided with open access through the website. The Thirukkural Ground Truth Corpus is released progressively whenever verified transcriptions are completed (with separate Zenodo DOIs for each palm-leaf manuscript, under a CC BY 4.0 license). To ensure that newly verified manuscript transcriptions and related research resources reach scholars and the general public without delay, this digital archive is continuously updated daily on a rapid and Mission Mode basis.

## Thirukkural Palm-leaf Manuscript Ground Truth (GT) Collection · 19 June 2026

| | | | | |
|---|---|---|---|---|
| **64**<br>GT specimens (GT-001–064) | **640 / 1,330**<br>kurals verified 48.1% | **30**<br>translation languages | **12**<br>CICT lexical-tool resources | **920**<br>digital-library items |

### Corpus Coverage

| Division | Chapters | Kurals | Status |
|---|---|---|---|
| Aṟattuppāl (Virtue) · அறத்துப்பால் | 38 / 38 | 380 | Complete |
| Poruṭpāl (Wealth) · பொருட்பால் | 26 / 70 | 260 | In progress |
| — Araciyal (Statecraft) sub-section | 20 / 25 | — | 80% covered |
| Kāmattuppāl (Love) · காமத்துப்பால் | 0 / 25 | — | Pending |
| **Total** | **64 / 133** | **640 / 1,330** | **48.10%** |

### Codicology & Paleography — a single scribal hand

The physical-leaf model holds across 16 consecutive recto/verso pairs (leaves 34–49): every recto bears its Tamil leaf numeral (௩௰௪ → ௪௰௯, unbroken), every verso none; chapter titles occupy 2–3 left-margin cells split by visual convenience (canonical title in metadata, scribal form verbatim). Folio verse order is scrambled and recorded line-by-line from curator-supplied kural numbers. Next predicted: GT-059 (Oṟṟāṭal), recto of leaf 50 — the round numeral ௫௰.

- **ந/ன substitution** — pervasive in body and titles, in both directions (மந்நவன்; title சிற்றிநம் vs body விநை under a canonical title)
- **ழ/ள merger** — bidirectional, even within one line (வாளும்/வாழுங், K542); ழ்+stop → ட்+stop hardening (வாட்கை, இகட்சி)
- **Old-Tamil notation preserved** — āytam (ஃ), alapeḍai (தழீஇ), grantha (க்ஷ); sandhi-assimilated junctions written solid (நிலற்கு, சிதையாமற்க்)

- **IIIF deep zoom** — OpenSeadragon viewer, clickable line bands; automatic fallback while Zenodo tiles process
- **Paleographic layer** — 5 toggleable highlights with live counts, grapheme-cluster safe
- **Corpus search (KWIC)** — all lines indexed; click-to-open with the hit highlighted in the transcription
- **HTR-ready export** — line-level JSONL (IIIF crops) + PageXML ZIP for Kraken / eScriptorium / TrOCR
- **30-language reader** — full Tirukkuṟaḷ translations.
- **Tamil TTS** — per-kural audio readback
- **Bilingual UI** — English/Tamil, command palette; 12 lexical-tool cards
- **Scholarly apparatus** — editorial-committee attribution, per-leaf Zenodo DOIs, CC-BY 4.0