

AthDGC: An Open Diachronic Greek Treebank with Indo-European Parallels

Nikolaos Lavidas¹ Kiki Nikiforidou¹ Dag Haug²
Leonid Kulikov³ Vassiliki Geka¹ Vassileios Symeonidis¹
Theodoros Michalareas¹ Sofia Chionidi¹ Anastasia Tsiropina¹
Eleni Plakoutsi¹ Evangelos Argyropoulos¹

¹*National and Kapodistrian University of Athens*

²*University of Oslo*

³*Ghent University*

14 June 2026

AthDGC (“Athens-PROIEL”) is an open, end-to-end workflow and dataset. It provides an openly licensed dependency-parsed treebank (a treebank is a collection of sentences whose grammatical structure has been analysed and stored in a machine-readable form) of Greek that spans eight diachronic periods, namely Archaic, Classical, Koine, Late Antique, Byzantine, Late Byzantine, Early Modern, and Modern Greek, under a single PROIEL XML 2.0 schema (the file format developed at Oslo for storing each sentence as a tree of word-by-word grammatical relations), with verse-level cross-alignment of the New Testament to Latin (Vulgate), Gothic (Wulfila), Old Church Slavonic (Marianus), and Classical Armenian. AthDGC builds on the PROIEL Treebank Family (Haug and Jøhndal 2008, 27–34; Eckhoff et al. 2018, 29–65), which established the schema and the Koine-Greek reference set for the project. Annotation uses the Stanford Stanza PROIEL-trained workflow; sentence-level alignment uses LaBSE, a multilingual sentence-embedding model; word-level alignment uses multilingual-BERT attention through the AwesomeAlign procedure. The v0.4 release provides curated samples and the open-source toolkit; the full annotated corpus partitions remain under v0.5 audit on the Greek national HPC. Quantitative scale, per-witness verse counts, and per-period annotated-row counts are reported as the v0.5 release notes, after the audit pass completes.

Please cite as:

Lavidas, Nikolaos, Kiki Nikiforidou, Dag Haug, Leonid Kulikov, Vassiliki Geka, Vassileios Symeonidis, Theodoros Michalareas, Sofia Chionidi, Anastasia Tsiropina, Eleni Plakoutsi, and Evangelos Argyropoulos. 2026. *AthDGC: An Open Diachronic Greek Treebank with Indo-European Parallels*.

1. Overview

1.1 Context and motivation

Greek is the longest continuously attested member of the Indo-European family, with a written record that spans roughly three millennia, from the Mycenaean tablets and the Homeric epics through the Classical and Hellenistic literature, the Koine of the New Testament and the Septuagint, the homiletic and hymnographic prose of the Late Antique and Byzantine periods, the increasingly vernacular registers of Late Byzantine and Early Modern Greek, and the standard Modern Greek that emerged after the orthographic reform of the late twentieth century. This continuity makes Greek a unique observatory for diachronic syntactic change: the same canonical text, whether the Iliad, the New Testament, the Septuagint Psalms, or a classical historiographic passage, is re-rendered into Greek again and again across periods. The structural choices made by each generation of readers and translators leave a recoverable trace in the syntax of the resulting text.

The PROIEL Treebank (Haug and Jøhndal 2008, 27–34; Eckhoff et al. 2018, 29–65) established the gold standard for syntactically annotated parallel Indo-European Bible corpora. PROIEL (the dependency-treebank standard for early Indo-European languages, developed at the University of Oslo) introduced a relation inventory specifically designed for the morphological richness and the free word order of the older Indo-European languages, and provided the Koine-Greek anchor sentence set against which any later diachronic-Greek effort is naturally measured. Its scope, however, is restricted to the New Testament and to a small number of language witnesses; it does not by design extend to Archaic, Classical, Byzantine, or Modern Greek, and it does not by design cover the cross-lingual alignment of the Bible to language families more distant than the four PROIEL witnesses (Greek, Latin, Gothic, Old Church Slavonic).

Accordingly, no openly licensed, end-to-end workflow currently covers the full diachronic record of Greek at the same level of annotation, with explicit argument-structure tagging (the recovery of who does what to whom for every verb in the corpus), and with reproducible cross-lingual alignment (the machine-readable matching of corresponding words and sentences between texts in different languages) to its sister Indo-European languages. AthDGC closes that gap. AthDGC is the Athens node of the PROIEL family: we adopt the PROIEL XML 2.0 schema verbatim, we publish in that schema only, and we extend it diachronically across the eight periods of Greek. Dag Haug, founding director of the PROIEL project at Oslo, is a member of the AthDGC team and a co-author of this report; the formal designation “Athens-PROIEL” in the title block records this affiliation.

The project’s specific scholarly focus is retranslation: the same canonical text is re-rendered into Greek across periods (Homeric, Koine, Byzantine, Modern) and into sister languages (Latin, Gothic, Old Church Slavonic, Classical Armenian) across the Indo-European family. The platform records these re-renderings as a retelling and retranslation chain (see §3.3 below), a machine-queryable

structure in which the same passage is annotated for syntax, argument structure, and morphology at every node of the chain, so that the diachronic researcher can ask, with a single query, which structural patterns persist from Homer to the Kakridis-Kazantzakis prose Iliad of 1955, and which collapse under the Byzantine epitome of Tzetzes.

1.2 Repository location

The public showcase of the platform is hosted at <https://athdgc.github.io>, and the v0.4.0 source-code snapshot is permanently deposited on Zenodo (an open-access research-data repository operated by CERN, which mints persistent DOIs for each deposit) under concept DOI 10.5281/zenodo.20439182. A concept DOI (a Digital Object Identifier that always resolves to the latest version of a deposited record) is the citation handle that does not change as the project versions advance; per-version DOIs are minted automatically on each release. The source repository AthDGC/Diachronic-Linguistics-Platform on GitHub remains private during the v0.5 audit pass, and becomes public at the v0.5 release, when the consolidated tools/ directory and the verified annotated partitions are in place.

1.3 Project resources (live)

Beyond the public showcase site at <https://athdgc.github.io>, the project maintains three working resources that track the day-by-day state of the corpus and the manual annotation pass.

The first is the AthDGC corpus repository, a closed-access GitHub repository at <https://github.com/AthDGC/athdgc>, the corpus that is the canonical Git-managed working location of the AthDGC corpus across the eight Greek diachronic periods and the four Indo-European parallels (with Sanskrit, Old English, Avestan, Old Persian, and Ukrainian queued at v0.7). Access is restricted to the eleven-member team during the v0.4 to v0.5 audit pass; the public release of the annotated partitions follows at v0.5 as a separate Zenodo dataset record under CC-BY-4.0.

The second is the per-language unique-identifier and metadata register, an Excel-format Google Sheet at <https://docs.google.com/spreadsheets/d/1MiXcxAedaHgdnj62q-zTQsS3Q4iGTyQU/edit?gid=99793223>, with one sheet per language (grc for Greek, lat for Latin, got for Gothic, chu for Old Church Slavonic, xcl for Classical Armenian, plus the queued IE witnesses). For each sentence in the corpus the register holds the unique identifier, the period assignment, the source archive, the edition, the licence, and the ingestion timestamp; this register is the bridge between the file-level partitions in the corpus repository and the per-sentence annotations in the PROIEL XML 2.0 layer.

The third is the live PROIEL annotation interface at <https://dialing.enl.uoa.gr/proiel/>, hosted at the National and Kapodistrian University of Athens (Division of Language-Linguistics, Department of English Language and Literature, School of Philosophy). This is the PROIEL XML 2.0 annotation web platform on which the AthDGC team performs and reviews the manual layer of PROIEL annotations on top of the Stanford Stanza pre-annotation pass, and on which the showcase samples published at <https://athdgc.github.io> are queried.

1.4 Institutional context

AthDGC is developed at the National and Kapodistrian University of Athens (NKUA), Division of Language-Linguistics, Department of English Language and Literature, School of Philosophy, under the direction of Prof. Nikolaos Lavidas. Compute is supplied by GRNET ARIS, the Greek national high-performance computing cluster (a shared cluster of fast machines operated by the Greek Research and Technology Network for academic computational workloads), under allocation pa260305.

1.5 Origin of the diachronic-Greek PROIEL line, Thessaloniki and Oslo, 2012

The diachronic-Greek PROIEL line that AthDGC continues began in 2012 as a Thessaloniki and Oslo collaboration between Prof. Dag Trygve Truslew Haug (University of Oslo, PROIEL Project Director) and Nikolaos Lavidas, then at the University of Thessaloniki. The first joint anchor text was George Sphrantzes, *Chronicon Sive Minus* (Chronicles, post-1453, ed. Grecu 1966), the principal historiographic account of the Fall of Constantinople and the only post-Koine, Late-Byzantine Greek text in the original PROIEL release series. The annotated edition was published in PROIEL Release 20180408 under CC-BY-NC-SA 4.0, with principal investigators Dag Trygve Truslew Haug and Nikolaos Lavidas, and funding from the University of Oslo and the University of Thessaloniki. The annotation and review team comprised Þorsteinn Vilhjálmsson, Anastasia Michali, Maria Geramani, Evgenia Klidona, Athina Papadopoulou, and Dag Haug. The release is available at the PROIEL Treebank GitHub repository at <https://github.com/proiel/proiel-treebank> and is browsable sentence by sentence at <https://syntacticus.org> under source identifier proiel:20180408:chron (for example, <https://syntacticus.org/sentence/proiel:20180408:chron:89063>). The Oslo side of the collaboration is housed in the Foni research group in linguistics (Forskergruppe i lingvistikk) at the Department of Philosophy, Classics, History of Art and Ideas, University of Oslo.

The Sphrantzes Chronicle is therefore the historical hinge between the original PROIEL programme and AthDGC. PROIEL covered Greek up to the Koine of the New Testament, with Sphrantzes as the lone post-Koine, Late-Byzantine extension. AthDGC takes that extension as a starting point and carries it through the entire Greek diachronic span (Archaic, Classical, Koine, Late Antique, Byzantine, Late Byzantine, Early Modern, Modern), under the same PROIEL XML 2.0 schema and the same relation inventory. The 2012 Thessaloniki and Oslo collaboration thus continues today, with the same PROIEL Project Director (Prof. Dag T. T. Haug, Oslo) and the same Greek PI (Prof. Nikolaos Lavidas), now at the National and Kapodistrian University of Athens, funded by HFRI Project No. 20577 and the Greece 2.0 National Recovery and Resilience Plan, and supported by GRNET ARIS allocation pa260305.

2. Method

The AthDGC workflow runs in six stages: discovery, filtering, conversion, annotation, argument-structure capture, and cross-lingual alignment. We describe each stage in turn and illustrate it with

a worked example carried through the opening line of the Iliad, μήνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος (“Wrath sing, goddess, of Peleides Achilles”), so that the reader can follow the same sentence from raw input to a fully annotated and cross-aligned record.

2.1 Discovery

Source material reaches AthDGC through three complementary channels, of which the first runs daily and the other two are continuous on a slower cadence.

The first channel is the daily online-archive harvest. Each day, the discovery stage probes online open-access repositories for new or updated Greek and parallel-language source material. The repositories currently probed include archive.org, the Perseus Digital Library (the standard online archive of classical-Greek and Latin texts maintained by Tufts University), the Open Greek and Latin First Thousand Years of Greek collection at Leipzig, Wikisource Greek (el.wikisource.org), the Diorisis Corpus (the lemmatised classical-Greek corpus by Vatri and McGillivray), and OpenGreekAnd-Latin / PerseusDL on GitHub. For the parallel languages, the discovery stage probes the Wulfila Project at the University of Antwerp (Gothic), TITUS at the University of Frankfurt (Old Church Slavonic and Classical Armenian), GRETEL at Goettingen (Sanskrit), TEAMS at the Robbins Library (Old English), and the National Library of Ukraine facsimile collection.

The second channel is in-house digitisation, namely the OCR (Optical Character Recognition, the conversion of scanned page images into machine-readable text) of out-of-copyright printed editions held by NKUA and partner institutions. This channel covers pre-1928 Teubner editions, the Patrologia Graeca volumes not yet available in clean text form, scanned Early Modern Greek printed books from the Anemi collection at the University of Crete, and equivalent printed material from collaborating Departments of Classical Studies in the CIVIS Alliance and at other Greek universities. Manuscript facsimiles supplied by project partners enter through the same channel and are flagged in the corpus metadata with the supplying institution and the corresponding editorial provenance.

The third channel is community contribution, namely the submission of prepared text or annotation by external researchers and by KEDIVIM continuing-education-course participants (see §5). Such contributions are versioned in the public GitHub repository as pull requests, pass the same PROIEL XML 2.0 schema validator and house-style check that govern internal contributions (see §3.4), and on acceptance receive co-authorship acknowledgement on the next release notes.

For the Iliad worked example, the discovery stage retrieves the Homeric text from the Perseus Digital Library through its CTS URN (Canonical Text Services Uniform Resource Name, a standard identifier scheme used by classical-text repositories for referring to a specific edition of a specific work), urn:cts:greekLit:tlg0012.tlg001, together with the Kakridis-Kazantzakis 1955 prose translation, mirrored on Wikisource, and the Tzetzes Byzantine epitome from Bibliotheca Augustana.

2.2 Filtering

Candidate texts are filtered by Greek-script ratio (at least 75% of alphabetic characters in the Greek Unicode blocks U+0370-U+03FF and U+1F00-U+1FFF), by a Path-B line filter (an in-house regular-expression filter that removes bilingual editorial apparatus such as Latin headers, page-number lines, and critical-edition footnote markers from the textual stream), and by content-hash deduplication (an SHA-256 hash of the normalised text is checked against a registry of previously ingested files, so that identical files discovered at multiple mirrors are ingested only once). A candidate that fails any of the three filters is rejected at this stage and logged for review; the v0.5 release notes will report per-archive rejection statistics.

2.3 Conversion

Surviving text is converted into the PROIEL XML 2.0 schema with sentence-level structure. The PROIEL XML 2.0 schema is XML (a plain-text format that stores structured information in tagged elements) with each sentence wrapped in a `<s>` element, each token wrapped in a `<token>` element, and every token carrying attributes for its form, lemma, part-of-speech tag (`pos`), ten-character morphological tag (`morphology`), syntactic head identifier (`head-id`), and dependency relation to that head (`relation`). For the worked example, the conversion stage produces a containing five tokens for `μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος` and writes the initial sentence header together with the empty token-attribute scaffolding that the next stage fills in.

2.4 Annotation

Annotation is performed sentence by sentence with the Stanford Stanza processor (Qi et al. 2020, 101–8). Stanza (Stanford’s open-source Python toolkit for automatic linguistic annotation, which performs tokenisation, lemmatisation, part-of-speech tagging, morphological analysis, and dependency parsing in a single workflow) provides PROIEL-trained models for several of the older Indo-European languages: `grc_proiel` for Koine Greek, `la_proiel` for Latin, `cu_proiel` for Old Church Slavonic, and `got_proiel` for Gothic. Classical Armenian is annotated through a PROIEL-style workflow currently under development; the v0.4 release reports the workflow configuration only, and the first audited Armenian alignment is planned for the v0.5 release (end of September 2026), using a fine-tuned Stanza model (a fine-tuned model is one that has been further trained on a smaller, target-specific dataset on top of its general-purpose training, so that it adapts to the particular text type in question) trained on TITUS Armenian material. The Stanza output for each sentence is then normalised to the PROIEL XML 2.0 schema; AthDGC publishes PROIEL XML 2.0 only, with no CoNLL-U (the line-by-line tab-separated format used by the Universal Dependencies project, with one token per line and one column per annotation attribute) or other format offered, since the analytical workflow downstream of the corpus (argument-structure extraction, alignment graph, classifier features) depends on the PROIEL relation inventory. For the worked example, Stanza assigns `μῆνιν` the lemma `μῆνις`, part-of-speech `Nb`, morphology `Nb-s—fa-` (a noun, singular, feminine, accusative), head identifier 2, and relation `obj` (direct object). The verb `ἄειδε`

receives lemma ἀείδω, part-of-speech V-, morphology V-spia— (verb, second person singular, present, imperative, active), head identifier 0 (sentence root), and relation pred (predicate).

2.5 Argument-structure capture

Beyond standard dependency annotation, AthDGC extracts an explicit argument-structure frame for every verb token, using the PROIEL relation inventory strictly. The argument-structure frame records, for one verb, the syntactic and semantic role of each of its arguments: the subject (sub, including the raised patterns xobj for raising objects and nonsub for non-subject controllers), the direct object (obj), the indirect and oblique arguments (iobj for indirect object, obl for oblique), the vocative addressee (voc), the voice (active, middle, or passive), and the aspect (perfective or imperfective). For the worked example, the verb ἀείδε receives the frame [sub:imp.2sg, obj:μῆνιν(acc.sg.f), voc:θεᾶ(voc.sg.f), voice:active, aspect:imperfective]. This frame is the unit on which downstream queries operate: a diachronic researcher can ask, for instance, across the Iliad reception, does the verb of singing retain its direct-object case-marking through the Byzantine epitome and the Modern Greek prose translation, or does it switch to a prepositional phrase? and receive an answer that does not require manual re-reading of the texts.

2.6 Cross-lingual alignment

Sentence-level alignment uses LaBSE embeddings (Feng et al. 2022, 878–91). LaBSE (Language-Agnostic BERT Sentence Embedding, a multilingual sentence-embedding model that maps sentences from over 100 languages into a single 1,500-dimensional vector space, so that translation pairs land close to each other regardless of the source language) provides the similarity score used to match a Greek New Testament verse to its Latin Vulgate, Gothic Wulfila, Old Church Slavonic Marianus, or Classical Armenian counterpart. Word-level alignment uses multilingual-BERT attention through the AwesomeAlign procedure (Dou and Neubig 2021, 2112–28). Multilingual BERT (Google’s neural language model pre-trained on text from 104 languages) emits cross-attention weights between the tokens of the source and target sentences; AwesomeAlign converts these weights into a discrete word-alignment matrix through a fine-tuning step on parallel corpora. Phonetic cognate scoring, finally, uses ASJP sound-class encoding (the Automated Similarity Judgment Program sound-class system, which reduces every phonetic segment to one of 41 equivalence classes that ignore fine phonetic detail) and LingPy edit distance (List 2014, 1–228) for detection of cognate pairs across the Indo-European witnesses.

3. Dataset description

Field	Value
Object name	AthDGC corpus, v0.4 release

Field	Value
Format	PROIEL XML 2.0; JSONL partitions; Neo4j alignment-graph dump; Qdrant vector store
Creation dates	2025-09 onwards; v0.4.0 minted 29 May 2026
Dataset creator	Lavidas, N., Nikiforidou, K., Haug, D., Kulikov, L., Geka, V., Symeonidis, V., Michalareas, T., Chionidi, S., Tsiropina, A., Plakoutsi, E., and Argyropoulos, E.
Languages	grc (Ancient Greek), gkm (Medieval Greek), ell (Modern Greek), with parallels in lat, got, chu, xcl; queued for v0.7: san, ang, ave, peo, ukr
Licence	Code Apache-2.0; metadata and alignments CC-BY-4.0; per-source raw text under its original licence
Repository	AthDGC/Diachronic-Linguistics-Platform on GitHub (private during v0.5 audit; public at v0.5)
Concept DOI	10.5281/zenodo.20439182
Publication date	2026-05-29 (v0.4.0 source-code snapshot on Zenodo)
Release status	v0.4 is samples-only on the public site; the full annotated partitions are under audit on GRNET ARIS and release at v0.5

3.1 Open-access source provenance

Every primary source text in AthDGC is open-access (public domain, CC-BY, CC-BY-SA, or equivalent). The annotation layer is AthDGC-original and released under CC-BY-4.0. Accordingly, the open-access chain is preserved from input through annotation to distribution.

Greek, per-period source map

Period	Source archive	Licence
Archaic, Classical, Hellenistic	Perseus Digital Library; Open Greek and Latin / First Thousand Years of Greek (Leipzig)	CC-BY-SA 4.0
Koine NT	SBL Greek NT; Tischendorf (1869); Westcott-Hort (1881)	SBL licence; public domain

Period	Source archive	Licence
Koine LXX	Rahlfs (1935) via openscriptures.org	public domain
Koine, documentary	Papyri.info / DDbDP	CC-BY 3.0
Late Antique, Byzantine	Patrologia Graeca via Documenta Catholica Omnia and Internet Archive	public domain
Byzantine, Late Byzantine	Bibliotheca Augustana; pre-1928 Teubner	public domain
Early Modern	Anemi (University of Crete); Wikisource el	public domain
Modern (19th c. to 1928)	Wikisource el; Anemi	public domain
Modern (post-1928)	publisher-licensed editions; not republished	in copyright

Indo-European parallels, per-language source map

Language	Source archive	Licence
Latin (Vulgate)	Clementine Vulgate via Vulsearch; Latin Library	public domain
Gothic (Wulfila)	Wulfila Project (University of Antwerp)	CC-BY-SA
Old Church Slavonic (Marianus)	TITUS (Frankfurt)	academic open access
Classical Armenian	TITUS; Digilib Armenian	academic open access
Sanskrit (Brahmana, Upanisadic)	GRETIL (Goettingen); SARIT; TITUS	academic open access
Old English (Wessex Gospels)	TEAMS; Bosworth-Toller; DOE extracts	public domain; CC-BY-SA
Avestan (Yasna, Yashts)	TITUS; Geldner (1886-95)	public domain; open access
Old Persian (Behistun)	TITUS; Kent (1953) PD transliteration	public domain
Ukrainian (Ostroh 1581)	National Library of Ukraine facsimile	public domain
Ukrainian (20th-c. revision)	Ohienko (1962); Khomenko (1963)	in copyright

In-copyright editions are never republished verbatim. Accordingly, AthDGC uses an open-access antecedent (for example, a pre-1928 Teubner edition, or the SBL Greek New Testament instead of the Nestle-Aland critical edition) or, where this is unavoidable, short quotation samples under fair

use with full attribution. The annotation layer is always AthDGC-original and released under CC-BY-4.0, so that the open-access chain is preserved from the input texts through the syntactic and morphological annotation to the public deposition.

3.2 Indo-European parallels covered

Language	Family	Period	Stanza model	Edition coverage	Status
Greek (Koine NT)	IE / Hellenic	1st-2nd c. AD	grc_proiel	SBL GNT (complete)	annotated
Latin (Vulgate)	IE / Italic	4th c. AD	la_proiel	Clementine Vulgate NT (complete)	annotated
Gothic (Wulfila)	IE / Germanic E	4th c. AD	got_proiel	Wulfila Project (Gospels and Pauline corpus, partial)	annotated
OCS (Marianus)	IE / Slavic	10th c. AD	cu_proiel	Codex Marianus (four Gospels)	sampled
Classical Armenian	IE / Armenian	5th c. AD	xcl_proiel (in dev.)	TITUS Armenian NT	ingestion (v0.5)
Sanskrit (Brahmana, Upanisadic)	IE / Indo-Iranian	1000-500 BC	sa_vedic (queued)	GRETIL / SARIT	queued (v0.7)
Old English (Wessex Gospels)	IE / Germanic W	10th c. AD	ang_proiel (queued)	TEAMS Wessex Gospels	queued (v0.7)
Avestan (Yasna, Yashts)	IE / Indo-Iranian	1000-500 BC	ae_proiel (queued)	TITUS Yasna and Yashts	queued (v0.7)
Old Persian (Behistun)	IE / Indo-Iranian	6th-5th c. BC	peo_proiel (queued)	TITUS Behistun	queued (v0.7)
Ukrainian (Ostroh and 20th-c. rev.)	IE / Slavic E	1581 and 1962	uk_dep adapted (queued)	Ostroh 1581 and Ohienko 1962	queued (v0.7)

Verified per-witness aligned-verse counts will be reported in the v0.5 release notes (planned end

of September 2026) once the ARIS audit pass closes. Accordingly, the v0.4 release provides the curated samples on the Samples page rather than full per-witness counts.

3.3 Retelling and retranslation chains: a worked case study

A retelling and retranslation chain is the central analytical structure of AthDGC. It is a directed sequence of nodes, where each node is the same canonical passage as it appears in a different period or a different language, annotated end-to-end in the PROIEL XML 2.0 schema, and connected by edges that record the type of relationship between adjacent nodes (retelling within Greek across periods, retranslation within Greek across registers, or retranslation cross-lingually across the Indo-European family). The chain is the unit on which the platform's diachronic and comparative queries operate.

Five chains are public on the Samples page at v0.4, namely the Iliad 1.1 across reception (Homer, Tzetzes, Kazantzakis), the Septuagint Psalm 1:1 across Greek periods, the New Testament John 1:1 patristic and Modern reception, the Plato Apology 17a across reception, and the Septuagint Genesis 1:1 across four Indo-European languages. We present the Iliad 1.1 chain in detail because it is the most familiar to the comparative linguist and because it shows the analytical pay-off of the chain structure most clearly.

The first node of the chain is the Homeric original, μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος. Under AthDGC annotation, μῆνιν is the object (obj, accusative singular feminine), ἄειδε the predicate (imperative second person singular, active, imperfective), θεὰ the vocative addressee (voc), and the genitive pair Πηληϊάδεω Ἀχιλῆος a coordinated attribute (atr) hanging from the object noun. The argument-structure frame of the verb is [sub:imp.2sg, obj:μῆνιν, voc:θεὰ, voice:active, aspect:imperfective], and the frame signature [obj:acc, voc:voc] is recorded as the chain's characteristic frame against which subsequent nodes are compared.

The second node of the chain is the Byzantine epitome of Tzetzes (twelfth century), in which the same opening is reframed as a third-person prose narrative. Under AthDGC annotation, μῆνιν survives as a direct object, but the imperative-and-vocative dialogic frame collapses: the goddess is no longer addressed, and the verb of singing is replaced by a verb of narration. The characteristic frame [obj:acc, voc:voc] therefore loses one slot and the chain records a frame collapse between Homer and Tzetzes, of the kind that typically signals a generic reframing rather than a syntactic change in Greek per se.

The third node is the Kakridis-Kazantzakis prose translation of 1955, the canonical Modern Greek prose rendering of the Iliad. Under AthDGC annotation, the wrath remains a direct object in the Modern Greek accusative, the vocative of the goddess is restored, and the verb of singing returns. The characteristic frame [obj:acc, voc:voc] is therefore stable across 2,800 years in the direct Homer-to-Modern path; the frame collapse at the Tzetzes node is shown to be a feature of the Byzantine epitomising genre, not of the Greek language. This is the kind of finding that the chain structure makes immediate: without the chain, the reader would need to read each text manually and reason

about the structural correspondence; with the chain, the platform produces the result in a single query of the form frame-stability(Iliad 1.1, Homer, Modern).

The Septuagint Psalm 1:1 retranslation chain shows the inverse pattern, namely the stability of pred-nom + sub + atr:rel-cl across the Septuagint, Byzantine, and Modern Greek nodes with two voice flips between the active and the mediopassive on the verb of walking, together with the lexical replacement of the Greek negation οὐκ by οὐ and then by δευ. The John 1:1 patristic and Modern chain shows the stability of the copular skeleton across 2,000 years, with the patristic nodes nesting under the xcomp of φησὶν (“he says”), which is a structural reflex of the Late-Antique commentary register. The Plato Apology 17a chain shows the preservation of the pred + adv:neg + comp[sub, obl] pattern across the Plato-to-Modern path, with the perfect aspect alternating between the synthetic Classical form and the analytic Modern Greek auxiliary plus participle. The Septuagint Genesis 1:1 cross-IE chain shows the stability of the pred + sub + obl/adv + obj + coord pattern across the Greek, Latin Vulgate, Gothic Wulfila, and Old Church Slavonic Marianus nodes, with only the obl and adv slots diverging in the OCS rendering, a divergence that reflects the well-known Slavic preference for adverbial expression where Greek and Latin use an oblique noun.

The retelling and retranslation chain is, in this respect, the central contribution of AthDGC to diachronic-Greek and Indo-European comparative syntax: it is the data structure that makes the diachronic stability of a syntactic pattern computationally observable.

3.4 Tools (open source)

The AthDGC toolkit is organised as a set of OSI-licensed modules in three readiness tiers. Five modules are LIVE in the public repository and can be cloned and run today. They are the Quarto pack that builds this Working Paper and the public site, the showcase generator that produces the browsable per-sample PROIEL trees, the corpus-fix toolkit that bulk-corrects annotation errors at scale, the PROIEL XML 2.0 exporter that runs inside the build workflow, and the Stanza annotation job script for the GRNET ARIS allocation. Eight modules are IN SETUP and become available with the v0.5 release, namely the PROIEL XML 2.0 schema validator (which enforces the relation inventory and rejects UD-style labels such as nsubj and dobj; UD here is Universal Dependencies, the alternative treebanking framework whose label inventory is incompatible with PROIEL), the house-style check, the argument-frame extractor, the Stanza fine-tuning scripts, the athdgc-tools PyPI package, and the three Hugging Face model repositories whose weights upload at v0.5. The remaining modules are FORTHCOMING at v0.5: the LightSIDE-AthDGC syntactic-feature fork and its export scripts, the NoSketch-style CWB indexer, the Neo4j alignment viewer, the valency-frame database client, the retranslation-pair browser, and the retelling-chain explorer. The complete tier matrix per module is on the public Tools page at <https://athdgc.github.io/tools.html> and on the Status page at <https://athdgc.github.io/status.html>.

Across the toolkit, the annotation layer and the corpus samples are CC-BY-4.0; the tool code is Apache-2.0 (most modules), MIT (the Quarto pack and the Neo4j viewer), BSD-3-Clause (inherited from the upstream LightSIDE project), and GPL-3.0 (the upstream CWB indexer, retained on the index builder).

Two automated gates run on every community pull request, namely a PROIEL XML 2.0 schema validator (which rejects UD-style labels such as `nsubj`, `dobj`, `nmod`, `case`) and a house-style check that enforces the project style guide.

4. Related work and positioning

AthDGC sits in an ecosystem of openly licensed treebanks for older Indo-European languages, and we briefly describe where it differs from and where it complements each of them.

The PROIEL Treebank (Haug and Jøhndal 2008, 27–34; Eckhoff et al. 2018, 29–65), as discussed in §1.1, established the schema and the Koine-Greek anchor against which AthDGC is naturally measured. AthDGC adopts the PROIEL XML 2.0 schema verbatim (and publishes in that schema only), so that any sentence in AthDGC can be read by any tool already in the PROIEL ecosystem. The complementarity is that PROIEL covers the New Testament across four IE witnesses at a high annotation density, whereas AthDGC covers the entire diachronic span of Greek across eight periods and extends the cross-lingual alignment with the Classical Armenian witness now, and with Sanskrit, Old English, Avestan, Old Persian, and Ukrainian at v0.7.

The GLAUx Corpus (the Greek Language Automated, a large lemmatised classical-Greek corpus released by Toon Van Hal and collaborators) and the Diorisis Corpus (the lemmatised classical-Greek corpus by Vatri and McGillivray) provide lemmatised classical-Greek text at scale, but they do not provide a dependency-syntactic annotation layer and do not provide cross-lingual alignment. AthDGC re-uses the lemmatisation effort of both projects where their licences permit and adds the syntactic-annotation layer on top.

The Ancient Greek Dependency Treebank (AGDT, the Perseus dependency treebank of classical and post-classical Greek) and the Pedalion treebank (the Leuven treebank of classical-Greek student-annotation material) provide dependency annotation for selected classical-Greek texts, but under different relation inventories from PROIEL (AGDT uses a hybrid Latin Dependency Treebank schema, Pedalion uses a pedagogical schema), and they do not provide cross-lingual alignment either. AthDGC publishes PROIEL XML 2.0 only, and we therefore do not re-ingest AGDT or Pedalion material at the syntactic-annotation layer; we treat them as comparison points for diachronic claims, not as input.

In broad terms, AthDGC’s contribution is therefore three-fold. First, the extension of PROIEL-schema syntactic annotation to the full diachronic span of Greek, from Archaic to Modern, under one schema. AthDGC is, in this respect, the Athens node of the PROIEL family: we adopt the schema verbatim and we extend it diachronically. Secondly, the systematic inclusion of retelling and retranslation chains as a first-class data structure, so that the diachronic stability of a syntactic pattern becomes a computable quantity. Thirdly, the extension of the cross-lingual alignment beyond the four PROIEL witnesses, with the Classical Armenian alignment planned for the v0.5 release (end of September 2026) and an Old English partition already present in the corpus at v0.4. The longer-horizon IE expansion (Sanskrit, Avestan, Old Persian, Ukrainian) is reported on the public Status page rather than here, since it is forthcoming rather than released.

5. Reuse potential

AthDGC supports four primary reuse classes, which we describe in turn with concrete query examples.

In the first place, historical syntactic research. Per-period dependency parses allow analysts to track argument-structure changes across three millennia of Greek. Examples include the active-to-mediopassive shifts of certain verbs of motion, the loss of the optative across the Hellenistic period, the rise of the periphrastic perfect over the synthetic perfect across the Byzantine and Early Modern periods, the accusative-to-genitive variation under verbs of perception across periods, and the gradual replacement of the postnominal genitive of possession by the prenominal genitive in Modern Greek. Accordingly, the colour-coded compact dependency overview on each sample page is designed for fast scanning of these patterns. A worked query example is for the verb ἀκούω (“hear”), is the object more often in the accusative or in the genitive at each period?, which the valency-frame database (v0.5) answers in a single line.

Secondly, Indo-European comparative syntax. The verse-level cross-lingual alignment of the New Testament to Latin (Vulgate), Gothic (Wulfila), and Old Church Slavonic (Marianus) provides aligned PROIEL parses across three sister Indo-European languages in v0.4, with Classical Armenian planned for v0.5 (end of September 2026). An Old English partition is present at v0.4; the further IE expansion is reported on the public Status page. In this respect, users can query the Neo4j alignment graph for typological correspondences. A worked query example is every Greek aorist active verb whose Latin Vulgate counterpart is passive, which is expressed in Cypher (Cypher is the query language of Neo4j, structurally analogous to SQL for relational databases, in which MATCH patterns are drawn directly with arrows between nodes) as `MATCH (gk:Token {lang:“grc”, tense:“aor”, voice:“act”})-[:TRANSLATED_AS]->(la:Token {lang:“lat”, voice:“pass”}) RETURN gk, la`, and which is answered across the entire NT-aligned partition in a single pass.

Thirdly, computational diachronic NLP. The fine-tuned Stanza checkpoints (a checkpoint is a saved snapshot of a trained model’s numeric parameters, which can be loaded back into Stanza to parse new sentences without re-training) (`grc_byz_proiel` for Byzantine Greek, `grc_lbem_proiel` for Late Byzantine and Early Modern Greek, `grc_mod_proiel` for Modern Greek) are trained on ARIS; the Hugging Face model repositories under AthDGC/* exist now as private during the v0.5 audit pass and become public at v0.5 with the weight payloads (the trained parameter files that the checkpoints comprise) uploaded. In broad terms, the PyPI package `athdgc-tools` reserves the name with the stub `0.4.0.dev0`, and the full toolkit replaces the stub at v0.5; the checkpoints can be loaded directly into Stanza for parsing new Greek diachronic data once the weights are published. Thereby the LightSIDE-AthDGC fork extends LightSIDE (the open-source text-mining workbench developed at Carnegie Mellon by Carolyn Rosé’s group) to operate on syntactic features, namely dependency arcs, argument-structure frames, and morphology bundles, opening a new feature space for the classification of diachronic stages. A worked query example is train an SVM classifier (Support Vector Machine, a standard classification algorithm) on `grc_proiel` argument-structure frames to distinguish Archaic from Classical from Koine Greek, and inspect the top-weighted frames per class, which is the kind of analysis that the LightSIDE-AthDGC fork makes a one-hour classroom exercise.

Fourthly, digital editions and pedagogy. The showcase generator (51_build_showcase_site.py) produces browsable per-sample PROIEL trees from any JSONL (JSON Lines, the line-per-record plain-text format) corpus; the Quarto template pack (Quarto is an open-source publishing system that builds websites, slides, papers, and posters from a single source file) generates the whole multi-output site, namely HTML, Reveal.js slides, .docx, .pptx, Beamer, and an A0-format academic poster, from a single source. Both are open-source and reusable by other historical-treebank projects. The platform is taught as a KEDIVIM continuing-education course at NKUA, Ψηφιακά Εργαλεία για τη Διαχρονική Ανάλυση της Γλώσσας / Digital Tools for the Diachronic Analysis of Language, autumn 2026; see <https://athdgc.github.io/training.html>.

6. Access and format

The public site at <https://athdgc.github.io> provides curated samples only for the v0.4 release. The full annotated corpus partitions remain under audit on the GRNET ARIS national HPC under allocation pa260305; they release at v0.5 as a separate Zenodo dataset record under CC-BY-4.0. The source-code snapshot at the v0.4.0 Zenodo record is Apache-2.0.

The toolkit's readiness per module is summarised in §3.4 and on the public Tools page at <https://athdgc.github.io/tools.html>. Five modules are LIVE today (the Quarto template pack, the showcase generator, the corpus-fix toolkit, the PROIEL XML 2.0 exporter inside the build workflow, and the Stanza annotation job script). Eight are IN SETUP and be included with the v0.5 release. The LightSIDE-AthDGC syntactic-feature fork, the NoSketch-style concordancer, the Neo4j alignment viewer, the valency-frame database client, the retranslation-pair browser, and the retelling-chain explorer are FORTHCOMING at v0.5.

7. Acknowledgements

Funded by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI Research Projects to support Post-Doctoral Researchers, Project No. 20577; with complementary support from the Greece 2.0 National Recovery and Resilience Plan. Compute supplied by GRNET ARIS, the Greek national high-performance computing cluster, under allocation pa260305.

The funded project is CVL-CDSAML: A Corpus-based Valency Lexicon for a Contrastive and Diachronic Study of Ancient and Medieval Languages.

The team comprises Prof. Nikolaos Lavidas (Principal Investigator, NKUA), Prof. Emerita Kiki Niki-foridou (NKUA; Co-Editor of the Genres and Influential Texts volume), Prof. Dag Haug (University of Oslo; PROIEL Project Director), Prof. Leonid Kulikov (Ghent University; diachronic typology, valency questionnaires), Dr. Vassiliki Geka (NKUA; Post-Doctoral Researcher; Co-Editor of the Genres and Influential Texts volume), Dr. Vassileios Symeonidis (NKUA; Post-Doctoral Researcher), Dr. Theodoros Michalareas (NKUA; Post-Doctoral Researcher), Sofia Chionidi (NKUA; PhD Candidate), Anastasia Tsiropina (NKUA; PhD Candidate), Eleni Plakoutsi (NKUA; PhD Candidate), and Evangelos Argyropoulos (NKUA; Research Assistant).

References

- Dou, Zi-Yi, and Graham Neubig. 2021. “Word Alignment by Fine-Tuning Embeddings on Parallel Corpora.” *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL 2021)*, 2112–28.
- Eckhoff, Hanne M., Kristin Bech, Gerlof Bouma, et al. 2018. “The PROIEL Treebank Family: A Standard for Early Attestations of Indo-European Languages.” *Language Resources and Evaluation* 52 (1): 29–65.
- Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. “Language-Agnostic BERT Sentence Embedding.” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, Volume 1: Long Papers, 878–91.
- Haug, Dag T. T., and Marius Jøhndal. 2008. “Creating a Parallel Treebank of the Old Indo-European Bible Translations.” *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34.
- List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf University Press.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2020)*, 101–8.