RETRANSLATIONS ACROSS MILLENNIA

A Diachronic Contrastive Corpus for Studying Interlingual and Intralingual Language Contact

The evolution of valency over time: Digital pathways uncovering grammatical structure

Nikolaos Lavidas, Theodoros Michalareas, Vassilios Symeonidis, Sofia Chionidi, Anastasia Tsiropina, Eleni Plakoutsi

Athens Digital Glossa Chronos (AthDGC)

This presentation introduces the AthDGC project, focusing on how translation reveals structural language change over 3,000 years.

1.1 Introduction: Key Question

What is the main focus of this presentation?

- What is the AthDGC project?
- How does it combine philology and computation?
- What makes this approach innovative?

We aim to define the project's goals, its interdisciplinary approach, and its unique contribution to historical linguistics.

1.2 Textual Foundation & Computational Architecture

A comprehensive approach to diachronic valency research combining traditional philological expertise with modern computational linguistics.

Key Components:

- Parallel diachronic corpus of influential texts
- Diachronic Valency Lexicon for historical languages
- Dependency annotation using PROIEL framework

The architecture supports both qualitative philological analysis and quantitative computational extraction of linguistic patterns.

The project combines text selection, computational annotation, and the creation of a structured lexicon to analyze linguistic change.

1.3 The Need for a Comprehensive Foundation

Why do we need a comprehensive foundation for diachronic research?

- What challenges do historical linguists face?
- How does text quality affect research outcomes?
- What factors must be controlled for valid analysis?

Establishing a controlled, consistent methodology is crucial for drawing valid and measurable conclusions from historical data.

1.4 Introduction: Key Foundational Elements

Text selection and preparation

Choosing representative and chronologically linked text samples.

Annotation standards

Ensuring consistent morphological and syntactic tagging across all varieties.

Cross-linguistic comparability

Enabling systematic, apples-to-apples comparison between languages.

Temporal coverage

Spanning millennia of language evolution for robust diachronic analysis.

Valid diachronic analysis requires meticulous planning and rigorous quality control across all stages of data preparation.

Athens

The institutional base at the National and Kapodistrian University of Athens.

Digital

The methodology, utilizing computational and digital humanities methods.

Glossa (γλώσσα)

Language, representing the core object of linguistic study.

Chronos (χρόνος)

Time, emphasizing the crucial diachronic perspective of the project.

1.6 The AthDGC Research Team

Nikolaos Lavidas

Vasiliki Nikiforidou

Theodoros Michalareas

Vasiliki Geka

Vassilios Symeonidis

Sofia Chionidi

Eleni Plakoutsi

Anastasia Tsiropina

The project is powered by a diverse team of researchers.

2.1 Drivers of Language Change

Internal vs. External Factors

Understanding the drivers of language change is essential for identifying the specific role of translation.

Linguistic evolution can be driven by forces originating within the language system or through contact with others.

2.2 Distinguishing Internal and External Factors

Internal Factors:

- Phonological simplification
- Analogical leveling (regularization)
- Grammaticalization (lexical to functional)
- Reanalysis (reinterpreting existing structures)

External Factors:

- Language contact
- Bilingualism
- **Translation effects (structural calques)**
- Prestige borrowing (lexical)

Our project examines how translation serves as a controlled venue for studying external contact effects on syntax.

Internal factors affect all texts, while external factors, particularly translation, are concentrated in specific, contact-affected texts.

2.3 AthDGC Research Questions

Primary Questions:

- How do valency patterns change across millennia?
- What role does translation play in introducing new constructions?
- Can we distinguish contact-induced change from natural evolution?
- What cross-linguistic patterns emerge in valency change?

Methodological Questions:

- How can we automate annotation for historical languages?
- What conversion standards work across annotation schemes?
- How do we maximize the comparability of disparate texts?

Our research addresses both the fundamental linguistic mechanisms of change and the computational methods required to study them.

3.1 Theoretical Framework: LCTT

Language Contact Through Translation (LCTT)

This framework defines our approach to studying external influences on linguistic structure.

The LCTT model allows us to treat the translator's desk as a controllable environment for contact-induced change.

3.2 LCTT: Translations as Laboratories

How can translations serve as laboratories for studying language change?

- What is Language Contact Through Translation (LCTT)?
- How does translation differ from spoken contact?
- What can translations reveal about language structure?

Translations offer a unique window into linguistic contact because the source text remains fixed while the target language evolves.

3.3 The LCTT Framework Explained

Key Framework:

- Translation as a venue for language contact (Kranich, Becher & Höder 2011)
- Translations may introduce source language features into target languages (McLaughlin 2011)
- Creates controlled environment for studying contact-induced change
- Distinguishes between direct (bilingual) and indirect (textual) contact effects

Unlike spoken contact, translation creates a textual interface where source language structures may systematically influence target language productions.

LCTT posits that the structural pressure from the source text can lead to systematic 'shining through' of foreign patterns.

4.1 Corpus Selection: Influential Texts

Definition:

Texts that have had great cultural and linguistic impact in the history of humanity across numerous periods and languages.

Key Criterion:

They must be retranslated and republished across many periods, enabling robust diachronic comparison.

Examples:

- The Bible (translated into nearly every language)
- Homeric Poems (foundational to Western literature)
- Byzantine Chronicles (connecting ancient and modern periods)

We selected texts whose continuous retranslation over millennia guarantees broad temporal and linguistic coverage.

4.2 The Value of Retranslation

What is a 'retranslation'?

- How do retranslations differ from original translations?
- Why are retranslations valuable for linguistic research?
- What motivates communities to retranslate texts?

Retranslations capture the language's evolution because they are adapted to conform to the contemporary linguistic norms of the new target period.

4.3 Interlingual vs. Intralingual Retranslation

Intralingual Retranslation:

Translations within the same language across different time periods.

Example: Modern Greek translation of ancient Greek Homer.

Value: Tracks evolution of the target language itself, serving as a control.

Interlingual Retranslation:

Translations between different languages (e.g., Greek NT to English).

Example: English translations of the Greek New Testament.

Value: Tracks contact-induced change across language families.

Both types allow us to track how the target language evolves while the source text remains constant.

The comparison between these two types of retranslation allows us to isolate contact effects from internal drift.

4.4 Translation Categories in AthDGC

Translation Categories:

- ### Ancient --> Modern (same language)

  Example: Ancient Greek Homer → Modern Greek translation (Intralingual).

- ### Source --> Target (different language)

  Example: Greek NT → Old English, Middle English, Modern English (Interlingual).

- ### Intermediate translations

  Example: Greek → Latin → Old French, tracking translation chains.

The corpus includes direct translations and intermediate links to map complex, historical translation pathways.

5.1 Focus: Greek Language History

From Ancient to Modern Greek

Tracking the structural evolution across 3,000 years, characterized by morphological simplification and increased analyticity.

Greek history provides a clear-cut case of an Indo-European language shifting from synthetic to analytic structure.

5.2 Greek Language Periods in the Corpus

- **Ancient Greek (8th c. BCE – 4th c. BCE)**

  Homer, Classical Attic prose (Highly synthetic).

- **Hellenistic/Koine Greek (4th c. BCE – 4th c. CE)**

  New Testament, Septuagint (Transitional phase).

- **Medieval/Byzantine Greek (4th c. CE – 15th c. CE)**

  Psellos, Chronicle literature (Loss of Dative).

- **Modern Greek (19th c. – present)**

  Standard Modern Greek (Fully analytic structure).

Each period is represented by influential texts that capture key stages of morphological and syntactic change.

5.3 Intralingual Case Study: New Testament

Original Source:

Hellenistic Koine Greek (1st c. CE).

Medieval Edition:

Byzantine scholarly editions (conservative grammar).

Modern Translation:

Alexandros Palles translation (20th c. vernacular style).

Research Value:

Tracking how Greek syntax and valency patterns evolved while translating the exact same content across 2,000 years.

Comparing different Greek versions of the NT allows us to measure internal syntactic evolution over time.

5.4 Why Intralingual Translation Matters

Controls for semantic content

Same meaning, different forms, allowing focus on structure.

Reveals grammatical change

Shows which constructions become obsolete or innovative in the target variety.

Acts as a "fossil" control

The ancient source text remains fixed, showing how translators adapt to contemporary norms.

Intralingual comparison is critical for distinguishing endogenous (internal) change from exogenous (external/contact) change.

6.1 Focus: Passive and Middle Voice

Tracking voice alternations across Greek history

Voice is a crucial morphological category that directly affects a verb's valency and argument realization.

The development of the verb's voice system is tightly linked to the overall morphological simplification of Greek.

6.2 Historical Development of Greek Voice

Historical Development:

- Ancient Greek: Distinct passive and middle morphology.
- Koine Greek: Beginning merger of middle-passive forms.
- Medieval Greek: Simplified voice system.
- Modern Greek: Syncretic medio-passive (one set of endings).

Valency Implications:

Voice alternations (e.g., active → passive) directly shift argument structure by promoting an internal argument (Patient) to the subject position.

Tracking the middle voice provides insight into changes in argument realization and argument suppression.

The collapse of the separate middle and passive voices into a single medio-passive voice simplifies verb paradigms.

6.3 Key Structural Change: Case System

Greek Case Evolution:

- Ancient Greek: 5 cases (NOM, GEN, DAT, ACC, VOC)
- Koine Greek: Dative begins declining in frequency.
- Medieval Greek: Dative lost (replaced by genitive/prepositions).
- Modern Greek: 4 cases (NOM, GEN, ACC, VOC).

Impact on Valency:

The loss of the Dative case necessitates structural changes in verb argument frames, primarily the required use of prepositions for oblique roles.

This morphological loss drives a fundamental syntactic shift toward analytic structures.

The demise of the Dative case is the most significant morphological change tracked in the Greek diachrony.

7.1 Corpus Overview: Languages and Structure

The AthDGC Corpus

The backbone of our research is the annotated corpus spanning 3,000 years and multiple Indo-European branches.

The corpus is designed to facilitate both fine-grained Greek diachrony and broad cross-linguistic comparison.

7.2 Corpus Criteria: Text Selection

What criteria should guide selection of texts for a diachronic corpus?

- What makes a text 'representative' of its period?
- How do we balance text types and genres?
- What role does manuscript tradition play?

Text selection requires careful balancing of chronological placement, stylistic register, and textual preservation quality.

7.3 Corpus Overview: Languages and Timespan

Languages Covered (Phase I):

Hellenic: Ancient, Koine, Medieval, Modern Greek

Germanic: Gothic, Old English, Middle English, Modern English

Romance: Latin, Old French

Timespan:

~3,000

Years of Linguistic History

From 8th century BCE (Homer) to 21st century CE (Modern Texts).

Phase I focuses on the Greek source and its most direct translation branches (Germanic and early Romance).

7.4 Greek Texts in the Corpus (Hellenic Branch)

- **Ancient Greek:**

  Homeric poems (Iliad, Odyssey)

- **Hellenistic Greek:**

  New Testament, Septuagint

- **Medieval Greek:**

  Psellos' Chronographia, Sphrantzes' Chronicle

- **Early & Modern Greek:**

  Digenis Akritas, Contemporary translations

This selection ensures representation across canonical, religious, historiographical, and vernacular genres.

7.5 English Texts in the Corpus (Germanic Branch)

- **Old English:**

  Ælfric's Lives of Saints, West Saxon Gospels (early contact effects).

- **Middle English:**

  Orrmin's Ormulum, Wycliffe Bible (period of massive change).

- **Early Modern English:**

  Tyndale NT, King James Bible (standardization effects).

- **Modern English:**

  Jane Austen's Pride and Prejudice (non-translated baseline).

7.6 Latin and Old French Texts (Romance Links)

Latin:

Vulgate Bible (Jerome's translation), serving as a crucial intermediate source text.

Classical Latin comparanda (providing structural contrast).

Old French:

Anglo-Norman texts and Medieval French translations.

Value: These serve as intermediate links in translation chains and contact scenarios impacting English.

Romance texts are included to analyze the structural transmission of Greek and Latin patterns into Western European languages.

7.7 Gothic Texts: The Earliest Germanic Link

Wulfila's Gothic Bible (4th c. CE):

The earliest substantial Germanic text.

A direct translation from Greek (Septuagint and New Testament).

Value: Provides the earliest Germanic valency patterns for comparison, revealing initial Greek structural influence.

Note: Already partially included in the PROIEL treebank (Haug & Jøhndal 2008).

Gothic offers a pristine, early example of interlingual contact between Greek and a Germanic language.

8.1 Methodology: The PROIEL Framework

Annotation Methodology and Tools

To ensure maximum consistency and comparability, all AthDGC data is annotated using the PROIEL scheme.

PROIEL (Pragmatic Resources in Old Indo-European Languages) is designed specifically for historical linguistics.

8.2 Dependency Annotation Benefits

What are the benefits of dependency annotation for historical languages?

- Why is dependency grammar preferred over constituency?
- How does PROIEL handle morphologically rich languages?
- What information is captured in the annotation?

Dependency Grammar is uniquely suited for the flexible syntax found in ancient and medieval Indo-European languages.

8.3 The PROIEL Treebank (Oslo Project)

PROIEL Overview:

University of Oslo project (Haug & Jøhndal 2008)

Open-source tools and data for collaborative annotation.

Uses a Dependency Grammar framework.

Target Languages:

Designed for free-word-order languages (FWO).

Already includes Greek, Latin, Gothic, Old Church Slavonic, Armenian.

PROIEL provides the standardized syntax and methodology necessary for cross-corpus comparability.

1. Tokenization

Word and sentence boundaries.

2. Lemmatization

Dictionary headwords.

3. Morphology

Case, number, gender, tense, mood, voice.

4. Part of Speech

Categorizing word classes.

5. Syntax

Dependency relations (subject, object, etc.).

8.5 Dependency Grammar for Valency Research

Why Dependency Grammar?

- Handles free word order naturally.
- No need for empty categories or movement (less complex).
- Direct representation of argument structure.

For Valency Research:

Dependencies directly encode verb-argument relations (Head → Dependent), which is exactly what valency theory describes.

This direct relationship simplifies the extraction of valency frames and their arguments.

Dependency grammar's focus on binary word relationships perfectly aligns with the argument structures central to valency theory.

8.6 AthDGC Annotation Workflow

Three-Stage Annotation Process:

1

Tokenization & Division

Fully automated.

2

Morphological Tagging

Partly automated, manual review for historical forms.

3

Syntactic Annotation

Manual annotation with automatic suggestions.

The PROIEL platform provides a web interface for distributed annotation and quality control across the team.

Annotation requires a blend of automated processing and crucial manual validation by expert linguists.

9.1 Focus: Valency Theory Fundamentals

Argument Structure Fundamentals

Valency theory centers on the verb's relationship with the participants it governs in the clause.

The French linguist Lucien Tesnière first established the verb as the structural center of the sentence (Tesnière 1959).

9.2 Valency: Definition and Significance

What is valency and why does it matter?

- How do we define verb valency?
- What determines how many arguments a verb takes?
- How does valency relate to meaning?

Valency defines the syntactic roles necessary for a verb to express its full semantic meaning.

9.3 Valency Classes (Tesnière 1959)

Avalent (0 arguments)

It rains.

Monovalent (1 argument)

Mary sleeps.

Divalent (2 arguments)

Mary likes John.

Trivalent (3 arguments)

Mary gave a letter to John.

The verb is the central element; all other required elements depend on it.

Valency is classified by the minimum number of participants required to complete the verb's meaning.

9.4 Arguments vs. Adjuncts

Arguments:

Required by the verb's meaning (Subject, Direct Object, Indirect Object).

Example: "John ate the apple." (Eater and eaten are arguments).

Tests: Semantic selection and subcategorization.

Adjuncts:

Optional additions that modify the verb or clause (e.g., adverbs, location phrases, time phrases).

Example: "John ate the apple quickly." (Manner is an adjunct).

Test: They can typically be omitted without making the sentence ungrammatical.

A key task in annotation is consistently distinguishing between the core, required arguments and the optional adjuncts.

10.1 Valency Alternations: Cross-Linguistic Variation

How do valency alternations differ cross-linguistically?

- What types of alternations exist?
- Are alternations universal or language-specific?
- How do alternations change over time?

Valency alternations show the flexibility of verb meaning and how arguments can be realized differently.

10.2 Common Valency Alternations I

Middle Alternation

"Jane broke the vase" (Transitive) → "The vase broke" (Intransitive).

Causative/Inchoative

"They stood the statue" (Causative) → "The statue stood" (Inchoative).

Dative Alternation (English)

"They gave him cake" (Double Object) → "They gave cake to him" (PP Dative).

These alternations, such as the Dative shift, reveal underlying semantic classes of verbs (Levin 1993).

10.3 Common Valency Alternations II

Unspecified Object Alternation

"Mike ate all the cake" → "Mike ate" ($\checkmark$)

Verbs of consumption allow their object to be omitted.

Locative Alternation

"She loaded hay onto the wagon" → "She loaded the wagon with hay."

Swapping the figure and ground arguments.

Reflexive Alternation

"John dressed the child" → "John dressed (himself)."

Analyzing these subtle shifts is key to understanding how grammatical patterns change across historical periods.

11.1 The Diachronic Valency Lexicon

Tracking Diachronic Verb Patterns

The lexicon is the final, searchable product of the corpus annotation, quantifying valency evolution.

The Valency Lexicon moves beyond raw text to present structured, quantifiable data on verb usage changes.

11.2 Lexicon Entry Structure

For Each Verb Entry:

- Lemma (dictionary form)
- Argument frames (e.g., NOM.V.ACC, NOM.V.DAT)
- Frequency attestation counts
- Examples (cited passages with annotation)

Metadata:

- Period (temporal classification)
- Source text reference (unique ID)
- Language (Greek, English, Latin, etc.)

Every entry links a verb's historical usage to precise quantitative data and philological context.

11.3 Lexicon Example: Modern English 'write'

Period: Modern English - Frequency: 1015

*Examples:

*

"You write uncommonly fast."

"...a person who can write a long letter with ease..."

"...and when I next write to her..."

Frames Attested:

NOM.V / NOM.V.ACC / NOM.V.to-DAT / Passive

Modern English shows a consistent reliance on the Prepositional Phrase (`to-DAT`) for the Recipient argument.

11.4 Lexicon Example: Middle English 'writen'

Period: Middle English - Frequency: 103

*Examples:

*

"Forr Moysæs wrat off himm..." (For Moses wrote of him...)

"...patt ure Laferrd Crist wrat o be grund..." (...that our Lord Christ wrote on the ground...)

Changes Observed:

Orthographic variation (wrat , writen) and emerging prepositional use alongside older case remnants.

Middle English marks a transitional period where spelling is unstable and the case system is getting lost.

11.5 Lexicon Example: Old English 'wrītan'

Period: Old English - Frequency: 10

*Example:

*

"...for eower heortan heardnesse he eow wrat þis bebod"

(...for your heart's hardness he wrote you.DAT this command)

Key Observation:

The Dative recipient (eow = you.DAT) is a bare noun form, requiring no preposition, typical of a synthetic language.

Old English provides the synthetic baseline for Germanic, showing a functional case system prior to its collapse.

11.6 Diachronic Pattern: The Evolution of 'write'

- **Old English:**

  wrītan + DAT recipient (fully synthetic system).

- **Middle English:**

  writen + to/unto + recipient (preposition emerging, competing with case).

- **Modern English:**

  write + to + recipient (preposition required, analytic structure).

The diachronic pattern for 'write' mirrors the change seen in Greek: case loss drives the emergence of prepositions to mark arguments.

12.1 Challenges Encountered

Technical and Linguistic Obstacles

Historical texts present unique difficulties that standard NLP pipelines cannot fully resolve.

The non-standard nature of historical data requires specialized tools and significant manual intervention.

12.2 Digitizing Historical Texts: Key Problems

What challenges arise in digitizing historical texts?

- How do we handle orthographic variation?
- What problems emerge with manuscript traditions?
- How accurate is automatic parsing?

Challenges stem from the lack of standardization in spelling, grammar, and text transmission across centuries.

12.3 Challenge: Orthographic Variation

Old and Middle English Examples:

Non-standard letters: $\mathbf{u}$ instead of $\mathbf{v}$; $\mathbf{i}$ instead of $\mathbf{j}$.

Same word, different spellings: gyue, yyue, yiue, giue all mean "give".

Historical letters: Thorn ($\mathbf{\mathbf{p}}$) and eth ($\mathbf{\mathbf{\delta}}$) vs. th .

Our Approach:

Keep original orthography BUT create custom dictionaries for each text to enable computational processing while preserving philological accuracy.

Custom dictionaries and normalization routines are essential to link variant spellings to a single lemma for analysis.

12.4 Challenge: Null Subjects (Pro-Drop)

Pro-drop Languages:

Greek (ancient through modern) is a pro-drop language; subjects are often unexpressed but understood.

Example:

$\gamma\rho\acute{\alpha}\phi\epsilon\iota$ (*graphei*) = "writes" or "he/she writes."

Extraction Challenge:

Automatic parsers struggle with null subjects; the inferred subject must be manually tagged.

If missed, the valency frame is under-represented (e.g., recorded as Monovalent instead of Divalent).

Accurate valency counting requires expert human annotation to correctly infer and tag missing grammatical arguments.

12.5 Challenge: Modern Greek Tagging

New Annotation Tags Needed:

Modern Greek requires tags for newly grammaticalized particles that lack equivalents in Ancient Greek:

$\nu\alpha$ (*na*) - subjunctive marker

$\alpha\varsigma$ (*as*) - hortative marker

$\theta\alpha$ (*tha*) - future marker (from $\dot{\eta}\theta\epsilon\lambda\alpha$)

Solution: We use an extended PROIEL tagset for Modern Greek particles, maintaining compatibility with the core Ancient Greek annotation scheme.

The annotation scheme must be flexible enough to handle both archaic morphology and modern functional particles.

12.6 Automatic Parsing Evaluation Results

Late Medieval Greek Test:

Testing 4 parsers (trained on Modern or Ancient Greek) against the Late Medieval Sphrantzes Chronicle.

Results: Very low F1 scores (< 0.42) .

Conclusion:

Semi-automatic annotation is currently as time-consuming as manual annotation due to high error rates.

We need trained models specific to each historical variety to maximize automation benefits.

The significant difference between historical varieties means existing models cannot be reliably transferred without custom training.

13.1 Cross-linguistic Perspectives

English Parallels to Greek Changes

The changes observed in Greek are mirrored by similar evolutionary patterns in the Germanic branch.

Comparing Greek and English allows us to test if the historical trend toward analyticity is a universal phenomenon.

13.2 English Valency Change (Trips & Stein 2019)

Case Loss:

Old English's robust case system collapsed into the system seen in Middle and Modern English.

Preposition Use:

Increased reliance on prepositions to realize arguments that were formerly case-marked (e.g., Dative).

Word Order:

Development towards a more rigid SVO (Subject-Verb-Object) fixed word order.

Key Differences:

English experienced almost complete case loss, while Greek retained the Nominative/Accusative distinction.

13.3 Contact-Induced Change in English

French Influence (Post-1066):

Massive lexical borrowing and some syntactic influence (e.g., word order, relative clauses) from Norman French.

New valency patterns emerged with borrowed verbs.

Parallel to Greek:

Just as Greek translations from Hebrew/Aramaic may show contact effects, English translations from Latin/French may import foreign constructions, accelerating internal trends.

External contact events, such as the Norman conquest, can rapidly accelerate internal tendencies like case loss in English.

13.4 Valency and Language Contact

Can valency patterns be "borrowed" through translation?

Shining through:

Source language patterns visible in translation.

Loan translations:

Calques of argument structures (e.g., Greek ditransitive structure copied into Germanic).

Frequency shifts:

Rare native patterns become more common under external pressure.

The key is that valency, as the core of syntactic structure, is susceptible to influence from foreign language patterns.

14.1 Theoretical Implications

For Historical Linguistics:

- Empirical basis for claims about syntactic change.
- Quantitative tracking of argument structure evolution.
- Testing theories of grammaticalization.

For Contact Linguistics:

- Controlled environment for studying contact effects.
- Distinguishing borrowing from parallel development.
- Understanding "shining through" phenomena.

The project provides quantitative data to support and refine major theories in both diachronic and contact linguistics.

14.2 Understanding Mechanisms of Change

What We Learn:

- How grammatical categories evolve over millennia.
- Which changes are gradual vs. abrupt.
- How contact accelerates or redirects change.
- The role of prestige and standardization.

Methodological Advances:

- Combining philological depth with computational scale.
- Developing replicable annotation protocols.

The methodology yields a deep understanding of the pace and causality of long-term linguistic evolution.

14.3 Unique Contribution: Long-term Documentation

Unique Contribution:

- 3,000+ years of continuous documentation.
- Multiple parallel translations enabling controlled comparison.
- Rich morphological data for tracking case/valency changes.

No Other Resource Provides:

Such temporal depth with annotation consistency, parallel texts across such diverse periods, or both intralingual and interlingual comparisons.

AthDGC creates a single, unified database for structural change across three millennia of Indo-European history.

14.4 Revealing Contact Effects in Written Language

Translation as Contact:

Translations create measurable contact situations where source language features may "shine through" due to repeated textual exposure.

Distinguishing Contact from Internal Change:

We compare translated vs. non-translated texts and track features that persist vs. disappear, measuring frequency shifts over time.

The project's controlled comparisons isolate the specific influence of textual contact on the target language's structure.

14.5 Building Empirical Foundation for Diachronic Theory

From Qualitative to Quantitative:

- Moves beyond anecdotal examples to statistically test hypotheses.
- Ensures reproducible methodology that can be validated by the community.

Testing Existing Theories:

We provide data to test theories on grammaticalization pathways, the unidirectionality hypothesis, and contact-induced grammaticalization.

The data provides a quantitative backbone to test and refine existing theories of language evolution.

15.1 Practical Applications

Beyond Academia

The rich, structured data has immediate value for technology, education, and cultural heritage.

The project's output serves as crucial infrastructure for digital humanities and language technology.

15.2 Real-World Uses of AthDGC Data

Translation Technology

Machine translation of historical texts (HMT).

Language Learning

Tools for ancient language education.

Cultural Heritage

Digital manuscript archives and accessibility.

15.3 Translation Technology Applications

Historical Machine Translation:

- Training data for ancient $\rightarrow$ modern translation models.
- Cross-period linguistic models for diachronic transfer.
- Automatic glossing and lemmatization for historical texts.

Translation Memory:

- Parallel aligned historical texts for pattern matching across centuries.
- Consistency checking for new translations of classical works.

The annotated, parallel data offers a unique, structured training set for historical NLP models.

15.4 Language Learning Resources

For Ancient/Historical Languages:

- Annotated reader texts with full syntactic parsing.
- Vocabulary frequency lists stratified by period.
- Interactive grammar pattern databases.

For Modern Language Education:

- Historical etymology resources showing word origins.
- Comparative syntax materials illustrating structural change.

The lexicon provides the raw data necessary for creating next-generation, historically-grounded educational tools.

15.5 Cultural Heritage Preservation

Digital Preservation:

TEI-XML encoding of manuscript texts and Linked Open Data standards for historical resources.

Accessibility:

Making ancient texts fully searchable and creating cross-linguistic discovery tools for public access to cultural patrimony.

The digital formats ensure the longevity of the texts and enhance public accessibility and research utility.

16.1 Future Directions

What Comes Next?

Our work continues with corpus expansion, automation improvement, and forging international partnerships.

Phase II focuses on scaling the project, both linguistically and computationally, to maximize impact.

16.2 Ongoing Work: Expansion and Automation

Corpus Expansion:

- Expand the least represented languages (Old and Middle French, Old English).
- Add new languages, such as Gothic expansion.

Automate Syntactic Annotation:

- More accurate conversion pipeline from UD-CoNLL to PROIEL.
- Training existing syntactic parsers on historical varieties.

Collaboration with Harvard CHS (ARCAS Team) is key to these efforts.

The current phase involves filling data gaps and developing better machine learning tools for annotation.

16.3 Phase II: Expanding Language Coverage

Phase II Languages:

Slavic: Old Church Slavonic (already in PROIEL)

Celtic: Old Irish, Middle Welsh

Indo-Iranian: Classical Sanskrit, Old Persian

Goals of Expansion:

Broader Indo-European coverage to test universality of change.

More translation chains to analyze contact transmission.

Typologically diverse comparison for more robust conclusions.

Expanding beyond the core Hellenic and Germanic languages will validate our theories on a global Indo-European scale.

16.4 Improving Automation: Addressing Parser Issues

Current Challenges:

- Low parser accuracy for historical varieties (< 0.42 F1 on some texts).
- Time-consuming manual correction of high error rates.
- Inconsistent tagset mappings across different corpora.

Solutions in Progress:

- Training period-specific models (e.g., dedicated Medieval Greek model).
- Transfer learning from related, higher-resource varieties.
- Active learning to reduce total manual annotation effort.

Investment in advanced NLP is essential to reduce the burden of manual annotation and accelerate corpus growth.

17.1 Harvard CHS Collaboration

ARCAS Project Partnership

Collaboration ensures that AthDGC data and methods are aligned with major global initiatives in classical studies.

17.2 Collaboration Areas with ARCAS

ARCAS Project:

Advanced Research Collaborative on Ancient Sciences.

Focuses on developing digital infrastructure for classical studies.

Collaboration Areas:

- Data sharing and format compatibility (PROIEL/UD).
- Joint annotation guidelines and standards.
- Combined training workshops for historical corpus linguists.

Partnering with Harvard enables the development of joint standards and the pooling of intellectual resources.

19.1 Additional Corpus Examples

More Diachronic Patterns

Further illustration of valency change using Greek, English, and Latin verb paradigms.

These examples reinforce the consistency of the synthetic-to-analytic shift across different language branches.

19.2 Greek verb $\delta\iota\delta\omega\mu\iota$ ('To Give')

Ancient Greek:

**$\delta\acute{\iota}\delta\omega\mu\acute{\iota}\
\sigma\omicron\iota\
\tau\omicron\tilde{v}\tau\omicron$**

(I give you.DAT this)

Modern Greek:

**$\sigma\omicron\upsilon\
\delta\acute{\iota}\nu\omega\
\alpha\upsilon\tau\acute{\omicron}$**

(you.GEN give.1SG this)

Observed Change:

Dative shifts to Genitive for recipient; verb form simplified (athematic $\rightarrow$ thematic); clitic pronoun position changed.

This case exemplifies the morphological loss (Dative) driving the substitution of one case (Genitive) for argument realization.

Say')

Classical Greek:

$\lambda\acute{\epsilon}\gamma\omega\ \sigma\omicron\iota\ \text{\quad}\ddot{o}\tau\iota\dots$ (I say to-you.DAT that...)

Koine Greek:

$\lambda\acute{\epsilon}\gamma\omega\ \pi\rho\acute{o}\varsigma\ \sigma\grave{\epsilon}\ \ddot{o}\tau\iota\dots$ (I say towards you.ACC that...)

Modern Greek:

$\sigma\omicron\upsilon\ \lambda\acute{\epsilon}\omega\ \ddot{o}\tau\iota\dots$ (you.GEN say.1SG that...)

Observed Pattern:

Progressive loss of dative marking, replaced by prepositional phrases in Koine before settling on the Genitive in Modern Greek.

19.4 English verb 'give' Diachrony

Old English:

Ic sealde him $\mathbf{p} \ddot{\mathbf{x}} t\ \mathbf{b} \mathbf{o}\mathbf{c}$ (I gave him.DAT the book)

Middle English:

I yaf him the book / I yaf the book to him

Modern English:

I gave him the book / I gave the book to him

Observed Change:

The Dative Alternation emerges in Middle English, where the case marking is replaced by the prepositional *to*-phrase.

The English development shows the emergence of two competing analytic valency frames for the Recipient argument.

19.5 Latin-Romance Comparison

Latin:

Do tibi librum (give.1SG you.DAT book.ACC)

Old French:

Je te done le livre (I you.DAT give the book)

Observed Pattern:

Synthetic case morphology (DAT) is replaced by analytic **clitic pronouns** (*te* in French), illustrating a parallel shift to Greek and English.

Romance languages show a similar shift from synthetic case to analytic expression, often through the use of clitics.

19.6 Cross-linguistic Valency Patterns

Common Diachronic Tendencies:

- Case $\rightarrow$ preposition/postposition (Analyticity).
- Synthetic $\rightarrow$ analytic marking.
- Free word order $\rightarrow$ fixed order.

Language-Specific Paths:

- Greek: Dative $\rightarrow$ Genitive or PP.
- English: Dative $\rightarrow$ *to*-PP or double object.
- Romance: Dative $\rightarrow$ *a* + pronoun (clitic).

While the overall trend toward analyticity is shared, each language finds unique structural solutions for the same problems.

20.1 Methodological Innovations

New Approaches Developed

The scope of the AthDGC project has necessitated the development of novel computational and philological integration methods.

Our work is defined by the necessary fusion of deep humanistic knowledge with large-scale computational power.

20.2 The Integration of Disciplines

Traditional Philology:

Deep textual expertise, manuscript knowledge, and historical context are essential for correct interpretation.

Computational Methods:

Large-scale pattern extraction, statistical validation, and reproducible analysis.

Integration:

Neither discipline alone is sufficient; our approach combines both strengths to ensure data accuracy and statistical power.

The synergistic relationship between the two disciplines is the project's greatest methodological asset.

20.3 Open Science Principles

Our Commitments:

- **Open Data:** Annotated texts freely available.
- **Open Tools:** Software under open-source licenses.
- **Open Methods:** Documented, replicable procedures.
- **Open Access:** Publications in OA venues.

Benefits:

- Community validation and improvement of the corpus.
- Broader research impact and educational reuse.

We adhere to Open Science principles to maximize transparency, reproducibility, and collaborative development.

20.4 Building Collaborative Networks

Network Effects:

Shared standards increase data value; distributed annotation scales faster; diverse expertise improves quality.

Our Role:

Serving as a hub for Greek historical linguistics, a bridge between classical and modern studies, and a training ground for digital philology.

Collaboration ensures our standards are globally compatible and our research is validated by the wider community.

23.1 Contact and Resources

Project Website:

🔗**https://athdgc.github.io/ (Coming soon)**

Principal Investigator:

Professor Nikolaos Lavidas
✉lavidas@enl.uoa.gr

Institution:

National and Kapodistrian University of Athens

Funding:

H.F.R.I. Excellence Grant - Project 15284

All project information, publications, and data will be made available through the official website and open access channels.

23.2 Selected References

Key Sources

This section lists the foundational texts and publications guiding the AthDGC research and methodology.

The methodology is built upon established frameworks in corpus linguistics, diachronic syntax, and valency theory.

23.3 Selected References (1/4)

- Eckhoff, H., et al. (2018). The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation*.
- Haug, D. T. T. & Jøhndal, M. L. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. *Proceedings of the 6th LREC*.
- Kranich, S., Becher, V., & Höder, S. (2011). A tentative typology of translation-induced language change. In S. Kranich et al. (Eds.), *Multilingual Discourse Production*. Amsterdam: John Benjamins.

These references establish the core corpus annotation methodology and the theoretical framework of translation-induced change.

23.4 Selected References (2/4)

- Lavidas, N. (2021). *The Diachrony of Written Language Contact: A Contrastive Approach*. Brill's Studies in Historical Linguistics.
- Lavidas, N., et al. (Eds.). (2023). *Internal and External Causes of Language Change: The Naxos Papers*. London: Bloomsbury.
- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- McLaughlin, M. (2011). *Syntactic Borrowing in Contemporary French: A Linguistic Analysis of News Translation*. Oxford: Legenda.

This section includes core publications on valency theory and the principal investigator's work on language contact and change.

23.5 Selected References (3/4)

- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.
- Trips, C. & Stein, A. (2019). Contact-induced changes in the argument structure of Middle English verbs. *Journal of Language Contact*.
- Trips, C. (2020). Copying of argument structure: A gap in borrowing scales. In B. Drinka (Ed.), *Historical Linguistics 2017*. Amsterdam: John Benjamins.
- Sitaridou, I. (2014). The Romeyka Infinitive: Continuity, Contact and Change in the Hellenic varieties of Pontus. *Diachronica*.

These references cover the foundation of valency theory and the crucial topic of argument structure borrowing.

23.6 Selected References (4/4)

- Sitaridou, I. (2016). Reframing the phylogeny of Asia Minor Greek: The view from Pontic Greek. *CHS Research Bulletin*.
- Nikiforidou, K. (2018). Genre and constructional analysis. *Pragmatics & Cognition*.
- Fried, M. & Nikiforidou, K. (Eds.). (2025). *The Cambridge Handbook of Construction Grammar*. Cambridge: Cambridge University Press.
- Mikros, G. (2005). Basic Quantitative Characteristics of the Modern Greek Language Using a Representative Corpus. *Journal of Quantitative Linguistics*.

This selection represents the broader theoretical engagement of the team with contemporary linguistic models and Greek varieties.

23.7 Additional References

- Malchukov, A. & Comrie, B. (Eds.). (2015). *Valency Classes in the World's Languages*. Berlin: De Gruyter Mouton.
- Zanchi, C. (2021). HoDeL: Homeric Dependency Lexicon. University of Bergamo.
- Luraghi, S., Cuzzolin, P., & Zanchi, C. (2024). PaVeDa: Pavia Verbs Database for Ancient Indo-European Languages.
- Collins, C. (2024). Argument Structure. In *Cambridge Handbook of Generative Syntax*.

We incorporate insights from related valency databases and argument structure theory in our lexicon design.

27.1 Acknowledgments

Funding:

Hellenic Foundation for Research & Innovation (H.F.R.I.) Excellence Grant, Project Number 15284.

Institutional Support:

National and Kapodistrian University of Athens, Greece 2.0 National Recovery and Resilience Plan.

Collaborators:

University of Oslo (PROIEL), Harvard CHS, and all other partner institutions contributing expertise and data.

We express gratitude to all partners and funding bodies that make this ambitious project possible.

Thank You!

Questions & Discussion

AthDGC Project - Athens Digital Glossa Chronos

nlavidas@enl.uoa.gr

We appreciate your time and interest in the Diachronic Contrastive Corpus research.

28.1 Questions?

Open Floor for Discussion

We are ready to address any specific inquiries regarding the project's methodology, data, or findings.