Retranslations Across Millennia: A Diachronic Contrastive Corpus for Studying Interlingual and Intralingual Language Contact

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

Athens Digital Glossa Chronos (AthDGC)

National and Kapodistrian University of Athens

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

H.F.R.I. Hellenic Foundation for Research & Innovation

Greece 2.0 NATIONAL RECOVERY AND RESILIENCE PLAN

DELPHI Workshop 2025 · November 26-29, 2025 · Delphi, Greece

KEY QUESTION

What does this presentation research focus on?

- What is the main architectural framework?
- How does it support linguistic research?
- What makes this approach unique?

Textual Foundation & Computational Architecture

A comprehensive approach to diachronic valency research

The AthDGC project represents a novel integration of traditional philological expertise with modern computational linguistics. This corpus contains carefully curated historical texts that enable systematic analysis of how verb valency patterns evolve over time. The architecture supports both qualitative philological analysis and quantitative computational extraction of linguistic patterns.

KEY QUESTION

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?

Introduction: The Need for a Comprehensive Foundation

Diachronic linguistic analysis requires careful consideration of multiple elements:

- **Text selection** and preparation
- Assessment of **historical evidence**
- Understanding **language development** contexts
- Recognition of **external factors** influencing linguistic change

Studying language change over millennia demands rigorous methodology in selecting and preparing texts. The AthDGC corpus addresses this need by establishing clear criteria for text inclusion and implementing systematic quality control measures. This foundation ensures that observed linguistic patterns reflect genuine historical developments rather than inconsistent data.

KEY QUESTION

What exactly is AthDGC?

- What does AthDGC stand for?
- How many texts does it contain?
- What time period does it cover?

What is AthDGC?

- **Purpose:** Foundation for a diachronic valency lexicon project
- **Scope:** historical texts covering **three millennia** of Greek language development
- **Inclusion:** Parallel traditions from **Germanic** and **Romance** languages
- **Approach:** Enables study of both internal and external language interactions

The AthDGC (**Athens Digital Glossa Chronos** — "Athens Digital Language Time") is the research team and corpus infrastructure at the National and Kapodistrian University of Athens dedicated to diachronic corpus linguistics. The project builds comprehensive textual databases for extracting valency information across languages and time periods. Each text is digitally encoded with linguistic annotations that enable automatic extraction of verb-argument structures. The corpus architecture supports both synchronic comparison across languages and diachronic tracking of changes within individual languages.

KEY QUESTION

What research questions can this corpus address?

- Internal vs. external language development
- Genre and register effects
- Historical transmission problems

Research Context & Methodology

The corpus allows researchers to analyze:

- **Internal language development** patterns
- **External language interactions** through contact
- **Genre effects** on linguistic features
- **Register variations** across time periods
- **Transmission problems** in historical linguistic evidence

The corpus is structured to allow comparison of the same text (e.g., Genesis 1) across five Greek translations spanning 2000 years, enabling precise tracking of how specific verbs like *ποιέω* "to make" change their argument structures over time.

KEY QUESTION

Who funds this research and what is the timeline?

- Which foundation supports the project?
- What is the funding period?
- What makes this project competitive?

Funding and Project Timeline

Hellenic Foundation for Research & Innovation (H.F.R.I.):

- **ELIDEK Excellence** grant program
- **2024-2025** funding period
- Competitive **national grant**
- Recognition of project **significance**

The H.F.R.I. ELIDEK Excellence grant provides substantial support for ambitious research projects in Greece. This competitive funding recognizes the innovative combination of traditional philological - linguistic methods with computational approaches. The grant enables hiring of researchers, acquisition of digital resources, and development of the computational infrastructure necessary for corpus processing and valency extraction.

KEY QUESTION

What translation theory underlies this research?

- What is LCTT theory?
- How does translation affect language change?
- What is "translationese"?

Language Contact Through Translation (LCTT)

A Key Theoretical Framework:

- Translation as a venue for **language contact** - Translations may introduce **source language features** into target languages- Creates a controlled environment for studying **contact-induced change**- Distinguishes between **direct** and **indirect** contact effects

LCTT theory, as developed by Kranich, Becher & Höder (2011) and McLaughlin (2011), provides a framework for understanding how translations serve as a unique form of language contact. Unlike spoken contact situations, translation creates a textual interface where source language structures may influence target language productions. This "translationese" effect can be measured and tracked over time, revealing patterns of influence that may eventually enter the (standard) target language.

KEY QUESTION

What types of influential texts are included?

- What genres are represented?
- Why are these texts "influential"?
- What languages do they come from?

Influential Texts in the Corpus

Biblical and Religious Texts:

- Hebrew Bible / Old Testament - Greek New Testament - Septuagint (Greek translation of Hebrew scriptures)

Historical and Literary Works:

- Homer's *Iliad* and *Odyssey* - Herodotus' *Histories* - Byzantine chronicles

Why "Influential"?

These texts shaped language use across centuries through liturgical reading, educational curricula, and cultural prestige. Their translation and retranslation created chains of linguistic influence traceable through the corpus.

KEY QUESTION

How does the corpus handle multiple translations?

- What is "retranslation"?
- Why translate the same text multiple times?
- What can we learn from comparing translations?

Multiple Translations and Retranslations

The corpus includes:

- **Original source texts** (Hebrew, Classical Greek) - **Multiple translations** of the same source into different target languages - **Retranslations** within the same language at different time periods

Example: Genesis 1

- Hebrew original - Septuagint (Greek, 150 BCE) - Byzantine Greek paraphrase - Modern Greek translation (20th century)

This structure enables tracking of how the same semantic content is expressed differently across time and languages.

KEY QUESTION

What is the difference between passive and active retranslation?

- What motivates retranslation?
- When does the original translation become obsolete?
- What role do ideological factors play?

Types of Retranslation

Passive Retranslation:

- Occurs when earlier translation is no longer accessible or intelligible - Language change makes original translation obsolete - Example: Modern Greek translations of the Septuagint

Active Retranslation:

- New translations created while earlier versions still accessible - Motivated by doctrinal, stylistic, or interpretive differences

Example: Multiple English Bible versions in the 20th century

Both types provide valuable data for understanding language change and translation practices.

KEY QUESTION

What specific retranslation examples does the corpus include?

- What Greek translations are included?
- What time span do they cover?
- How do they differ linguistically?

Greek Biblical Retranslations in the Corpus

Old Testament Translations:

- **Septuagint** (c. 150 BCE) - **Aquila's translation** (2nd century CE) - **Byzantine paraphrases** (Medieval period) - **Modern Greek versions** (19th-20th century)

New Testament:

- **Original Greek** (1st-2nd century CE) - **Byzantine textual tradition** - **Modern Greek adaptations**

This range enables tracking of Greek language development from Hellenistic through Byzantine to Modern periods.

KEY QUESTION

What is intralingual vs. interlingual translation?

- Translation within the same language vs. between languages
- When is intralingual translation necessary?
- What linguistic changes does it reveal?

Intralingual and Interlingual Translation

Interlingual Translation:

- Translation **between different languages** - Example: Hebrew → Greek (Septuagint) - Reveals cross-linguistic contact effects

Intralingual Translation:

- Translation **within the same language** across time periods - Example: Koine Greek → Modern Greek - Reveals internal language change patterns

The corpus systematically includes both types to distinguish contact-induced change from natural language evolution.

KEY QUESTION

Why is intralingual translation linguistically significant?

- What changes between language stages?
- How do translators update archaic features?
- What remains constant despite time?

Significance of Intralingual Translation

Intralingual translation reveals:

- **Grammatical changes** (case system simplification) - **Lexical changes** (obsolete vocabulary replacement) - **Syntactic changes** (word order preferences) - **Stylistic changes** (register shifts)

Research Value:

- Controlled comparison (same semantic content) - Clear temporal boundaries - Identifiable translator decisions - Direct evidence of perceived archaism

When a translator updates "archaic" Greek to "modern" Greek, their choices reveal what features have become obsolete or marked.

Passive Retranslation Explained

Occurs when: A previous translation becomes inaccessible or unintelligible to contemporary speakers

Example: Modern Greek Translations of the Old Testament

- The Septuagint (c. 150 BCE) became no longer intelligible to modern speakers
- This necessitated new translation work for contemporary audiences

Active Retranslation Explained

Occurs when: New translations offer competing perspectives based on different theological, doctrinal, or interpretive preferences

Example: English Bible Translations (19th-20th centuries)

- Protestant versions
- Catholic versions
- Each reflecting different theological and doctrinal preferences

Language Families in the Corpus

The AthDGC corpus includes texts from **three Indo-European language families:

The Three Language Families

1. **Hellenic**

- Greek (across all historical periods)

2. **Germanic**

- English (various historical stages)

3. **Romance**

- Latin
- French

Why These Languages?

Strategic Selection Based On:

- Significant **attested history** of written contact
- Long **documented traditions**
- Rich **translation practices**
- Extensive **historical records**

The Role of Original Texts

Original influential texts are included as:

- **Source texts** for translation analysis
- **Control group** to identify and filter:
- Source language influences
- Unique characteristics of translations
- "Translationese" phenomena

(Toury 1995)

The Core Hellenic Tradition

Greek: The Longest Historical Record

- Written documents from **Mycenaean era** to present day
- Unparalleled **continuity** of documentation
- Critical for understanding **long-term language change**

Focus on Post-Classical Development

The AthDGC corpus examines:

- **Koine Greek** (Hellenistic and Roman periods)
- **Byzantine Greek** (Medieval period)
- **Modern Greek** varieties

Why?

This period represents a crucial time for studying Greek grammar shift

Grammatical Shift: Synthetic to Analytic

Major Transition in Greek:

- From **synthetic structures** (morphologically complex)
- To **analytic structures** (syntactically complex)

Key Research Question:

How and when did this shift occur? *(Browning 1983; Horrocks 2010)*

Dual Methodology Approach

Philological Curation:

- Expert text selection - Critical edition standards - Historical authenticity

+

Computational Processing:

- Systematic analysis - Large-scale pattern detection - Reproducible results

Classical Greek Baseline

Herodotus' *Histories* provides:

- Evidence for **register variation** in Classical Greek
- Contextual baseline for **later developments**
- Comparison point for **linguistic change**

The Koine Greek Core

Foundation of the Early Corpus:

- **Septuagint** (Greek Old Testament, c. 150 BCE)
- **New Testament** (1st-2nd century CE)
- Other Koine texts from the **Hellenistic and Roman periods**

Significance of Koine Greek

Why Koine is Central:

- Transition period between Classical and Medieval Greek - Extensive **translation activity** (Hebrew → Greek) - **Widespread use** across the Mediterranean - Rich evidence for **language contact** phenomena - Foundation for **Byzantine** and **Modern Greek**

Summary: A Comprehensive Resource

The AthDGC corpus provides:

- **More than 100 texts** across multiple languages - **Three millennia** of language history - **Systematic methodology** combining philology and computation - **Translation-based** approach to language contact - Foundation for **diachronic valency research**

Detailed Example: The Septuagint as a Translation Corpus

Historical Context:

- Translated in Alexandria, Egypt (3rd-2nd century BCE) - Hebrew Bible → Koine Greek - Major cultural and linguistic bridge

Research Value:

- Evidence of **Hebrew-Greek language contact** - Translation strategies and techniques - Lexical and syntactic innovations

Example: Septuagint Translation Patterns

Hebrew Source Influence:

```
Hebrew: wayhi → Greek: kai egeneto (literal calque)
"and it came to pass"
```

Translationese Features:

- Word-for-word translation approach - Preservation of Hebrew syntax patterns - Creation of new Greek semantic extensions

New Testament: A Koine Greek Case Study

Corpus Characteristics:

- Written in **Koine Greek** (1st-2nd century CE) - Multiple **authors** with varying linguistic backgrounds - Different **registers** and styles

Linguistic Features:

- Vernacular Koine grammar - Semitic influences from bilingual authors - Range from "sophisticated" to "simple" Greek

Example: Gospel Language Variation

Luke's Gospel:

- More **literary** Koine style - Classical influences - Sophisticated vocabulary

Mark's Gospel:

- More **colloquial** style - Simpler syntax - Oral narrative features

Same content, different linguistic realizations

Byzantine Greek: The Medieval Transition

Time Period: 4th-15th centuries CE

Key Changes:

- Loss of **dative case** (replaced by prepositional phrases)
- Simplification of **infinitive** system
- Development of **periphrastic constructions**

Example: Dative Case Replacement

Classical/Koine Greek:

```
τῷ ἀνθρώπῳ (tō anthrōpō)
"to the man" (dative case)
```

Byzantine/Modern Greek:

```
στον άνθρωπο (ston ánthropo)
"to the man" (preposition + accusative)
```

Modern Greek in the Corpus

Varieties Included:

- **Katharevousa** (formal, archaic) - **Dimotiki** (vernacular, spoken) - Modern translations of ancient texts

Research Focus:

- Completion of the synthetic→analytic shift - Standardization processes - Contemporary translation practices

Example: Multiple Greek Translations Compared

Source: Old Testament passage

Septuagint (150 BCE): Literal from Hebrew

Byzantine translation (10th century): Simplified syntax

Modern translation (20th century): Contemporary idiom

Analysis: How do valency patterns change across 2000+ years?

Germanic Language Family: English Corpus

Historical Stages Included:

- **Old English** (Anglo-Saxon Bible translations)
- **Middle English** (Wycliffe Bible, 14th century)
- **Early Modern English** (King James Bible, 1611)
- **Modern English** (20th-21st century translations)

Example: English Bible Translation Evolution

Same Verse Across Time:

Wycliffe (1382):

"*In the bigynnyng God made of nouyt heuene and erthe*"

King James (1611):

"*In the beginning God created the heaven and the earth*"

NIV (1978):

"*In the beginning God created the heavens and the earth*"

Romance Languages: Latin Foundation

Latin Corpus Components:

- **Vulgate** (Jerome's Latin Bible, 4th century CE) - Medieval Latin texts - Original Latin works for comparison

Why Latin?

- Source language for Romance translations - Direct influence on ecclesiastical Greek - Control for translation directionality

Example: Latin Vulgate Influence

Vulgate Latin:

```
"Fiat lux" (let there be light)
```

Influenced translations in:

- French: "Que la lumière soit"
- Italian: "Sia la luce"
- Spanish: "Hágase la luz"

French in the Corpus

Stages Represented:

- **Old French** Bible translations
- **Middle French** versions
- **Modern French** translations

Research Questions:

- How did French develop its analytical structures?
- What role did Latin play in French syntax?

KEY QUESTION

Digitization process

- How are printed texts digitized?
- What quality control measures are in place?

Computational Processing Pipeline

Stage 1: Text Digitization

- OCR for printed sources - Quality verification (double-checking, error correction)

Digitization is the foundation of the entire project: physical texts must be converted to machine-readable format. For printed texts (19th-21st century), we use OCR technology trained on historical fonts; OCR output is manually reviewed. For manuscripts and early printed books, researchers create electronic editions preserving orthographic features. Some texts undergo re-digitization. The corpus includes metadata tracking digitization method and quality for each text.

KEY QUESTION

Annotation layers

- How are words segmented?
- How are lemmas identified?
- How is morphology tagged?

Computational Processing Pipeline (2)

Stage 2: Linguistic Annotation

- **Tokenization** (word segmentation using language-specific rules) - **Lemmatization** (dictionary form identification via lookup tables) - **Morphological tagging** (grammatical features: case, tense, number, etc.) - **Syntactic parsing** (sentence structure: dependency or constituent analysis)

Linguistic annotation adds layers of linguistic information to raw text. Tokenization segments text into words, handling clitics and contractions appropriately for each language. Lemmatization links inflected forms to lexicon entries (e.g., ἔδωκεν → δίδωμι). Morphological tagging assigns grammatical features using taggers trained on historical language data. Syntactic parsing identifies grammatical relations between words, which is crucial for valency extraction. All annotation undergoes manual quality control, with inter-annotator agreement.

KEY QUESTION

Lexicon design

- What information does each entry contain?
- How are frames linked to examples?
- What enables complex queries?

Computational Processing Pipeline (3)

Stage 3: Valency Extraction

- Identify predicates (verbs, adjectives, nouns requiring arguments) - Extract **argument structures** (what depends on the predicate) - Code **valency frames** (abstract patterns, e.g., NOM+ACC+DAT) - Track changes across time periods (compare frames diachronically)

Each predicate instance is checked for its arguments (subject, objects, complements, adjuncts), and these are abstracted into valency frames. For example, from 100 instances of δίδωμι in a Koine text, the system extracts frame "NOM+ACC+DAT" occurring 87 times and "NOM+ACC" occurring 13 times. This quantitative data will flow into the lexicon, documenting both dominant and minority patterns.

Example: Valency Pattern Extraction

Predicate: δίδωμι (didōmi) "to give"

Classical Greek Valency:

- SUBJ (nominative) + OBJ1 (accusative) + OBJ2 (dative)
- "*X gives Y to Z*"

Modern Greek Valency:

- SUBJ + OBJ1 + prepositional phrase (σε + accusative)
- "*X gives Y to Z*" (different realization)

Valency Lexicon Structure

Each Entry Contains:

- Lemma (base form) - Time period - Source text - Valency frame(s) - Example sentences - Frequency data

The valency lexicon is structured as a relational database where each predicate entry links to multiple frames, and each frame links to attested examples from the corpus. The database schema will include tables for predicates, frames, arguments, and attestations, allowing complex queries like "all verbs that gained a prepositional complement between Classical and Byzantine periods." This structured approach will enable both linguistic research and computational applications.

Valency Theory

Foundational concepts from Tesnière (1959)

What is Valency?

Valency describes the presence/absence, type and number of arguments licensed by a verb:

- *Mary sleeps.* — 1 argument (monovalent)
- *Mary likes John.* — 2 arguments (divalent)
- *Mary gave a letter to John.* — 3 arguments (trivalent)

Underlying Idea (Tesnière 1959):

The verb is the central structural element of the sentence; all other elements depend on it (dependency grammar). Number of participants in semantics = number of arguments in syntax.

Basic Argument Frame

Basic argument frame = the most frequent/default argument set of a verb

e.g. *He saw her.* → 1-nom.V.2-acc

But verbs can have multiple frames:

- *I see.* (= I understand.) — different argument count
- Variation in: number of arguments, case/type of arguments, word order

KEY QUESTION

What are valency alternations?

- How do verbs change their argument structure?
- What patterns exist cross-linguistically?
- How can we track these changes diachronically?

Valency Alternations

(cf. Levin 1993: up to 50 alternations for English)

Middle alternation:

Jane broke the crystal. // The crystal broke.

Causative/Inchoative alternation:

They stood the statue on the pedestal. // The statue stood on the pedestal.

Dative alternation:

They gave him cake. // They gave cake to him.

More Valency Alternations

Unspecified Object Alternation:

Mike ate all the cake. // Mike ate. ✓

*Mike put the flowers in the vase. // *Mike put.* ✗

Understood Reflexive Object:

Jill dressed herself. // Jill dressed. ✓

*Jill cut herself. // *Jill cut.* ✗

Locative Preposition Drop:

Martha climbed up a mountain. // Martha climbed a mountain. ✓

Valency Lexicon Examples

Diachronic tracking of 'write' across English periods

Lexicon Example: Modern English 'write'

Modern English — Frequency: 1015

- "*You write uncommonly fast.*"
- "*...because he does not write with ease.*"
- "*...a person who can write a long letter with ease...*"
- "*...and when I next write to her...*"
- "*...it had been written five days ago.*"

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

Lexicon Example: Middle English 'writen'

Middle English — Frequency: 103

- "*Reum Beel Theem and Samsai the scryuen writen sich oon epistle fro Jerusalem to the kyng Artaxerses*"
- "*Forsothe of his vnclennesse and his vnreligioustee it is writen in the book of the tymes of kyngis.*"

Note: Orthographic variation (writen/written), passive constructions common

Lexicon Example: Old English 'wrītan'

Old English — Frequency: 10

- "*Moyses lyfde þæt man write hiwgedales boc & hi forlete.*"
(Moses allowed the man to write a bill of divorce and abandon her.)
- "*for eower heortan heardnesse he eow wrat þis bebod.*"
(for your heart's hardness he wrote you this command)

Note: Dative recipient (*eow* = you.DAT), no preposition needed

Challenges Encountered

Technical and linguistic obstacles in corpus construction

KEY QUESTION

What challenges arise in historical corpus work?

- Orthographic variation across periods
- Automatic extraction of null subjects
- Annotation schemes for different language stages

Challenge: Orthographic Variation

Old and Middle English Examples:

- *u* instead of *v*, *i* instead of *j*
- Same word, different spellings: *gyue, yyue, yiue, giue* = "give"

Our Decision:

Keep original orthography BUT create custom dictionaries for each text

Consequences:

- Automatic dictionary does not recognize different tokens of same word
- Harder tokenization and morphological parsing

Challenge: Automatic Subject Extraction

Problems:

- Null subjects in pro-drop languages (e.g. Greek)
- Null subject verbs vs avalent verbs (impersonal, weather) — hard to distinguish automatically
- Null objects — even harder to locate

Our Decision:

- Extract all overt subjects
- If no overt subject → mark as expressed by verb suffix (V[subj])
- Manually delete subject from avalent verbs

Challenge: Modern Greek Annotation

Problem:

PROIEL uses traditional metalanguage for Ancient Greek morphology — conflates tense with aspect (aorist = [+past, +perfective])

Modern Greek differs:

- Many periphrastic verbal forms — don't fit 6 traditional categories
- Different tense-aspect system

Our Solution:

New annotation scheme: particle tags for *na/as, tha*; non-finite for perfect periphrases; no subjunctive tag; imperative unspecified for tense

Challenge: Automatic Parsing (Late Medieval Greek)

Evaluation:

Tested 4 parsers trained on Modern or Ancient Greek against reference text (Sphrantzes)

Results:

- Very low F1 scores (<0.42) — poor performance on Late Medieval Greek
- Better results on shorter sentences
- PROIEL has more fine-grained distinctions than UD-CoNLL

Conclusion:

Semi-automatic annotation currently as time-consuming as manual; need better conversion pipeline and trained models

KEY QUESTION

What will the database schema look like?

The relational structure will allow efficient querying across time periods, languages, and semantic classes. Researchers will extract all frames for a given predicate, all predicates with a specific frame type, or all changes between periods. The schema separates abstract frame descriptions from concrete corpus attestations, enabling statistical analysis of frame frequencies.

KEY QUESTION

Statistical overview

- How many texts total?
- How many tokens?
- What is the distribution across periods?

Corpus Statistics Overview

Total Corpus Size:

- **More than 100 texts** across all languages - Approximately **4 million tokens** (words) - Coverage: **3000+ years** of linguistic history

Greek Subcorpus:

- **50+ texts** from Classical to Modern Greek - Dense coverage of **Koine** and **Byzantine** periods - Multiple versions of key texts (translations, retranslations)

KEY QUESTION

Genre diversity

- What genres are represented?
- Why is genre diversity important?
- How does genre affect valency?

Genre Distribution

Major Categories:

- **Religious texts** (Bible, liturgy, patristics): 40% - **Historical/narrative** (chronicles, histories): 25% - **Technical/scientific** (medical, legal): 15% - **Literary** (poetry, fiction): 15% - **Documentary** (papyri, inscriptions): 5%

Rationale:

Genre affects language use significantly. By including multiple genres, we can examine the role of genre in variation and change.

KEY QUESTION

Alignment methodology

- How are parallel texts matched?
- What level of alignment is used?
- What tools support alignment?

Parallel Text Alignment

Alignment Levels: (mainly future work)

- **Document level:** Matching entire texts (e.g., Genesis in Hebrew vs. Greek) - **Chapter/section level:** Subdivisions for navigation - **Verse/sentence level:** Fine-grained correspondence - **Word level:** For detailed translation analysis

Tools Used:

- Manual verification - Existing biblical reference systems (chapter:verse) - Custom alignment for non-biblical texts

KEY QUESTION

Standards compliance

- What international standards apply?
- How is interoperability ensured?
- What metadata is recorded?

Annotation Standards

Key Standards:

- **TEI** (Text Encoding Initiative) for text markup - **PROIEL** tagset for historical languages - **ISO 639** for language codes

Benefits:

- Interoperability with other corpora - Reusability of tools and methods - Long-term data preservation - Community standards compliance

KEY QUESTION

PROIEL integration

- What is PROIEL?
- How does it support this project?
- What data does it provide?

PROIEL Framework Integration

PROIEL (Pragmatic Resources in Old Indo-European Languages):

(Haug & Jøhndal 2008; Eckhoff et al. 2018)

- Established corpus of ancient Indo-European texts - High-quality dependency annotation - Verified morphological analysis - Existing valency research

Integration Benefits:

- Leverages existing annotation work - Builds on proven methodologies - Enables comparison with related projects - Contributes back to PROIEL community

KEY QUESTION

Quality assurance

- How is annotation quality verified?
- What inter-annotator agreement is achieved?
- How are errors corrected?

Quality Control Measures (ongoing and future work)

Annotation Quality:

- **Double annotation** for critical texts - **Expert third review** for difficult cases - **Automatic consistency checks**

Error Correction:

- Regular quality audits - Feedback integration - Version control for all changes - Documented correction history

KEY QUESTION

Multiglossia concept

- What does "grammatical multiglossia" mean?
- How does it affect historical texts?
- Why is it important for this research?

Grammatical Multiglossia

Definition:

The coexistence of multiple grammatical registers or varieties within a single speech community or textual tradition.

In Greek Context:

- Learned texts using archaic grammar - Vernacular texts using contemporary forms - Mixed texts showing both features

Research Implications:

- Cannot assume uniform grammar for any period - Must track register alongside time - Translations may show register shifts

KEY QUESTION

Byzantine evidence

- What registers coexist in Byzantine Greek?
- How do texts vary?
- What does this tell us about language change?

Byzantine Multiglossia Example

High Register (Learned):

- Atticizing style imitating Classical Greek - Preserved case system - Archaic vocabulary

Low Register (Vernacular):

- Simplified case system - Analytic constructions - Contemporary vocabulary

Mixed Texts:

- Chronicles often show both features - Religious texts vary by purpose - Official documents vs. private correspondence

KEY QUESTION

Chronicle significance

- Why are chronicles valuable?
- What linguistic features do they show?
- Which chronicles are included?

Byzantine Chronicles in the Corpus

Included Chronicles:

- **Malalas' Chronicle** (6th century) - **Theophanes' Chronicle** (9th century) - **Chronicle of Morea** (14th century)

Linguistic Value:

- Narrative prose (natural syntax) - Range of registers - Datable compositions - Evidence for vernacular features

Chronicles are particularly valuable because they often show more colloquial features than theological or literary texts.

KEY QUESTION

What are the key findings so far?

- What patterns have emerged?
- What changes are documented?
- What surprises have been found?

Preliminary Research Findings

Documented Patterns:

- **Dative loss:** Systematic across verb classes - **Infinitive replacement:** By να + finite verb - **Preposition increase:** Compensating for case loss - **Word order shifts:** More fixed SVO patterns

Unexpected Findings:

- Earlier vernacular features in "learned" texts - Translation influence on standard grammar - Register-specific preservation of archaic forms

KEY QUESTION

Case study

- Which verb shows clear change?
- What are the before/after patterns?
- When did the change occur?

Case Study: The Verb πείθω (peíthō) 'to persuade'

Classical Greek:

- Active: NOM + ACC (persuade someone) - Middle: NOM + DAT (obey someone)

Byzantine Greek:

- Active: unchanged - Middle: NOM + ACC (dative lost)

Modern Greek:

- NOM + ACC only - Middle: NOM + PP (semantic change)

This verb illustrates the broader pattern of dative loss affecting verb valency.

KEY QUESTION

Verb class patterns

- Which semantic classes change most?
- Are some classes more stable?
- What predicts change?

Verb Classes and Valency Change

Most Changed Classes:

- **Ditransitive verbs:** Dative recipient → prepositional phrase - **Verbs of perception:** Genitive → accusative - **Impersonal verbs:** Dative experiencer → nominative subject

Most Stable Classes:

- **Basic transitive verbs:** NOM + ACC preserved - **Motion verbs:** Directional arguments stable - **Speech verbs:** Core arguments unchanged

Semantic factors (animacy, volitionality) influence which argument marking changes.

KEY QUESTION

Translation effects

- Do translations show different patterns?
- Can translation introduce new structures?
- How long do translation effects last?

Translation Effects on Valency

Observed Effects:

- **Source language calques:** Hebrew patterns in Septuagint Greek - **Conservatism:** Translations often preserve archaic structures - **Innovation:** New constructions to render foreign structures

Example:

The Hebrew cognate accusative construction (e.g., "he feared a great fear") appears in Septuagint Greek but is rare in native Greek texts.

Research Question:

Did Septuagint Greek influence later Byzantine developments?

KEY QUESTION

English comparison

- How has English valency changed?
- What parallels Greek?
- What differs?

English Valency Change: Parallels to Greek

Similar Patterns:

(cf. Trips & Stein 2019; Trips 2020)

- Case loss (Old English → Middle English) - Increased preposition use - Fixed word order development

Differences:

- English: Almost complete case loss - Greek: Retained nominative/accusative distinction - English: Stronger word order constraints

Research Value:

Comparison reveals which changes are language-specific vs. typologically common.

KEY QUESTION

Technical challenges

- What are the main processing difficulties?
- How are historical languages handled?
- What NLP limitations exist?

Computational Challenges

Historical Language Processing:

- **Limited training data** for NLP models - **Orthographic variation** (no standardization) - **Ambiguity** without native speaker intuition

Solutions:

- Custom models trained on annotated historical texts - Rule-based preprocessing for normalization - Human-in-the-loop verification - Cross-linguistic transfer learning

Modern NLP achieves 90%+ accuracy on modern languages; historical languages can reach 80% with specialized models???

KEY QUESTION

Long-term planning

- How will data be preserved?
- How will the project continue after funding?
- What open access provisions exist?

Sustainability and Open Access

Data Preservation:

- Multiple backup locations - Standard formats for longevity - Institutional repository hosting

Open Access Commitment:

- Corpus freely available upon completion - Documentation and tools shared - CC-BY licensing for derived works

Continuation Plans:

- Integration with existing digital humanities infrastructure - Community maintenance model - Ongoing annotation contributions

KEY QUESTION

Collaborative network

- What institutions are involved?
- What expertise is contributed?
- How is collaboration structured?

International Collaborations

Partner Institutions:

- **University of Oslo** (PROIEL project)
- **Harvard CHS** (Digital resources)

Collaboration Benefits:

- Shared methodological expertise - Access to existing resources - Broader linguistic coverage - International validation

KEY QUESTION

Teaching uses

- How can the corpus support education?
- What learning resources will be created?
- Who are the target learners?

Educational Applications

For Students:

- Interactive valency exploration - Historical grammar exercises - Translation comparison tools

For Teachers:

- Ready-made example sets - Progression tracking resources - Cross-linguistic comparison materials

For Researchers:

- Training data for NLP - Benchmark datasets - Reproducible research resources

KEY QUESTION

Future plans

- What languages might be added?
- What features will be developed?
- What is the long-term vision?

Future Extensions

Planned Additions:

- **More languages:** Slavic, Armenian - **More genres:** Technical, legal, epistolary - **More periods:** Archaic Greek, Late Latin

Feature Development:

- Enhanced visualization tools - Automated change detection - Machine learning integration

Future Phases:

- Corpus expansion

Broader Impact: Digital Humanities

Contribution to Field:

- Model for **large-scale** historical corpus projects
- Integration of **traditional** and **digital** methods
- **Sustainable** infrastructure design
- **Replicable** methodology

Theoretical Significance

Why This Matters:

- Understanding **mechanisms** of language change
- Documenting **long-term** syntactic evolution
- Revealing **contact effects** in written language
- Building **empirical foundation** for diachronic theory

Practical Applications Beyond Academia

Real-World Uses:

- **Translation technology** (machine translation of historical texts)
- **Language learning** resources
- **Cultural heritage** preservation
- **Digital archive** infrastructure

Contact and Resources

Project Website: https://athdgc.github.io/ [Coming soon]

Institution: University of Athens

Funding: H.F.R.I. - ELIDEK Excellence Grant

Selected References (1/2)

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O. E., & Jøhndal, M. (2018). The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1), 29–65.

Haug, D. T. T. & Jøhndal, M. L. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the 6th LREC*. ELRA.

Kranich, S., Becher, V., & Höder, S. (2011). A tentative typology of translation-induced language change. In S. Kranich et al. (Eds.), *Multilingual Discourse Production* (pp. 11–43). Amsterdam: John Benjamins.

Lavidas, N. (2021). *The Diachrony of Written Language Contact: A Contrastive Approach*. Brill's Studies in Historical Linguistics. Leiden: Brill.

Lavidas, N., Bergs, A., van Gelderen, E., & Sitaridou, I. (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.

Sitaridou, I. (2014). The Romeyka Infinitive: Continuity, Contact and Change in the Hellenic varieties of Pontus. *Diachronica*, 31(1), 23–73.

Selected References (2/2)

McLaughlin, M. (2011). *Syntactic Borrowing in Contemporary French: A Linguistic Analysis of News Translation*. Oxford: Legenda.

Sitaridou, I. (2016). Reframing the phylogeny of Asia Minor Greek: The view from Pontic Greek. *CHS Research Bulletin*, 4(1), 1–17.

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*, 12(1), 232–267.

Trips, C. (2020). Copying of argument structure: A gap in borrowing scales. In B. Drinka (Ed.), *Historical Linguistics 2017* (pp. 409–430). Amsterdam: John Benjamins.

Nikiforidou, K. (2018). Genre and constructional analysis. *Pragmatics & Cognition*, 25(3), 543–575.

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. *Journal of Quantitative Linguistics*, 12, 167–184.

Thank You!

Questions & Discussion

AthDGC Project · Athens Digital Glossa Chronos