

AI Applications for Language Analysis

Instruments, analysts, and verifiers

Stergios Chatzikyriakidis

ILSP

CLARIN:EL Summer School, 24 September 2026

Language Technology and Artificial Intelligence for Digital Humanities

Which AI?

Four families of methods, often used together. Name the configuration behind a result.

Symbolic: rules, logic, ontologies, proofs

Rhyme: Greek G2P and the Topintzi taxonomy.

Zeugma: Prolog validates extracted events.

RAG-to-Coq and semantics: proofs or refusals.

HeptaTAX: discriminative legal formulae.

TermGuard: ISO 1087-1 checks, alongside an LLM.

Statistical and machine learning

Linguistic Distance and PhylProGramm: 16 languages, stability weights and phylogenies over ASJP, WALs, Grambank and UD. Perplexity: an LM's surprise as distance. Detectors: classifiers and probability scores.

Neural analysis: taggers and encoders

NATS and Voyant-NLP: spaCy entities, character networks and text statistics. RoBERTa: a trained detector.

Generative language models

Claude, GPT-4o, Gemini, Llama and Krikri, plus prompts, retrieval, tools and verifiers.

Rhyme, HeptaTAX, BookForge and chatbots: propose, draft and extract.

A number belongs to the whole configuration.

Much of the checking and measuring here uses symbolic rules or statistics.

Three ways to put AI to work on language

1. Instrument

The system **measures**.

Feature-based distance between Ancient and Modern Greek; perplexity as dialectal distance; detector scores as authorship evidence.

2. Analyst

The system **labels**.

Entities and character networks with spaCy; rhyme types, legal genres and historical events with LLMs.

3. Proposer and verifier

A model **proposes**, symbolic AI **verifies**.

Rules, gazetteers, logic programs, proof assistants.

The question is the same in all three: **how do we know it did it right?**

The answer comes from the humanities. LLMs are one tool in the box, and rarely the one doing the checking.

The frame: injective and federative neuro-symbolic systems

Chatzikyriakidis & Lappin (2026), *Neuro-symbolic NLP: taxonomy, assessment, and directions*. *Frontiers in Artificial Intelligence*, **9**, article 1797587.

Injective

Symbolic structure is pushed *inside* the network:
as training signal, as architecture, as a loss.

Dominates the literature. Gains are usually marginal.

Federative

Neural and symbolic components stay *separate*
and exchange messages: the model proposes, a
solver, a rule base or a proof assistant checks.

Fewer papers. Larger, more reliable gains. Not yet
scaled to foundation-model size.

- Combined taxonomy: Kautz's composition codes plus Lappin's injective/federative distinction.
- Everything in Part III of this talk is federative. The symbolic side is where **your** knowledge goes.

The model as instrument

Perplexity as dialectal distance

- **GRDD+**: 6.3M words, 10 Greek varieties (Chatzikyriakidis et al., 2026).
- Score each variety's text with a model. **Perplexity** is how surprised the model is by the text.
- The model's surprise *ranks* varieties by their distance from the Greek it was trained on.
- Seven models, three sample sizes. The table shows eight varieties plus Standard Modern Greek.

Caveats: tokenisation drives the absolute numbers (Llama 2 splits Greek into 5–6 tokens per word, Krikri into 1.5–2.4); a distance is not a genealogy.

Variety	PPL, Krikri-8B
Standard Modern Greek	8.8
Heptanesian	12.4
Cypriot	24.7
Maniot	24.7
Pontic	37.2
Northern	41.6
Cretan	62.1
Tsakonian	96.5
Griko	121.6

25,000-word samples.

What to do about the distance: Markantonatou, Bompolas, Stamou & Dimakis, today 11:30

- Sixteen Indo-European languages in ancient-to-modern pairs: Ancient → Modern Greek, Latin → Romance, Old Church Slavonic → Slavic, Gothic → German, Sanskrit → Hindi.
- Seven dimensions of distance, each from a curated resource.
- Stability-based weighting (F-ratio): which features resist change across the transitions?
- Distance-based and character-based phylogenies, with uncertainty quantification.

Dimension	Resource
lexical	Swadesh lists
typological	WALS, Grambank
featural	URIEL
syntactic	UD treebanks
phonological	ASJP
cognates	CLDF cognate sets

What resists change

- **Cognate distance** carries 48% of the stability-based weight. The comparative method, vindicated by the numbers.
- Only **26 of 72** WALS features change in fewer than 20% of the historical transitions.
- **Six** features never changed in the sample, among them adposition order and possessive marking.
- Family-specific: case marking is stable in Slavic and unstable in Romance.

Lesson

“AI for language analysis” is not only large language models. Curated databases plus statistics answer questions no LLM can, and they say what they measure.

Detection as authorship attribution

- The oldest question in philology: *who wrote this?* The 2026 version: *was it a model?*
- **Corpus:** 1,550 baseline texts from eight model families (GPT-4o, GPT-5.5, Llama 3.3 70B, Mistral Large 3, DeepSeek V4 Pro, Kimi K2.6, Phi-4 reasoning, Grok 4.1) plus a human reference corpus.
- **Perturbation:** round-trip machine translation, $\text{en} \rightarrow \text{X} \rightarrow \text{en}$, pivots fr, el, zh, ja, ar, and double hops ($\text{en} \rightarrow \text{X} \rightarrow \text{ja} \rightarrow \text{en}$). 13,950 round trips in the supplied results.
- **Detectors:** Binoculars (perplexity ratio) and Fast-DetectGPT (probability curvature), both zero-shot and a fine-tuned RoBERTa classifier, an encoder, not a generative model.
- **Calibration:** true-positive rate at a fixed false-positive rate on the human corpus. A detector that flags humans is not a detector.

What a detector measures

Texts by Llama 3.3 70B, true-positive rate

Condition	Binoculars	RoBERTa
original	93%	57%
en→fr→en	85%	57%
en→el→en	82%	59%
en→zh→en	54%	42%
en→ja→en	50%	45%
en→ar→ja→en	26%	43%

- Evasion **scales with typological distance** of the pivot and with hop count.
- Probability-based detectors measure “how model-like is this surface”. Translation rewrites the surface.
- The trained classifier does not collapse: it picks up something translation preserves. Joint evasion needs paired predictions.
- English-only detector performance on Greek requires a separate evaluation.

For the humanities

A detector score is a measurement with a theory behind it. Know the theory before trusting the number.

The model as analyst

Rhyme in Modern Greek: the task

By stress position (Topintzi et al., 2019)

- M, oxytone: καρδιά / φωτιά
- F2, paroxytone: ελπίδα / πατρίδα
- F3, proparoxytone: στόματα / σώματα

Independent features

- Rich: onset of the stressed syllable matches (καλά / μαλά)
- IDV: the pre-stress vowel matches (ξανθή / γραφή)
- Mosaic: the rhyme domain spans a word boundary (όνομά της / ο μπάτης)
- Imperfect: systematic vowel or consonant variation (χάνετε / γίνετε); Copy: same lemma

- The domain is the **phonological word**: a host plus its weak pronoun. The three-syllable rule applies.
- A token-based model sees letters in chunks; rhyme lives in syllables, stress and sound.
- **Corpus**: 40,000+ rhymes from the Anemoskala and Interwar Poetry corpora, cleaned and released.

Chatzikyriakidis & Natsina, *LLMs Got Rhyme? Hybrid Phonological Filtering for Greek Poetry Rhyme Detection and Generation*, LaTeCH-CLfL 2026 (EACL). Code: github.com/StergiosCha/Greek_Rhyme

What models hear

Identification accuracy, best configuration per model

Model	Acc.	Needs
Claude 4.5	53.8%	CoT + RAG
GPT-4o	50.0%	CoT (7.7% without)
Claude 3.7	42.3%	RAG or CoT
Gemini 2.0	42.3%	RAG
Mistral Large	26.9%	CoT
Llama 3.1 8B	15.4%	RAG

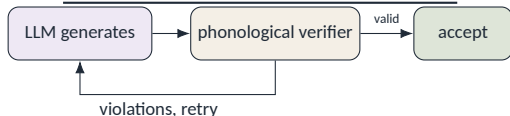
By stress, Claude 4.5: M 65.6%, F2 47.5%, **F3 21.9%**

- **F3** (proparoxytone) is the hardest for every model.
- **Mosaic** rhymes: **0%** across all models and prompts. These tested configurations missed rhyme domains crossing word boundaries.
- Homophones flagged as copies: κρίνοι / κρίνει, “lilies” and “judges”, judged the same word.
- GPT-4o: 7.7% to 50% with chain-of-thought. It knows more than it lets on, but must be walked through it.
- **Everyone in this room can see these errors without a leaderboard.**

Generation: propose, verify, refine

Valid poems, rhyme constraints fully satisfied

Model	Alone	With verifier loop
Claude 3.7	3.8%	73.1%
GPT-4o	0.0%	42.3%
Claude 4.5	0.0%	34.6%



- Ask for a poem with a rhyme scheme and features: almost nothing comes back fully valid.
- Add a deterministic verifier (Greek grapheme-to-phoneme rules plus the taxonomy) and loop on its report.
- The verifier is a 2019 linguistics paper turned into code. No training.
- Still unsolved: combinations such as IDV + mosaic stay at 0% even with the loop.

HeptaTAX: notarial acts of sixteenth-century Corfu

- **1,088 acts** by five Corfiot notaries, 1500–1567; editions digitised at the Laboratory of Southeastern Mediterranean Linguistics (Aegean).
- Non-standard orthography, Greek/Latin/Italian code-switching, formulaic but variable legal language.
- **Three tiers**: 17 core genres (Venditio, Testamentum, Emphyteusis, ...), extensions, hybrid tags. 40-act gold benchmark.
- Twelve LLMs $\times$ four settings: zero-shot, few-shot, full-context, **neuro-symbolic**.

After the image becomes text (Gatos, 14:30), this is the next step.

The symbolic engine

Deterministic rules detect the discriminative legal formulae; their output goes into the prompt. The model may override.

- Gap between strongest and weakest model: **47.5 $\rightarrow$ 12.5 points**.
- Llama 3.1 8B gains 47.5 points and matches frontier models running without symbolic support.
- Ceiling on the core tier: 72.5%. Extensions about 63%, hybrids about 35%.
- Errors cluster in procedurally dense acts: lexical cues do not identify *legal effect*.

More analysts, same lesson

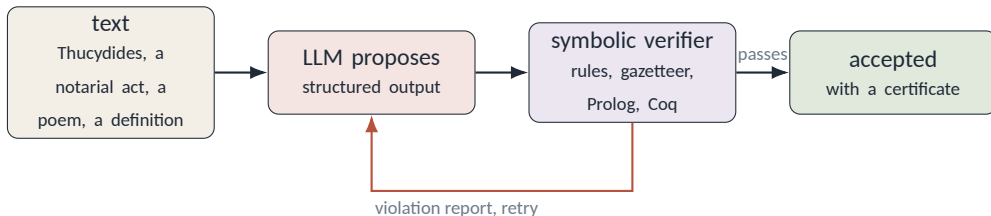
- **NATS**: entities, character networks and text statistics for Greek literature with spaCy. No LLM, no invented entities, its own error profile.
- **Greek parliamentary speech**: stance and rhetoric at scale, labels audited (doctoral work under supervision).
- **Curriculum compliance**: scoring textbooks against the curriculum by thematic field, from the ministry's PDFs.
- **OYXOY NLI** and *Reverse-Engineering NLI* (COLM 2025): models pass benchmarks by routes never intended.
- **Gender bias in Greek pronoun resolution** (Llama, Krikri): the analyst has priors. *Friday, Gkirtzou.*
- **Interpretable mixture of experts** for dialect text: which expert fires on which feature. *Friday, Paraskevopoulos.*

Common thread

A label from a model is a hypothesis. It becomes a finding when a philologist, a rule, or a benchmark with teeth has checked it.

**The model as proposer, the
humanities as verifier**

The pattern



- Federative: the two sides never share weights.
- The verifier cannot be charmed by fluent prose.
- The verifier encodes a taxonomy, a gazetteer, a chronology, a definition theory.
- That is **your** knowledge, written so a machine can refuse.

Zeugma: reasoning over Thucydides

- Task: extract events from the *History* as RDF (actor, action, place, year); enrich with Wikidata, DBpedia and ConceptNet; resolve places to coordinates.
- Model RDF is dirty: duplicate identifiers, malformed literals, “Here are the events...” preambles. Apache Jena parses **0 of 32** outputs.
- **Zeugma**: fault-tolerant parser (53% strict, 100% in recovery mode), RDF → Prolog, user-defined rules. 12× faster than Jena.
- Claude Sonnet 4, GPT-4o, Llama 70B, Llama 8B, eight extraction configurations each.

Why this text

Precisely dated, densely located, argued: a stress test for keeping events, places and years consistent.

κτῆμά τε ἐς αἰὲ μᾶλλον ἢ ἀγώνισμα ἐς τὸ παραχρῆμα ἀκούειν ξύγκειται. (1.22)

Workshop interface in MEDEA-NEUMOUSA. Paper venue: CLASP Concluding Conference.

Rules catch what syntax cannot

A temporal rule in Prolog

```
suspicious_date(E, Y) :-  
    hasYear(E, S), atom_number(S, Y),  
    Y > -100, Y < 0.
```

“Battle of Sybota in year −8”: a well-formed triple, an impossible date.

Negation as failure

Missing locations can be tested with Prolog negation or SPARQL FILTER NOT EXISTS. Prolog also supports conditional recursive rules.

- 333 catalogued errors (13 temporal, 290 geographic, 30 agency). The paper reports no false positives on clean data.
- Simulated self-correction: a violation report motivates changing “2024” to −433 for Sybota. The supplied demonstration scripts the correction.

The point

The rules are the historian’s knowledge: chronology, geography, who can act on whom. The model never sees them as weights.

- Same source, stronger verifier: extracted events become **Coq propositions**, and the proof assistant checks their consistency with what is already established.
- Five configurations to see what each component buys: **zero-shot**, **base**, **KG-only**, **RAG-only**, **full**.
- Token-aware chunking, multi-knowledge-graph integration, caching; an interactive proving interface with the live goal state.

Why a proof assistant

A proof either closes or it does not. There is no partial credit, no fluent wrong answer, and the certificate can be inspected by someone who does not trust the model.

Why it matters for a historian

The pipeline is forced to say which claims it relies on. That is a bibliography, mechanised.

- ISO 1087-1: a definition names the **genus** (the superordinate concept) and the **differentiae**. Definition by genus and differentia: *Topics*, *Posterior Analytics*.
- **TermGuard** checks a candidate definition: is the genus a recognised concept (grounded in open lexical and ontological resources)? Is it circular? Do the differentiae distinguish?
- Then terminology-aware translation and export as TBX (ISO 30042), for Greek and the other EU languages.

The proposer

A model drafts terms, equivalents and definitions from domain text.

The verifier

A definitional theory 2,300 years older than the technology, plus a reference termbase built only from authoritative, licensed, machine-readable collections.

BookForge: the pattern at production scale

- Greek children's picture books: **story bible** → **drafting board with critics** → **symbolic validation** → reference-conditioned illustration → PDF in publishing-house presets (Οσελότος, Υδροπλάνο, ...).
- «Το Σχολείο μέσα στο Παραμύθι»: exercises per primary-school grade, aligned with the curriculum, grounded in the story. Every arithmetic answer is **computed by code**, not by the model.
- An author's own story can be imported verbatim: the pipeline then does illustration and layout only.
- Built on Azure Foundry models; the critics and validators are the same propose-and-verify loop as the rhyme system, one level up.

Same pattern, different verifier

Rhymes: a phonological rule set. Thucydides: Prolog and Coq. Terms: ISO 1087-1. Children's books: continuity checks, vocabulary level, a calculator.

Formalising semantics in Coq: the rigor gap

- A Coq library mechanising influential papers in formal semantics: 31 files, more than 235 theorems.
- Montague, Kratzer, Rooth, Dowty, Link, Champollion, Barwise & Cooper, MTT adjectives and mass/count, Greek polydefinites, presupposition projection, DPL.
- **File Change Semantics** resists mechanisation: non-determinism, global state, implicit choices.

What formalisation shows

What a theory actually commits to. Kratzer's modal semantics turns out to be an interface to existing modal logic rather than new machinery. Some "formal" semantics is less formal than it looks.

A method for the humanities

Write the theory down so a machine can refuse it. The refusals are the findings.

Reaching people

When grounding is a safety requirement

SimasiaAI, Greek-language applications

- Assistant for ΚΑΠΑ3 (cancer patient guidance centre): oncology patients' social and welfare rights. Delivered free.
- BPAN Companion, for a rare-disease patient association (pro bono).
- Assistant for the multiple-sclerosis patient federation (ΠΟΑμΣΚΠ).
- SimasiaEdu for tutoring centres; DialogosAI.

The constraint

A wrong answer about a welfare deadline hurts a real person. So: retrieval over verified documents only, refusal when the sources do not support an answer, and people in the loop.

- The same architecture as Part III, with a human as the last verifier.
- EU AI Act compliance is a design input, not an afterthought.

Takeaways, and where the day goes next

1. **Instrument**: know what the number measures. Perplexity measures surprise; a detector measures surface likeness to a model.
2. **Analyst**: the label is a hypothesis. Verify it with something that cannot be charmed.
3. **Proposer**: let the model propose; let a rule, a gazetteer, a proof assistant, or a philologist decide.
4. “AI” is four families of methods. Name the one you used. Most checking and measuring here is symbolic or statistical.

Your expertise is the verifier.

Today

11:30 Low-resource varieties:

Markantonatou, Bompolas, Stamou, Dimakis

12:15 Ancient Greek: Platanou

14:30 Historical document images:
Gatos

Friday

Social biases: Gkirtzou

Explainability: Paraskevopoulos

Greek NLP Swiss Knife (Azure Container Apps)

- MEDEA-NEUMOUSIA, with Zeugma and the knowledge-graph extractor
- Greek Rhyme System
- TermGuard
- Dialect Generator
- Linguistic Distance
- NATS, PlotAnalyzer, Voyant-NLP

Svarna lab site

- Greek Corpus Workbench
- Datasets: GRDD+, OYXOY, the rhyme corpus
- One page that links every demo

Entry point: stergioscha.github.io/svarna/demos.html

You can keep using all of this after Thursday. The workshop apps provide access; an external chat comparison can use your existing account.

A. The analyst who cannot hear (15 min)

Paste a stanza of Solomos into a chat model and ask for the rhyme types. Then run it through the Greek Rhyme System. Find the stress and mosaic errors yourselves.

B. Propose, then verify (25 min)

Pick a passage of Thucydides on the Zeugma page of MEDEA-NEUMOUSA. A model extracts events; Prolog rules check chronology, geography and agency. Read the violation report. Feed it back and check whether the next graph improves.

Start here: stergioscha.github.io/svarna/demos.html

Workshop access is provided in your browser.

A: Greek Rhyme System B: MEDEA Zeugma

If time remains: the Dialect Generator (write a Cretan sentence, judge it) and Linguistic Distance (Ancient to Modern Greek across dimensions), as a bridge to 11:30.

Thank you

`schatzikyriakidis@athenarc.gr`

`stergioscha.github.io/svarna`

Funding acknowledgment

Funding from Microsoft's LINGUA project supported CoDAG (Computational Dialectal Atlas of Greek), making Svarna, GPT-4.1 fine-tuning and the Greek NLP Swiss Knife possible.