HeptaTAX: A Neuro-Symbolic Pipeline and Benchmark for Classifying 16th-Century Heptanesian Notarial Acts

Stergios Chatzikyriakidis[1], **Eleni Karantzola**[2], **Vasiliki Makri**[3]
[1] University of Crete
[2] University of the Aegean
[3] National and Kapodistrian University of Athens
stergios.chatzikyriakidis@uoc.gr, karantzola@aegean.gr, vasmakri@enl.uoa.gr

Abstract

This study originates in the investigation of lexical bundles and formulaic language within sixteenth-century Corfiot notarial documents. The observed functional variation across identical formulaic sequences motivated the development of a document classification framework designed to support the structural interpretation of such language. Given that 16th-century Corfiot notarial acts represent a rich, albeit understudied, dialectal resource, their systematic categorization into subgenres is essential for their full exploration. However, this task requires substantial manual work, while NLP tools for this task and dialect do not exist. In this paper, we attempt to take an initial step in this direction. First, we present a corpus of 1,088 notarial acts from 5 notaries spanning 1500-1567, a 3-tier annotation schema (17 core genres, extension subcategories, hybrid cross-cutting tags), and a 40-act benchmark with gold annotations at all three tiers. Then, we evaluate 12 LLMs across 4 architectures, zero-shot, few-shot, full-context and Neuro-Symbolic. For the latter, we introduce a symbolic engine comprising a set of deterministic rules for identifying discriminative legal formulae, whose output is then injected into the neural (LLM) engine. The results show that the NeSy architecture compresses the accuracy gap between stronger and weaker models from 47.5 pp to 12.5 pp, with the smallest model (Llama 3.1 8B) gaining 47.5% and matching frontier models that operate without symbolic support. Three models reach a ceiling of 72.5% on the core tier. However, consistent errors in procedurally dense material reveal the limits of lexical and formulaic cues for identifying legal effect, motivating the use of symbolic signals in the NeSy pipeline. Extension and hybrid classification remain open challenges, with best scores of $\sim$63% and $\sim$35% respectively.

Keywords: NLP, neuro-symbolic, Heptanesian Greek, Corfiot, notarial acts, document taxonomy

1. Introduction

The sixteenth century marks a crucial phase for the study of post-Byzantine vernacular / Early Modern Greek, characterized by extensive linguistic variation and the absence of a standardized norm (Karantzola, 2024; Kakoulidi-Panou et al., 2023). Within this context, notarial documents from Corfu provide a uniquely valuable corpus: they are precisely dated, geographically localized, and shaped by recurring formulae that capture the interaction between professional writing practices and regional linguistic features. Classifying these notarial acts into legal genres (Venditio, Testamentum, Emphyteusis, etc.) is essential both for historical interpretation and for the linguistic analysis of formulaic language. However, this process requires domain expertise, remains extremely time-consuming when performed manually, and is further complicated by editorial inconsistency: not all editors classify documents in the same way, and proposed categorizations may be guided by surface lexical cues rather than the legal effect of the instrument. On the other hand, performing this task computationally using NLP techniques is not straightforward either, since systems have to face non-standard orthography, code-switching (Greek/Latin/Italian), formulaic but variable legal language, and domain vocabulary absent from LLM training data. We present a first attempt to tackle this problem via HeptaTAX offering the following three research contributions:

1. A new corpus of $\sim$1,088 notarial acts from 5 sixteenth-century Heptanesian[1] and in particular Corfiot notaries along with a 40-act benchmark including 3-tier genre annotations.
2. A taxonomy of 17 core legal genres with extensions and hybrid tags based on the close reading of notarial acts by experts.
3. A neuro-symbolic pipeline that combines deterministic symbolic rules with LLMs.

We then evaluate 12 LLMs (both open and close weight ones) across 4 architectures (zero-shot, few-shot, full context, and NeSy), effectively showing that symbolic knowledge injection can help in reducing the gap in performance between stronger

[1] Despite its acknowledged contribution to the formation of Standard Modern Greek, Heptanesian Greek, which is spoken in the Ionian Islands (as well as Kythira), forms a distinct and cohesive dialectal group characterized by a set of grammatical and lexical features that set them apart both from the standard language and from mainland dialects; internal variation among the islands further supports their independent status (Liosis, 2024).

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 265–273
16 May 2026.

and weaker models.[2]

2. Related Work

Within the Early Modern period (16th–18th centuries), notarial acts constitute a key linguistic source for the study of vernacular Greek, as their precise dating, geographical localization, and formulaic structure enable reliable contextualization and systematic corpus-based analysis (Karantzola et al., 2012; Manolessou, 2003; Makri, 2020; Wagner et al., 2013). The classification of notarial documents into documentary types has long been a central concern of diplomatics, where distinctions such as *instrumenta* and *acta* reflect functional differentiation in written practice (Gasparini, 2023; Duranti, 2015). In digital corpora, document typology constitutes essential metadata for both historical interpretation and computational processing, motivating recent work on the automatic classification of historical documents (Ehrmann et al., 2023). At the same time, the strong formulaicity of notarial writing has prompted linguistic research on recurring lexical bundles and formulaic sequences (Biber et al., 2004; Wray, 2002).

NLP work for Greek dialects is not abundant but has seen a surge in recent years. For example, Chatzikyriakidis et al. (2023) introduce GRDD, a large-scale raw text dataset for Greek dialectal NLP that covers four MG dialects, i.e., Cretan, Pontic, Northern Greek, and Cypriot. This has developed into GRDD+ (Chatzikyriakidis et al., 2026a), further extending the resource to 10 varieties (6.3M words), including Heptanesian, and also performing fine-tuning experiments on Llama and Krikri models on four dialects. There is also important work on treebanks for Greek dialects: Eastern Cretan (Vakirtzian et al., 2025) and the dialect of Lesbos (Bompolas et al., 2025), both using cross-dialectal knowledge transfer from Standard Modern Greek. Vakirtzian et al. (2024) also introduced a speech recognition benchmark for four Greek varieties. Chatzikyriakidis et al. (2026b) apply perplexity as a metric for measuring dialectal distance across Greek varieties. Earlier attempts to provide computational resources include the work by Anastasopoulos et al. (2018), who present a POS-tagged dataset for Griko, and the Multi-CAST version for Cypriot Greek by Hadjidas and Vollmer (2015). There is also dialectometric work for Greek dialects. There, we find the work by Bompolas (2023) that applies traditional dialectometric distance-based methods to Asia Minor varieties (Cappadocian, Pharasiot, and Silliot) and also Psaltaki et al. (2025) that deals with loanword detection of Turkish and Italian in four dialects of Modern Greek. These are very important initiatives to develop resources and models for Greek dialectal NLP, but none of them deal with the classification of historical documents, let alone documents in 16th-century Heptanesian.

On the neuro-symbolic (NeSy) side, our system is a Federative-Reasoning (F-R) NeSy system under the Chatzikyriakidis-Lappin taxonomy (Chatzikyriakidis and Lappin, To appear), i.e. one in which the symbolic and neural components operate as fully autonomous modules that communicate only through their outputs, without any modification of the internals of either component. Similar systems have been proposed for various NLP tasks, but a detailed discussion of such systems falls outside the scope of this paper. However, there is a closely related F-R system that uses a hybrid phonological system to deal with the identification and generation of Greek rhyme (Chatzikyriakidis and Natsina, 2026). There, a deterministic rhyme-detection module verifies and refines LLM-generated Modern Greek poetry, raising accuracy from 4% to 73%. The symbolic component in the system described in this paper plays an analogous role: it preprocesses the input and passes structured evidence to the LLM, which retains the freedom to override it.

3. The HeptaTAX Resource

3.1. Corpus

The corpus comprises five published registers of Corfiot notaries, originally edited from manuscript sources and digitized at the Laboratory of Southeastern Mediterranean Linguistics (University of the Aegean). The texts were processed via OCR, normalized to monotonic orthography, and verified against the editions, including manual correction and even typing where required. The corpus contains 1088 acts, distributed per notary as follows:

- Alexakis: 104 acts (1513–1516)
- Farmakis: 173 acts (1515–1525)
- Spyris[3]: 375 acts (1560–1567)
- Toxotis: 282 acts (1500–1503)
- Varagas: 154 acts (1541–1545)

The sixteenth-century Corfiot notarial documents reflect an early stage in the formation of the Heptanesian varieties and provide a valuable source for historical dialectology despite the constraints of

[2]The data and the code needed to reproduce the results can be found here: https://osf.io/vj2fr/overview?view_only=a9cccb01cae148bd9c095e50d687a261

[3]It should be noted that a subset of 19 acts was authored by the notary Photios Palatianos.

their legal-administrative register. Although historical orthography may obscure phonological developments in progress, the texts combine conservative features with emerging morphophonological, morphological, and lexical innovations, indicating an ongoing phase of linguistic reorganization rather than a stabilized dialectal system (Karantzola and Lavidas, 2016; Karantzola and Makri, in print). Variation across scribes and communities points to both diatopic (geographical) and diastratic (sociolinguistic) differentiation, while pervasive orthographic fluctuation reflects the absence of a fixed written norm. Overall, the corpus occupies an intermediate position between administrative convention and vernacular usage, preserving regional traits while documenting broader processes of linguistic change.

3.2. Benchmark

In order to test the ability of NLP models to classify the genre of notarial acts, we sample 40 acts from the corpus, corresponding to 8 acts per notary. Two experts then annotate them using a 3-tier label scheme that includes:

- Core layer: one of 17 legal genre categories (Table 1)
- Extension layer: subcategory where applicable (e.g., Venditio → Venditio cum pacto retrovendendi)
- Hybrid tags: cross-cutting features (e.g., "cum pacto", "agricultural credit")

Annotation was performed collaboratively by two domain experts, who jointly assigned labels through discussion until consensus was reached.

3.3. Taxonomy

The two experts also built a taxonomy of 17 core legal genre categories and organized them into 6 thematic groups (Duranti, 2015; Fellmeth and Horwitz, 2021):

I. *Contractus de rebus* (property): Venditio, Donatio, Permutatio, Emphyteusis, Locatio conductio rei

II. *Contractus de credito* (credit): Mutuum, Instrumentum obligationis, Quietantia, Pignus/Hypotheca

III. *Contractus operarum* (production): Locatio conductio operis, Societas, Praevenditio

IV. *Acta familiaria* (family/personal): Testamentum, Dos, Divisio/Partitio

V. *Acta proceduralia* (procedural): Procuratio, Renuntiatio, Revocatio

VI. *Acta ecclesiastica*: Concessio ecclesiastica, Beneficium

The Extension layer then captures subtypes of the main taxonomy (e.g., emphyteusis perpetua, locatio vineae), while the hybrid layer captures cross-cutting features that span multiple core types. The taxonomy is presented in trilingual form: Latin / English / Greek.

4. Methodology

We evaluate four prompting strategies across 12 LLMs. The first one is a **zero-shot** setting, where the model receives only the act text together with the list of 17 valid core categories and an instruction to classify. The second is a **few-shot** setting, where the prompt is augmented with 5 manually annotated examples covering diverse genres (Venditio, Praevenditio, Emphyteusis, Procedural acta, and Concessio ecclesiastica), each accompanied by its gold 3-tier label. The third is a **full-context** setting, where the model additionally receives the complete genre taxonomy manual, including definitions of all 17 core categories, extension subcategories, and hybrid tags. The fourth strategy is a **neuro-symbolic (NeSy)** setting. Here, we pair the full-context prompt with the output of a deterministic symbolic component. This component comprises ∼70 hand-crafted rules that pattern-match on formulaic phrases in the normalised act text. The symbolic rules map a phrase to a genre and carry a weight, ranging from 3 (near-unique to one genre) to 1 (weakly associated). Any matched rule is collected and are then used to rank the candidate genres by cumulative weight, assigning, at the same time, a confidence level based on the top score and the margin over the runner-up. The result is a symmary that includes the predicted genre, confidence, and supporting phrases, that is passed to the LLM with the instruction to treat this information as defeasible. It is important to note that the symbolic component runs locally before the API call and, as such, imposes no extra inference cost.

The NeSy system proposed is an instance of an **Federative-Reasoning (F-R)** architecture in the Chatzikyriakidis-Lappin taxonomy (Chatzikyriakidis and Lappin, To appear), where the symbolic engine is totally autonomous from the neural component. The idea is that the output of the symbolic engine is passed to the neural component (the LLM in our case) without the involvement of any modification of the internals of each component. The motivation for using a NeSy system in our case is quite standard: LLMs generalise well across a remarkable number of tasks but sometimes fail because they lack domain-specific knowledge, while, on the other hand, hand-crafted rules encode precise formulaic

Group	Core Genres	N
I. Contractus de rebus (property)	Venditio, Donatio, Permutatio, Emphyteusis, Locatio conductio rei	5
II. Contractus de credito (credit)	Mutuum, Instrumentum obligationis, Quietantia, Pignus/Hypotheca	4
III. Contractus operarum (production)	Locatio conductio operis, Societas, Praevenditio	3
IV. Acta familiaria (family/personal)	Testamentum, Dos, Divisio/Partitio	3
V. Acta proceduralia (procedural)	Procuratio, Renuntiatio, Revocatio	3
VI. Acta ecclesiastica	Concessio ecclesiastica, Beneficium	2
	Total	**17**

Table 1: The 17 core legal genres organised in 6 thematic groups following the Latin legal tradition adapted for Venetian-ruled Greek practice. The extension layer captures subtypes (e.g., *emphyteusis perpetua*, *locatio vineae*); the hybrid layer captures cross-cutting features that span multiple core types.

Normalised phrase	Genre	W
ψιχικις μοι σοτιριας *'for my soul's salvation'*	Testamentum	3
εποιλισεν *'sold'*	Venditio	3
χανονιν *'canon / annual rent'*	Emphyteusis	3
πλιρομενος χε ειχαριστιμενος *'paid and satisfied'*	Quietantia	3
ις τιν ερχομενιν εσοδιαν *'in the coming harvest'*	Praevenditio	3
χριτε αλπιτρε *'judges arbiters'*	Procedural acta	3
δια τιμιν *'for a price'*	Venditio	2
επαρεδοσεν *'delivered' (generic transfer)*	Venditio	1

Table 2: Sample symbolic rules. Each rule maps a normalised formulaic phrase to a target genre with a discriminative weight (W): 3 = near-unique to one genre, 2 = strongly associated, 1 = weakly associated (shared across genres). The full rule set contains ∼70 rules.

knowledge but break easily when faced with open text (Ebrahimi et al., 2024).

Table 2 presents a sample of the symbolic rules, illustrating how genre-diagnostic formulaic phrases from the notarial tradition are mapped to target genres with varying discriminative weights.

All 12 models ran through the OpenRouter API at temperature 0.0. We have a good selection of models that cover both proprietary and open models but also of different sizes and architectures (thinking vs non-thinking). Most specifically we use 6 proprietary models (Claude 3.7 Sonnet, Claude 4 Sonnet, Claude 4 Opus, GPT-4o, GPT-5.2 Chat, and Gemini-2.5-Flash) and six open-weight models (Llama 3.3 70B, Llama 4 Maverick, DeepSeek V3, DeepSeek R1, Qwen 2.5 72B, and Llama 3.1 8B). Each model is asked to produce a structured JSON response with three fields as output: `core_layer`, `extension_layer`, and `hybrid_tags`. In total, 12 models × 4 strategies × 40 acts = 1,920 classifications.

5. Results and Analysis

The results are summarized in two tables: Table 3 reports the accuracy for the core layer only, while Table 4 reports the results for the 3-tier evaluation (core, extension, hybrid). We first discuss the non-symbolic strategies before turning to the NeSy results.

Zero-shot and few-shot. Starting with the zero-shot setting, models range from 17.5% (Llama 3.1 8B) to 65.0% (GPT-5.2 Chat and Claude 3.7 Sonnet), with a mean of 53.3%. The main factor here seems to be model scale. Smaller models (Qwen 2.5 72B, Llama 3.3 70B) as well as DeepSeek V3 score below 55%, while frontier models cluster in the 57.5–65% range. The addition of five annotated examples in the few-shot setting gets the mean up to 56.2%. However, the effect is not stable across models. Some models benefit considerably (DeepSeek V3: 45.0% → 57.5%; Qwen 2.5 72B: 45.0% → 55.0%), while for others it is detrimental to their performance (e.g. Llama 4 Maverick drops sharply from 60.0% to 37.5%, and GPT-4o from 62.5% to 57.5%). This suggests that the five examples may introduce a distributional bias that misleads some models.

Full-context. This is the case in which the models receive the whole taxonomy manual. Accuracy mean is raised to 58.1%. An interesting fact is that the increase is most visible in models that seem to struggle with the few-shot setting (Llama 4 Maverick recovers to 60.0%; Qwen 2.5 72B remains flat at 45.0%). However, the strongest gains are for the two Claude models used (Claude 3.7 Sonnet reaches 70.0%; Claude 4 Opus also reaches 70.0%). Full-context is the preferred strategy for Claude 4 Opus (70.0% vs 67.5% NeSy) and a competitive one for DeepSeek R1 (60.0%).

Neuro-symbolic (NeSy). The NeSy setting achieves the highest mean accuracy at 67.7% and is the best or tied-best strategy for 10 out of 12 models, with the two exceptions being Opus 4 and DeepSeek R1 as noted above. The top three NeSy models—GPT-4o, GPT-5.2 Chat, and Llama 4 Maverick—all reach 72.5%, followed by Claude 3.7 Sonnet and Claude 4 Sonnet at 70.0%. In the extension layer, the full-context and NeSy settings are close (best in 5/12 and 6/12 models, respectively). In the hybrid tags layer, accuracy remains low across the board (25–35%) with no clear winner among strategies. Notably, the NeSy setting considerably improves small models on extensions (Llama 8B: 15.4% full-context $\rightarrow$ 48.7% NeSy).

One of the main findings is that the NeSy approach acts as an equalizer for weaker and stronger models. For example, Llama 3.1 8B goes from an accuracy of 17.5% in the zero-shot setting to 65.0% in the NeSy setting (47.5-point increase), suggesting that good domain-specific rules can compensate for model scale.

There is an accuracy ceiling at 72.5%. Three architecturally different models (GPT-4o, GPT-5.2, Llama 4 Maverick) all plateau there. The remaining errors might point to task-inherent limitations. In particular, the performance ceiling we observe at 72.5% is likely due to the the inherent ambiguity of the historical legal register, where multiple juridical effects overlap within a single instrument, rather than from architectural constraints of the models.

DeepSeek-R1, the only model with mandatory chain-of-thought generation, is not helped by NeSy and achieves its best score in the few-shot setting. Moving to the frontier instruction-following models, results are mixed: Sonnet 3.7 and Opus 4 perform best in the full-context setting, while Sonnet 4 and GPT-5.2 benefit from the symbolic component.

Llama 3.1 8B in the NeSy setting (65.0%) presents a very interesting cost-efficiency ratio case. The 8B model is $\sim$150$\times$ and $\sim$750$\times$ cheaper per input and output token, respectively, than Claude Opus 4 (70% in full context setting, 67.5% in the NeSy setting). Proper NeSy architectures can make these much smaller models useful for similar tasks at a much lower cost.

Comparing full-context vs NeSy, we see evidence that providing the full taxonomy helps, but curated symbolic evidence on top of the full-context (NeSy) provides better performance in 10/12 models for the core layer. Moving on to the extension layer results, we find a slightly different picture. There, full-context settings outperform NeSy in 5/12 models (GPT-5.2: 62.9% vs 59.5%; Opus 4: 57.9% vs 48.7%; R1: 60.0% vs 51.4%). However, NeSy still improves the performance of small models on extensions. For example, Llama 8B goes from 15.4% (full-context) to 48.7% (NeSy), and LLama 3.3 70B. Lastly, the Hybrid tag remains challenging. Models struggle there within the 25-35% across all strategies, with no clear winner.

6. Error Analysis

Looking at patterns of confusion, we see that the errors cluster around categories with overlapping legal vocabulary. Examining the 48 attempts per act (12 models $\times$ 4 strategies), the most frequent confusions are: Procedural acta mistakenly labeled as Procuratio (47 cases), Locatio conductio rei as Emphyteusis (47 cases), Concessio ecclesiastica as Procuratio (46 cases), Procedural acta as Mutuum (46 cases), and Procuratio as Quietantia (44 cases).

Acts involving multiple legal transactions (e.g., sale combined with pledge) are among the hardest to classify, since the core layer forces a single label. It is also important to note that the symbolic rules do not cover a number of rare categories (e.g., Divisio or Beneficium), and as such, classification of these types relies entirely on the LLM's knowledge.

6.1. Procedural Registers as a Stress Test: The Toxotis Case

The Toxotis documents are different w.r.t the rest of the corpus. It is packed with procedural documents and relies heavily on institutional and delegation language. Unlike registers that mainly involve transactional acts, Toxotis' documents mostly deal with procedural maneuvering, for example they handle issues like appointments, declarations, setting up arbitration, or representing someone in a specific dispute.

Given this setup, the models, no matter the configuration, seem to assume they are looking at transactional documents, and as such, they greatly over-predict categories like Procuratio, property deals, and obligation-related classes. The results are quite telling: 5/8 Toxotis acts in the benchmark were missed by all 12 models across all 4 strategies (scoring 0 out of 48 attempts for each). The remaining 3 acts managed 2 correct predictions (out of 144 total attempts). The result is a 99.5% failure rate overall. It is important to note that these are not scattered mistakes, given that in most cases, all 12 models zero in on the exact same incorrect category, no matter the prompting strategy. For example, Act 130 (gold: Procedural acta / Declaratio) is classified as Mutuum by every model in every setting; Act 170 (gold: Procuratio) is unanimously labelled Quietantia; and Acts 156 and 137 are both called Procuratio by all 48 attempts despite their distinct gold labels (Procedural acta and Concessio ecclesiastica, respectively).

Misleading vocabulary appears to be the main

Model	Zero-shot	Few-shot	Full-context	NeSy
GPT-4o	62.5	57.5	62.5	**72.5**
GPT-5.2 Chat	65.0	70.0	70.0	**72.5**
Llama 4 Maverick	60.0	37.5	60.0	**72.5**
Claude 3.7 Sonnet	65.0	65.0	70.0	**70.0**
Claude 4 Sonnet	57.5	67.5	62.5	**70.0**
Claude 4 Opus	60.0	62.5	**70.0**	67.5
DeepSeek V3	45.0	57.5	55.0	**67.5**
Qwen 2.5 72B	45.0	55.0	45.0	**67.5**
Gemini 2.5 Flash	60.0	57.5	62.5	**65.0**
Llama 3.1 8B	17.5	30.0	22.5	**65.0**
Llama 3.3 70B	45.0	52.5	57.5	**62.5**
DeepSeek R1	57.5	**62.5**	60.0	60.0
Mean	53.3	56.2	58.1	**67.7**
Range (max–min)	47.5	40.0	47.5	**12.5**

Table 3: Core layer accuracy (%) across 12 LLMs and 4 prompting strategies on the 40-act benchmark. Bold indicates best strategy per model. NeSy = neuro-symbolic pipeline with symbolic evidence injection.

	Zero-shot			Few-shot			Full-context			NeSy		
Model	Co	Ex	Hy	Co	Ex	Hy	Co	Ex	Hy	Co	Ex	Hy
GPT-4o	62.5	40.0	25.0	57.5	16.0	30.0	62.5	43.8	32.5	**72.5**	46.9	30.0
GPT-5.2 Chat	65.0	26.9	9.1	70.0	25.0	27.0	70.0	**62.9**	34.2	**72.5**	59.5	**35.0**
Llama 4 Maverick	60.0	33.3	18.9	37.5	16.0	30.8	60.0	51.4	**37.5**	**72.5**	**55.3**	25.0
Claude 3.7 Sonnet	65.0	36.8	25.0	65.0	23.1	22.2	70.0	52.8	26.3	**70.0**	**53.8**	**30.0**
Claude 4 Sonnet	57.5	8.3	14.3	67.5	20.8	27.8	62.5	**56.8**	25.6	**70.0**	45.0	**27.5**
Claude 4 Opus	60.0	14.8	23.1	62.5	20.8	27.0	**70.0**	**57.9**	**30.0**	67.5	48.7	27.5
DeepSeek V3	45.0	40.0	25.0	57.5	28.6	26.3	55.0	48.6	27.5	**67.5**	**46.2**	**27.5**
DeepSeek R1	57.5	38.5	21.1	**62.5**	31.0	27.5	60.0	**60.0**	**32.5**	60.0	51.4	26.3
Qwen 2.5 72B	45.0	36.8	21.1	55.0	16.7	21.2	45.0	51.5	30.0	**67.5**	**57.1**	25.0
Gemini 2.5 Flash	60.0	41.0	25.0	57.5	20.0	28.9	62.5	**55.9**	**30.0**	**65.0**	52.8	25.0
Llama 3.1 8B	17.5	40.0	3.2	30.0	12.5	19.4	22.5	15.4	25.0	**65.0**	**48.7**	**25.0**
Llama 3.3 70B	45.0	42.5	21.1	52.5	16.0	19.4	57.5	50.0	27.5	**62.5**	**52.8**	**35.0**

Table 4: Complete 3-tier evaluation (%) across 12 LLMs and 4 strategies. Co = core genre; Ex = extension subcategory (24/40 acts have ground truth, percentage is calculated on this number); Hy = hybrid tags. Bold = best strategy per model per tier.

reason behind misclassification. The texts are full of verbs about giving, receiving, or appointing, as well as references to property, which can easily fool the model. The Toxotis documents present a consistent pattern of procedural misidentification. The models seem to get carried away by standard diplomatic wording, particularly the delegation formulas, and immediately assume that they are dealing with a transaction or a representation act. They completely miss the fact that the document is merely setting up someone's participation in a specific legal procedure. This inflates the representation of certain categories and under-identifies procedural acts such as Declaratio, Arbitrium, and Renuntiatio.

A related phenomenon is what may be termed "frame dominance," in which the diplomatic structure of an instrument outweighs its legal effect in the model's prediction. Despite frequently containing formulaic language associated with contracts, the texts from Toxotis are procedural in terms of the resulting act. This issue reveals a central problem for legal NLP in historical diplomatic materials. It shows that the surface wording of an instrument does not reliably reflect its true legal function.

The Toxotis register is also important from the standpoint of the NeSy system presented in this paper, i.e. it showcases that symbolic rules keyed to legal effect bypass the diplomatic framing that misleads the classifier by surface lexical cues and formulaic language. Procedural texts are a useful stress test for NeSy precisely because improvements there cannot be explained by pattern memorization alone.

Concluding, we can say that the Toxotis examples can function as a stress test for the design of our taxonomy, as well as annotation stability, since it reveals category boundaries that are more difficult to see in more standard transactional registers. Furthermore, these findings elucidate the overall performance trends, as they show that the

performance of a model might shift systematically depending on the diplomatic register and that a high density of procedural language heavily impacts category recognition. In the end, the Toxotis examples seem to back up our broader point: in order to reliably classify historical legal documents, one has to identify the primary legal effect instead of just relying on vocabulary or standard formulas.

7. Conclusion and Future Work

In this paper, we presented HeptaTAX, which consists of the following: a) a corpus of 1,088 sixteenth-century Corfiot notarial acts, b) a 40-act benchmark with 3-tier genre annotations, c) a taxonomy of 17 core legal genres, and d) a neuro-symbolic pipeline for automatic classification. The NeSy pipeline achieves 72.5% core accuracy, a 14.4% improvement over the mean zero-shot baseline, and compresses the accuracy gap between the strongest and weakest models from 47.5 pp to 12.5 pp. The symbolic engine acts as an equaliser across model sizes: Llama 3.1 8B gains 47.5 pp with NeSy and matches frontier models operating without symbolic support. To our knowledge, HeptaTAX is the first computationally accessible corpus of Heptanesian notarial acts with structured genre annotations. The NeSy approach is applicable to other historical document classification tasks where domain expertise can be formalised as rules, representing a form of knowledge transfer from humanities to NLP.

Beyond its computational contribution, HeptaTAX provides a structured framework for linking documentary typology with formulaic language, offering a reproducible basis for future linguistic and dialectological investigation of Early Modern Greek.

For future work, we will provide a full annotation of the whole corpus (1088 acts) in a semi-automated way, using the best-performing NeSy pipeline for pre-annotation followed by human correction. We will expand the symbolic rules to cover rare categories and procedural acts. Furthermore, we would like to extend this approach to other territories that were under Venetian rule and have similar notarial traditions (e.g. Crete, Cyprus. Lastly, we would like to attempt fine-tuning of a small model on the annotated data to see the effect of fine-tuning and how well this fares when compared to prompt engineering and our NeSy system.

Another relevant research direction is systematically modelling lexical bundles and diplomatic formulas in the form of signals that link surface textual patterns to juridical effect. The results presented in this paper point to formulaic language playing a central role in shaping classification behaviour (particularly in procedurally dense registers). The explicit representation of formula structures, whether statistically and/or symbolically, can provide an intermediate layer that is able to connect textual variation with legal function. Lastly, implementing juridical effect as a computationally tractable category allows one to connect historical diplomatics with NLP and shows that legal function (and not merely surface wording) can be effectively used in document classification.

In a broader sense, this type of research points to a research agenda where historical legal NLP goes beyond document classification and moves towards modelling the interaction between formulaic expressions, diplomatic register, and juridical effect. Thus, this approach can contribute to more robust effect-oriented classification and to legal NLP systems that are more explainable and transparent, given their grounding in historical documentary practice.

8. Limitations

The 40-act benchmark is small; a larger annotated set would strengthen the claims. The rules are hand-crafted and specific to the Venetian-Greek notarial tradition and would need adaptation for other legal traditions or periods. Temperature was set to 0.0 throughout, meaning each model produced a single deterministic run with no variance analysis. It would be interesting to check temperature variations across many different runs. The extension and hybrid classification tasks require more refinement, with their best scores being at $\sim$63% and $\sim$35%, respectively. Lastly, the Toxotis case shows that procedural registers require dedicated annotation and rule development that are lacking in the current work.

9. Bibliographical References

Antonios Anastasopoulos, Alessandro Cattelan, Kevin Duh, Christian Federmann, and Chris Callison-Burch. 2018. Part-of-speech tagging on an endangered language: A parallel Griko-Italian resource. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 29–35. Association for Computational Linguistics.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and curriculum-based writing. *Applied Linguistics*, 25(3):371–405.

Stavros Bompolas. 2023. *Computational Dialectology in the Linguistic Varieties of Cappadocian, Pharasiot, and Silliot*. Ph.D. thesis, University of Patras.

Stavros Bompolas, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2025. Crossing dialectal boundaries: Building a treebank for the dialect of Lesbos through knowledge transfer from Standard Modern Greek. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 39–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Stergios Chatzikyriakidis and Shalom Lappin. To appear. Neuro-symbolic NLP: Taxonomy, assessment, and directions. *Frontiers in Artificial Intelligence*. Research Topic on Neural-Symbolic NLP: Bridging Theory and Practice.

Stergios Chatzikyriakidis and Anastasia Natsina. 2026. LLMs got rhythm? Hybrid phonological filtering for Greek poetry rhyme detection and generation. In *Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2026)*. Association for Computational Linguistics.

Stergios Chatzikyriakidis, Dimitris Papadakis, Sevasti-Ioanna Papaioannou, and Erofili Psaltaki. 2026a. GRDD+: An extended Greek dialectal dataset with cross-architecture fine-tuning evaluation. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026)*.

Stergios Chatzikyriakidis, Erofili Psaltaki, Dimitris Papadakis, Erik Henriksson, and Veronika Laippala. 2026b. Perplexity as a metric for dialectal distance: A computational study of Greek varieties. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2026), EACL 2026*.

Stergios Chatzikyriakidis, Chatrine Qwaider, Ilias Kolokousis, Christina Koula, Dimitris Papadakis, and Efthymia Sakellariou. 2023. GRDD: A dataset for Greek dialectal NLP. In *Proceedings of the 16th International Conference on Greek Linguistics (ICGL16)*.

Luciana Duranti. 2015. *Diplomatics: New Uses for an Old Science*. Rowman and Littlefield, Lanham, MD.

Monireh Ebrahimi, Pascal Hitzler, Md Kamruzzaman Sarker, et al. 2024. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web*, 15(4):1291–1329.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2).

Aaron X. Fellmeth and Maurice Horwitz. 2021. *Guide to Latin in International Law*, 2 edition. Oxford University Press.

Silvia Gasparini. 2023. Notaries and the law in Venice: Development of a discipline. *Italian Review of Legal History*, 9:1–33.

Harris Hadjidas and Maria C Vollmer. 2015. Multi-CAST Cypriot Greek. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*.

Eleni Kakoulidi-Panou, Eleni Karantzola, and Katerina Tiktopoulou. 2023. *Vernacular Prose of the 16th Century*. Greek Language Center, Athens. In Greek.

Eleni Karantzola. 2024. What do we know about and what can we learn from Early Modern Greek? In *Proceedings of the 15th International Conference on Greek Linguistics*, volume I, pages 49–68, Belgrade. University of Belgrade. In Greek.

Eleni Karantzola and Nikolaos Lavidas. 2016. Characteristics of 16th-century Corfiot. *Studies on the Greek Language*, 36:129–150. In Greek.

Eleni Karantzola and Vasiliki Makri. in print. A comparative analysis of 16th-century Corfiot and Cephalonian: Evidence from notarial records. In *Proceedings of Modern Greek Dialects and Linguistic Theory 10*.

Eleni Karantzola, Katerina Tiktopoulou, and Katerina Frantzi. 2012. Notarial documents as sources for Early Modern Greek. In *Early Modern Greek Public Secretariat Language, Tradition, and Poetics*, pages 473–501. Vikela Municipal Library, Heraklion. In Greek.

Nikos Liosis. 2024. Heptanesian dialects. In G. Giannakis, editor, *Encyclopedia of Greek Language and Linguistics Online*. Brill.

Vasiliki Makri. 2020. *The Formal Expression of Grammatical Gender in a Modern Greek Dialect Affected by Italo-Romance*. Ph.D. thesis, University of Patras, Department of Philology.

Io Manolessou. 2003. Non-literary sources as evidence for the language of the medieval period. *Lexicographic Bulletin*, 24:61–88. In Greek.

Erofili Psaltaki, Yves Scherrer, and Stergios Chatzikyriakidis. 2025. Italian and Turkish loanwords detection in Greek dialects. In *Proceedings of the 17th International Conference of Greek Linguistics (ICGL 2025)*, Cambridge, UK.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint*

25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 776–784, Tallinn, Estonia. University of Tartu Library.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. Speech recognition for Greek dialects: A challenging benchmark. In *Interspeech 2024*, pages 3974–3978.

Esther-Miriam Wagner, Ben Outhwaite, and Bettina Beinhoff, editors. 2013. *Scribes as Agents of Language Change*. Mouton de Gruyter, Berlin.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge.