From Corpus to Concept Scheme: Towards Developing a SKOS Vocabulary for Armenian Epigraphic Heritage

Hamest Tamrazyan
DHLAB, EPFL, Lausanne, Switzerland
hamest.tamrazyan@epfl.ch

Kamal Nour
EPFL, Lausanne, Switzerland
kamal.nour@epfl.ch

Emanuela Boros
DHLAB, EPFL, Lausanne, Switzerland

Abstract

Armenian epigraphy, one of the world's oldest and most diverse inscriptional traditions, remains largely absent from digital research infrastructures due to a lack of basic linguistic and conceptual resources. No machine-readable corpus, standardized terminology, or controlled vocabulary exists for describing Armenian inscription types, preventing indexing and interoperability. This paper addresses this gap by constructing the first dataset of Armenian inscription-type terminology and by developing a computational pipeline for analyzing it at scale. We digitize and preprocess a broad corpus of authoritative printed publications; curate a culturally grounded terminology list; and train transformer-based NER models to identify both attested inscription types and potential terminological variants across unseen texts. The resulting resources form the first empirical foundation for modelling Armenian epigraphic concepts needed for further developing a SKOS vocabulary aligned with, yet culturally distinct from, existing international epigraphic ontologies.

1 Introduction

In the last decade, digital technologies have transformed the way cultural heritage is preserved, accessed, and studied. As inscriptions, manuscripts, monuments, and other heritage artefacts increasingly enter digital ecosystems, the question of how cultural knowledge is organized has become a central concern for the digital humanities (DH) (Yan et al., 2020; Bianchini, 2023). Classification systems such as vocabularies, ontologies, and semantic models, play a decisive role in shaping what becomes visible, searchable, interoperable, or even thinkable within digital infrastructures (Liu et al., 2023). However, the act of classification is far from being neutral and determines *which cultural traditions are accurately represented*, *which are simplified or misread*, and *which remain invisible altogether*. A direct consequence of this shift toward digital heritage is that traditions lacking structured, machine-readable vocabularies cannot be effectively indexed, searched, or linked within modern research infrastructures. This is the case for Armenian epigraphy: *despite the depth and historical significance of its corpus, the field remains only minimally integrated into existing digital standards*.

Thus, the need for a dedicated SKOS vocabulary[1] for Armenian epigraphy arises from the fact that, despite their historical significance (Figure 1), Armenian inscriptions remain largely absent from the semantic infrastructures used in digital humanities (Greenwood, 2004). This gap reflects structural barriers that limit the visibility and preservation of Armenian epigraphic heritage in digital environments (Tamrazyan et al., 2026; Tamrazyan and Hovhannisyan, 2024a,b).

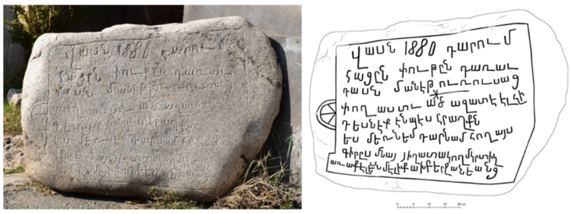

Figure 1: Inscription of Vahravar (Figure 5a, b in (Harutyunyan, 2023a)).

A first challenge is the **lack of standardized, machine-readable descriptive terminology** in Armenian epigraphy, which prevents inscriptions from being systematically indexed, linked, or compared across databases. In the absence of controlled vocabularies, Armenian material cannot be effectively integrated into international corpora, limiting its visibility for researchers working in epigraphy, archaeology, linguistics, and cultural heritage studies.

[1] https://www.w3.org/2004/02/skos/

Proceedings of SIGHUM 2026, pages 1–10
March 28-29, 2026 ©2026 Association for Computational Linguistics

Second, the **structural bias of existing epigraphic ontologies toward Greco-Roman traditions** prevents Armenian inscriptions from being accurately modelled (Espinosa Espinosa and Velázquez Soriano, 2021). Ontologies such as EAGLE[2] and FAIR Epigraphy[3] rely on typologies and conceptual hierarchies that do not reflect the cultural logic or material diversity of Armenian monuments. This misalignment leads to inappropriate categorizations and forces Armenian material into external frameworks that lack cultural specificity.

Third, the highly dispersed nature of Armenian heritage further increases **the need for interoperable digital infrastructures**. With more than 80% of Armenian monuments located outside the Republic of Armenia (Tigranyan, 2023), often in regions facing conflict, neglect, or political pressure, the ability to represent inscriptions in a consistent, structured, and internationally accessible format becomes crucial for cultural preservation and scholarly continuity.

Fourth, **the absence of interoperable vocabularies** has broader implications for computational research (Tamrazyan and Hovhannisyan, 2025). Without a standardized terminology layer, the application of natural language processing, machine learning, semantic web technologies, or automated annotation workflows becomes severely limited. Establishing a SKOS vocabulary therefore functions not only as a heritage documentation effort but as an enabling infrastructure for future digital research in Armenian studies.

Finally, there is a growing recognition within the DH of the **importance of supporting underrepresented cultural traditions** and challenging the implicit Eurocentrism embedded in many digital heritage standards. Developing Armenian-specific vocabularies contributes to this broader movement by ensuring that local epistemologies and scholarly traditions are formally encoded rather than marginalized.

This article contributes to this broader effort by offering the first corpus-driven analysis of inscription-type terminology in Armenian epigraphy and by identifying the requirements for developing a culturally grounded SKOS vocabulary. Rather than importing externally defined categories, the study derives its insights from distinctions that emerge within Armenian inscriptional practice and considers how these may be aligned with international standards. The data and code underlying this study are openly available[4].

Specifically, this work contributes with:

- **A curated terminology dataset** derived from foundational Armenian epigraphic publications and produced through a dedicated OCR and text-processing workflow.

- **A terminology-aware NER model** trained to identify and classify inscription-type expressions across Armenian-language scholarship[5].

- **A corpus-based foundation for conceptual modelling** that identifies key terminological patterns and conceptual distinctions required for further constructing a SKOS vocabulary of Armenian inscription types.

2 Data Normalization

To examine how inscription-type terminology functions across Armenian epigraphic scholarship, we develop a computational workflow designed to digitize the corpus, extract terminology, and analyze its distribution at scale. The workflow proceeds through several stages: assembling a representative set of printed publications, digitizing them into machine-readable text, preprocessing the OCR output into a clean sentence-level corpus, constructing a ground-truth NER dataset using an expert terminology list, and training transformer-based NER models to evaluate both in-domain performance and generalization to unseen books.

2.1 Adoption of the EAGLE Model

As a conceptual starting point, we adopt the seven-domain organizational structure of the EAGLE vocabularies, one of the most widely used international models for describing inscriptions in digital epigraphy. These domains—*Material*, *Execution Technique*, *Type of Inscription*, *Object Type*, *Decoration*, *Dating Criteria*, and *State of Preservation*—provide a shared descriptive template that

[2] https://www.eagle-network.eu/resources/vocabularies/

[3] https://ontology.inscriptiones.org/type_of_inscription

[4] https://github.com/dhlab-epfl/dhlab-epigraphy-studies

[5] The pipeline is designed to be transferable to other under-resourced epigraphic traditions, with adaptation mainly required for OCR, sentence segmentation, and expert terminology curation.

Book	Abbrev.	Citation
Divan Hay Vimagrut’, yan 1	DHV–1	Orbeli (1965b)
Divan Hay Vimagrut’, yan 2	DHV–2	Orbeli (1960)
Divan Hay Vimagrut’, yan 3	DHV–3	Barkhudaryan (1967b)
Divan Hay Vimagrut’, yan 4	DHV–4	Barkhudaryan (1973b)
Divan Hay Vimagrut’, yan 5	DHV–5	Barkhudaryan (1982b)
Divan Hay Vimagrut’, yan 6	DHV–6	Avagyan and Janpoladyan (1977b)
Divan Hay Vimagrut’, yan 7	DHV–7	Grigoryan (1996)
Divan Hay Vimagrut’, yan 8	DHV–8	Grigoryan (1999)
Divan Hay Vimagrut’, yan 9	DHV–9	Barkhudaryan et al. (2012b)
Divan Hay Vimagrut’, yan 10	DHV–10	Sargsyan et al. (2017b)
Historical Monuments of Akunk and Katnaghbyur	KAR–2014	Karapetyan (2014)
Khodjavank Monastery	KAR–2024a	Karapetyan (2024)
Northern Artsakh	KAR–2021a	Karapetyan (2021c)
Lapidary Inscriptions of Bun Aghvank	KAR–2021b	Karapetyan (2021b)
Armenian Collection of the Caucasian Museum	KAR–2004	Karapetyan (2004)
Previously Unpublished Inscriptions of Yerevan Katoghike	HAR–2019a	Harutyunyan (2019a)
Epigraphic Heritage of Armavir Province	HAR–2017	Harutyunyan (2017)
Newly Discovered Tombstones of Holy Ejmiatsin	HAR–2021	Harutyunyan and Melkonyan (2021)
Settlement of Noragavit and St. Gevorg Church	HAR–2019b	Harutyunyan (2019b)
Inscribed Artefacts of the N. Adonts Museum of Sisian	HAR–2019c	Harutyunyan (2019b)
Epigraphic Heritage of Tatev Hermitage	HAR–2023	Harutyunyan (2023b)
Natural Disasters and Celestial Phenomena in Inscriptions	HAR–2022/2023	Harutyunyan (2022, 2023a)
Melik Mansions of Artsakh and Syunik	GHU–2001	Ghulyan (2001)
Inscriptions of Armenian Settlements in India	KOR–2024	Kortoshyan (2024)
Inscriptions of Aleppo	KOR–2013	Kortoshyan (2013a)
Tsaghatskar Monastery: Historical and Archaeological Study	MEL–2024	Melkonyan (2024)
Birthplaces of Genocide Survivors in Lebanese Funerary Inscriptions	TAS-2018	(Tashjian, 2018)
Georgian State Policy and Armenian Cultural Monuments (1988–1998)	KAR-1998	(Karapetyan, 1998)
Reconstruction of Dadivank	AYV-2011	(Ayvazyan, 2011)

Table 1: Publications included in the corpus, with abbreviations used throughout the paper. Light-gray rows indicate the volumes used to construct the full training dataset for the main experiments. The earliest baseline model was trained solely on **DHV–10**. The remaining publications (non–highlighted rows) were held out as unseen material for evaluating the model's ability to generalize and to detect potential new inscription-type terminology.

would allow Armenian inscriptions to be positioned within the existing epigraphic infrastructures.

For the present study, we focus specifically on the *Type of Inscription* domain that captures the functional category of an inscription (e.g., "commemorative", "funerary", "votive", "legal", "donor", "construction-related") and interacts closely with linguistic, material, and historical characteristics. To operationalize this conceptual layer for Armenian epigraphy, we first required a stable and culturally grounded vocabulary of inscription-type terminology. For this, a team of domain specialists compiled an initial inventory of terms that represent inscription types. They also used a large language model (LLM) as an assistive tool (GPT–5) to expand this inventory by generating additional variants drawn from the digitized corpus; however, because these automatically proposed terms varied in accuracy, each term candidate was manually reviewed by experts, who retained only those terms that were terminologically sound, historically attested, and conceptually distinct.

The resulting curated list comprises **41 Armenian inscription-type multiword terms**[6]. These include general descriptors such as *վիմագիր* and *վիմական արձանագրություն*; functional classes such as funerary (*տապանագիր*), donor (*նվիրագիր*), commemorative (*հուշագիր*), legal (*կանոնագիր*), and construction-related inscriptions (*շինարարական արձանագրություն*); architectural subtypes specifying the placement of inscriptions on a monument, including *գավիթային*, *սյունային*, *բարավորի*, and *որմնագիր*; administrative and socio-economic categories such as tax-exemption inscriptions (*ապահարկման արձանագրություն*) and water-management inscriptions (*ջրօգտագործման արձանագրություն*); and rare or specialized types, including cryptographic inscriptions (*շրջագիր / ծածկագիր*), royal and princely inscriptions, monastic inscriptions, and manuscript-associated

[6] A term can be composed of only one term e.g. *վիմագիր* or multiple words (multiword) e.g. *վիմական արձանագրություն*

colophonic writings. We refer to this set of terms as our **expert-defined inscription type reference list** and will serve us in creating an annotated dataset for detecting new and undocumented inscription-type terms.

2.2 Corpus Construction

Because no digital corpus of Armenian epigraphy currently exists, we construct our dataset manually from authoritative printed publications. The complete set of works forming the corpus is listed in Table 1. These sources include multi-volume corpora, regional surveys, architectural studies, thematic monographs, and peer-reviewed articles, and together represent the most comprehensive and methodologically rigorous body of scholarship available for Armenian inscriptions.

Our primary source is the monumental *Divan Hay Vimagrut'yan* (DHV) series, published by the National Academy of Sciences of Armenia since the 1960s (Orbeli, 1965a,a; Barkhudaryan, 1967a, 1973a, 1982a; Barkhudaryan et al., 2012a; Avagyan and Janpoladyan, 1977a; Barkhudaryan, 1999; Sargsyan et al., 2017a). The DHV volumes remain the most detailed and methodologically consistent corpus of Armenian inscriptions, offering reliable transcriptions, functional classifications, and palaeographic descriptions. They therefore serve as the backbone of our terminology inventory.

To avoid a DHV-centric vocabulary and to reflect the broader diversity of Armenian epigraphic scholarship, we also incorporate terminology from regional corpora and thematic studies authored by leading specialists such as Karapetyan, Kortoshyan, Petrosyan, Barkhudaryan, and Harutyunyan (Karapetyan, 2021a; Kortoshyan, 2013b; Karapetyan, 1998, 2004; Kortoshyan, 2013a, 2024; Petrosyan, 2008; Harutyunyan, 2017). These works capture terminological variation across regions, generations, and disciplinary perspectives, and contribute essential conceptual depth to the corpus. The publications in Table 1 were selected according to four criteria designed to ensure representativeness, conceptual precision, and cultural fidelity.

Authorship Diversity. We include works authored across different generations, institutions, and scholarly traditions. This diversity reduces reliance on any single terminological style and captures diachronic variation in Armenian epigraphic discourse.

Genre Balance. To reflect both stable and emerging terminology, we intentionally balance different scholarly genres. Books and monographs provide detailed historical, architectural, and linguistic analyses; non-digital corpora of inscriptions, especially the DHV volumes, offer authoritative transcriptions and typological classifications; and peer-reviewed articles and conference proceedings capture recent conceptual innovations and methodological developments.

Regional Coverage. Because more than 80% of Armenian cultural heritage lies outside the Republic of Armenia, we include publications documenting inscriptions from Armenia, Artsakh, Georgia, Turkey, Iran, India, Ukraine, Moldova, and other historically Armenian regions.

Temporal Coverage. The corpus spans publications from the 1960s to 2024, enabling us to incorporate both foundational terminology and more recent shifts in scholarly interpretation.

2.3 Digitization

After assembling the corpus, all publications were digitized and converted into machine-readable text using an OCR pipeline for printed Armenian. Although most content was preserved, the OCR output introduced structural fragmentation, punctuation errors, and inconsistent line breaks, necessitating substantial preprocessing before terminology extraction and NER annotation.[7]

2.4 Preprocessing

We, thus, build a dedicated preprocessing pipeline to restore the structural and linguistic coherence of the corpus. The raw OCR output was highly fragmented: paragraphs were split into short line segments, Armenian punctuation, especially the full stop (:), was frequently misrecognized, and the hierarchical organization of the printed editions was largely lost. These issues made sentence segmentation and downstream terminology extraction unreliable without substantial normalization.

To address this, we designed a three-stage preprocessing workflow. (1) We segmented the OCR output by page in order to preserve the structural

[7] PDFs were converted to page-level images using `pdf2image` and processed with the Calfa `hye-tesseract` OCR engine (`https://github.com/calfa-co/hye-tesseract`), which supports Mesropian and Grabar orthographies.

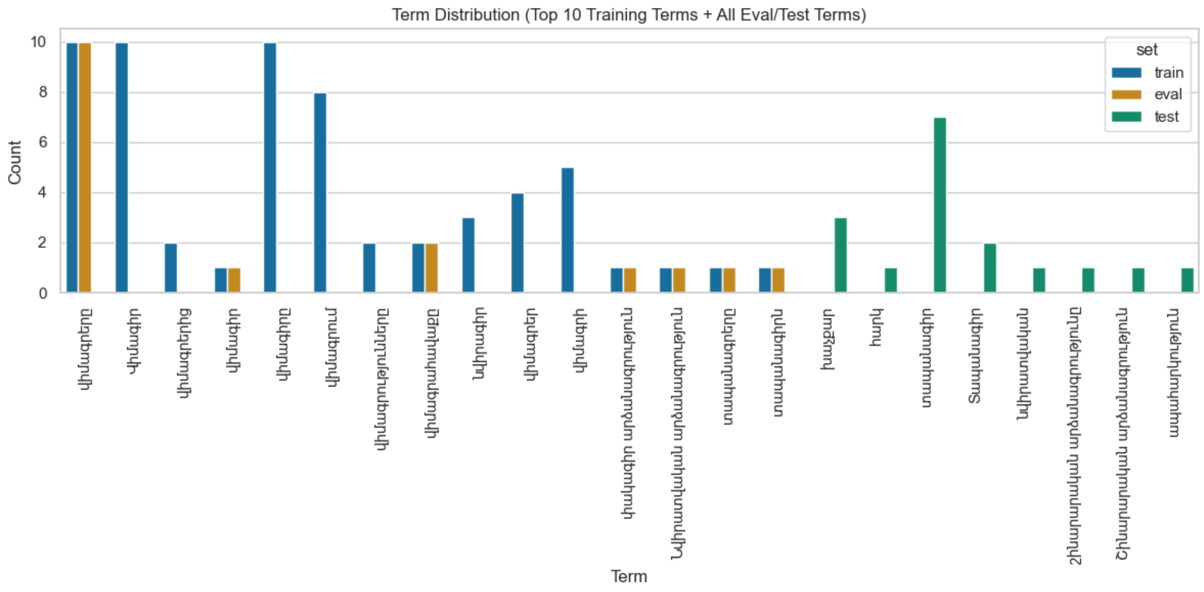

Figure 2: Distribution of terms across splits for DHV-10 data.

boundaries of the original publications. (2) Within each page, we regrouped the fragmented lines into coherent paragraphs by leveraging recurrent layout patterns characteristic of Armenian epigraphic editions that we noticed in DHV: inscription headers, line-count descriptors, uppercase inscription blocks, and commentary sections commonly introduced by the marker ծանոթ. (“note”). This stage also included systematic correction of punctuation and the merging of artificially split sentences. (3) We segmented the reconstructed paragraphs into sentences using a custom Armenian sentence splitter, which is adapted to the Armenian full-stop “:”.

3 Experiments

With the sentence-level corpus prepared, the next step is to identify and annotate occurrences of inscription-type terminology using our expert-defined inscription type reference list. These annotations will form the supervised dataset required for training and evaluating the NER models.

Because Armenian inscription terminology is morphologically rich and appears in highly variable scholarly contexts, simple string matching is insufficient. We therefore combine rule-based detection with GPT-5’s assistance[8]. The prompt guides term extraction by focusing on predefined terms, recording each occurrence separately with surrounding context for clarity. Metadata like file name and page number ensure traceability. We organized the extracted data hierarchically, with occurrences grouped under terms and terms compiled into an extraction result. The prompt emphasizes Armenian language handling to account for inflectional and orthographic variations, without relying on explicit morphological rules, ensuring robust zero-shot extraction.

Each sentence is then tokenized and labelled using the standard inside–outside–beginning (IOB) tagging scheme[9], which marks the boundaries of multiword terms. Because Armenian exhibits rich inflection, flexible word order, and modifier-heavy expressions, the automatic tagging often produced misaligned spans. Therefore, we did a manual verification and targeted rule-based adjustments to ensure consistent token-level annotation and transformed it in the IOB format. These rules are especially adapted to Armenian, where rich inflectional morphology and the frequent use of suffixes can cause surface forms to deviate from their base forms.

We then partitioned the data into training, development, and test sets. To evaluate true generalization beyond the expert-defined inscription type reference list, the test set contains only terms absent from the training and development sets. This design prevents memorization effects and allows us to measure the model’s ability to recognize unseen

[8] Specifically, we applied GPT-5 over predefined text chunks together with the reference term list (Appendix A). GPT-5 is used for assisted annotation, while extraction is performed with reproducible fine-tuned NER models.

[9] https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging)

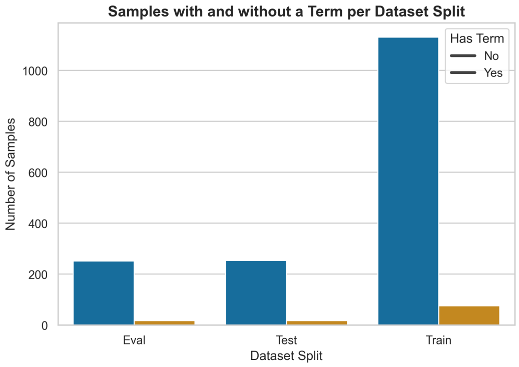

Figure 3: Distribution of sentences across dataset splits for DHV-10 data.

terminology, while the training and development splits include overlapping term types to support stable learning.

Figures 2 and 3 summarize the distribution of sentences and terms across the splits, showing a consistent 85/15 ratio of sentences with and without annotated entities and confirming the strict separation of term types between the test set and the remaining subsets. The figures also highlight the strong class imbalance inherent to the corpus, a characteristic of epigraphic scholarship that makes the task particularly challenging for NER models, which must learn to recognize sparse and morphologically variable terminology in predominantly non-annotated text.

3.1 Results

As our main baseline system, we chose **spaCy**, a neural NLP toolkit optimized for industrial use. We extend its pipeline with a custom NER component configured to recognize inscription-type entities and train it on the DHV–10 split using stochastic gradient descent with shuffled batches, periodic evaluation, and early stopping based on validation loss. For comparison, we also train three transformer-based models under identical conditions: Multilingual BERT (mBERT)[10], XLM-RoBERTa[11], and HyeBERT[12], an Armenian-specific transformer.

As shown in Table 3, the spaCy baseline achieves moderate precision (0.50) but extremely low recall (0.06), indicating that while its few predictions are often correct, it fails to identify the vast majority of inscription-related entities. HyeBERT, despite being tailored to Armenian, performs similarly poorly, suggesting that limited pretraining data and a smaller architecture restrict its ability to model the highly variable descriptive language of epigraphic scholarship.

More unexpectedly, HyeBERT, despite being specifically pretrained for Armenian, exhibits similarly poor performance, with an F1 score of 0.09 and recall of only 0.07. This result suggests that language-specific pretraining alone is insufficient for this task, likely due to limited pretraining data (Armenian subset from OSCAR dataset[13]), and possibly a mismatch between HyeBERT's training domain and the highly variable, descriptive language of epigraphic scholarship.

By contrast, the multilingual transformer models (mBERT and XLM-RoBERTa) substantially outperform both spaCy and HyeBERT, reaching F1 scores of 0.52 with recall values more than six times higher than spaCy. We attribute this performance gap to the richer contextual representations learned by multilingual transformers, which better capture Armenian's inflectional complexity and flexible word order. In addition, cross-lingual subword sharing appears beneficial for sparse, domain-specific terminology, allowing multilingual models to generalize more effectively across diverse epigraphic contexts despite limited task-specific training data.

3.2 Scaling to the Complete Corpus

The baseline experiments on DHV–10 demonstrate that multilingual transformer models offer a clear advantage over lightweight architectures, but they also reveal an important limitation: the restricted size and lexical diversity of the initial dataset prevent the models from fully learning the range of inscription-type terminology present in Armenian epigraphic scholarship. To develop a model capable of generalizing beyond a narrow subset of terms and adapting to highly variable scholarly contexts, we decided to train on a substantially broader corpus.

To evaluate model performance under more realistic and diverse conditions, we expand the dataset to include sentence-level annotations from twenty volumes of the corpus (Table 1). This extension increases the scale of the experiment by more than an order of magnitude: the baseline setup

[10] https://huggingface.co/google-bert/bert-base-multilingual-cased

[11] https://huggingface.co/FacebookAI/xlm-roberta-base

[12] https://huggingface.co/aking11/hyebert

[13] https://oscar-project.org/

Split	# Sentences	# Sentences w/ Term	# Unique Terms	# Overlap w/ Train
Train	1,208	76	27	-
Development	269	17	7	7
Test	271	17	8	0

Table 2: Dataset statistics for the NER training, development, and test splits, drawn from **DHV–10**. The test set contains only unseen terms.

Model Name	Precision	Recall	F1
spaCy	0.50	0.06	0.11
mBERT	**0.78**	0.39	**0.52**
XLM-RoBERTa	**0.78**	0.39	**0.52**
HyeBERT	0.34	0.07	0.09

Table 3: Baseline performance of spaCy and transformer models on the DHV–10 subset.

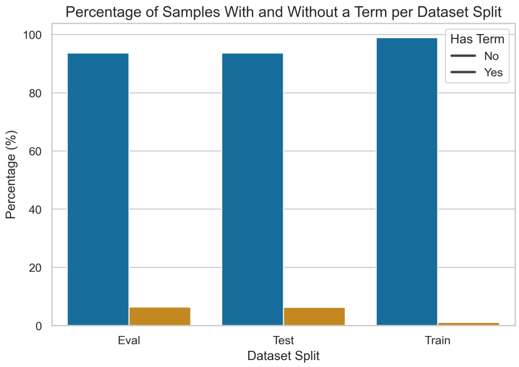

Figure 4: Distribution of sentences across dataset splits for complete data in percentages.

contained roughly 1,200 sentences, while the full dataset exceeds 56,000. The enlarged corpus introduces both a wider range of positive examples (sentences containing inscription terminology) and a substantially larger proportion of negative sentences, thereby mirroring the natural sparsity of terminology in scholarly prose. Crucially, the development and test sets remain unchanged, ensuring that all results are directly comparable across experimental settings.

Figure 4 shows the distribution of sentences across the training, development, and test splits for the complete dataset. Table 4 summarizes the key statistics of the expanded corpus.

We train the same set of models (spaCy, HyeBERT, mBERT, and XLM-RoBERTa) on this enlarged dataset. The results are presented in Table 5. We notice that, while maintaining the same trend, using a larger amount of data leads to notable improvements in transformer performance. The multilingual BERT model achieves the highest F1 score (0.63), with precision rising to 0.82 and recall improving to 0.50. XLM-RoBERTa also benefits from the expanded data, though to a lesser degree. These gains suggest that access to a richer and more heterogeneous set of contexts significantly enhances the models' ability to recognize inscription terminology, especially given that such terms occur in only about one percent of sentences in the corpus. The superior performance of mBERT may reflect the alignment between its subword vocabulary and the morphological patterns of Armenian, allowing it to identify relevant terms even under highly imbalanced conditions.

Overall, these results confirm that multilingual transformer models not only outperform smaller architectures but also scale effectively with larger and more realistic training data. This motivates their use as the primary models for terminology discovery in the unseen-books experiments that follow.

4 Generalization to Unseen Books and Term Discovery

To assess generalization beyond the training corpus, we applied the best-performing NER model to a set of books excluded from all training, development, and benchmark annotation stages. The volumes were processed using the same preprocessing and sentence-segmentation pipeline, and the model was used to infer inscription-type terminology from previously unseen texts.

Figure 5 illustrates the relationship between three term sets: the *expert reference vocabulary* (54 terms), the *benchmark terms* extracted during annotation (136 terms), and the *inferred terms* predicted by the model (98 terms).

The overlaps reveal two main behaviors. First, the model reliably recognizes known inscription-type terminology in new contexts: 40 inferred terms overlap with the benchmark set, indicating robust contextual generalization across grammatical realizations and textual environments. Second, the model shows limited capacity for independent

Split	# Sentences	# Sentences w/ Term	# Unique Terms	# Overlap w/ Train
Train	51,223	530	136	-
Development	269	17	7	7
Test	271	17	8	0

Table 4: Statistics of the complete dataset.

Model Name	Precision	Recall	F1
spaCy	0.50	0.06	0.11
mBERT	**0.82**	**0.50**	**0.63**
XLM-RoBERTa	0.67	0.44	0.53
HyeBERT	0.50	0.06	0.10

Table 5: Performance of all models on the complete dataset.

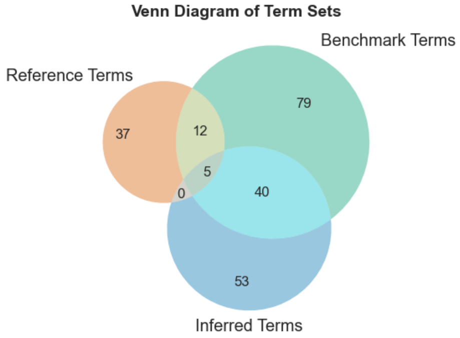

Figure 5: Venn diagram comparing reference terminology (54 terms), benchmark new terms (136 terms), and terms inferred by the model on unseen books (98 terms).

term discovery. None of the inferred terms overlap exclusively with the expert reference list, and only five terms are shared across all three sets. Furthermore, 53 inferred terms are unique to the model output but correspond to morphological variants, partial matches, or contextually ambiguous expressions rather than genuinely new inscription types.

Overall, these results indicate that while transformer-based NER models generalize well to unseen texts, they do not autonomously expand the conceptual terminology. Their primary value lies in supporting corpus exploration and expert analysis by surfacing candidate expressions and contextual variation rather than replacing expert-driven terminology formation.

5 Conclusions and Future Work

This study presented a first systematic, data-driven analysis of inscription-type terminology in Armenian epigraphic scholarship and established the foundational resources needed for developing a culturally specific SKOS vocabulary. Through the digitization of authoritative publications, the creation of a curated terminology list, and the training of transformer-based NER models, we produced the first empirical infrastructure for identifying and examining Armenian inscription-type concepts at scale.

Our results show that automatic terminology extraction alone is insufficient for this domain. While multilingual transformer models successfully generalize to unseen texts and reliably detect attested terms, they rarely identify genuinely new concepts; most model-generated candidates are morphological variants or contextually ambiguous expressions. This highlights the necessity of expert involvement, as Armenian inscription-type categorization is deeply embedded in cultural, architectural, and functional knowledge that cannot be inferred from surface patterns alone. At the same time, the corpus reveals both categories familiar from international ontologies (such as funerary, donor, or legal inscriptions) and types highly specific to Armenian tradition (such as գավիթային, բարավորի, or ջրօգտագործման արձանագրություն), many of which are absent from EAGLE and FAIR Epigraphy vocabularies.

Eventually, we would want to continue to focus on moving from terminology extraction toward concept modelling. This involves validating terminology through contextual and distributional evidence; identifying synonymy, variation, and fine-grained distinctions; constructing hierarchical relations and selecting preferred or alternative labels; detecting inscription-type concepts expressed only implicitly through descriptive context; and aligning Armenian-specific categories with international ontologies while preserving cultural specificity. These steps will advance the project from surface-level term identification toward a structured conceptual model grounded in Armenian scholarly practice and will ultimately support the development of the first Armenian SKOS vocabulary for inscription types, enhancing the interoperability and visibility of Armenian epigraphic heritage within global digital infrastructures.

References

S. Avagyan and H. Janpoladyan. 1977a. *Corpus Inscriptionum Armeniacarum*, volume 6.

Suren Avagyan and Hovhannes Janpoladyan. 1977b. *Divan Hay Vimagrutyan, Prak VI*. Academy of Sciences of the Armenian SSR, Yerevan.

Ashot Ayvazyan. 2011. *Dadivanki Verakangnumy (1997–2011 tt.)*. HChU, Yerevan.

M. Barkhudaryan. 1999. *Aghvants Yerkir yev Dratsik: Artsakh (The Land of Aghvank and Its Neighbors: Artsakh)*. Gandzasar, Yerevan.

S. Barkhudaryan. 1967a. *Corpus Inscriptionum Armeniacarum*, volume 3.

S. Barkhudaryan. 1973a. *Corpus Inscriptionum Armeniacarum*, volume 4.

S. Barkhudaryan. 1982a. *Corpus Inscriptionum Armeniacarum*, volume 5.

S. G. Barkhudaryan, K. Ghafadaryan, and S. T. Saghumyan. 2012a. *Corpus Inscriptionum Armeniacarum*, volume 9.

S. G. Barkhudaryan, K. Ghafadaryan, and S. T. Saghumyan. 2012b. *Divan Hay Vimagrutyan, Prak IX*. Institute of Archaeology and Ethnography, NAS RA, Yerevan.

Sedrak Barkhudaryan. 1967b. *Divan Hay Vimagrutyan, Prak III*. Academy of Sciences of the Armenian SSR, Yerevan.

Sedrak Barkhudaryan. 1973b. *Divan Hay Vimagrutyan, Prak IV*. Academy of Sciences of the Armenian SSR, Yerevan.

Sedrak Barkhudaryan. 1982b. *Divan Hay Vimagrutyan, Prak V*. Academy of Sciences of the Armenian SSR, Yerevan.

Francesco Bianchini. 2023. Looking beyond the text: Opportunities and challenges in the digitisation of sanskrit inscriptions. *Can't Touch This*.

David Espinosa Espinosa and Isabel Velázquez Soriano. 2021. Epigraphy in the digital age: opportunities and challenges in the recording, analysis and dissemination of inscriptions.

Artak Ghulyan. 2001. *Artsakhi yev Syuniki Melikakan Aparanknere*. HChU, Yerevan.

Tim Greenwood. 2004. A corpus of early medieval armenian inscriptions. *Dumbarton Oaks Papers*, 58:27–?

Garnik M. Grigoryan. 1996. *Divan Hay Vimagrutyan, Prak VII*. National Academy of Sciences of Armenia, Yerevan.

Garnik M. Grigoryan. 1999. *Divan Hay Vimagrutyan, Prak VIII*. National Academy of Sciences of Armenia, Yerevan.

Arsen Harutyunyan. 2017. Armaviri marzi vimagrakan zharangutyan usumnasirutyune. In *Metsamoryan Yntertsumner I*. Patmamshakutayin Argelots-Thangaranneri Tsarayutyun, Yerevan.

Arsen Harutyunyan. 2019a. Antip vimagrer yerevani katoghike yekeghetsuts. In G. G. Sargsyan and A. E. Harutyunyan, editors, *Sedrak Barkhudaryan – 120. Gitakan Hodvatsneri Zhoghovatsu*. HAI, Yerevan.

Arsen Harutyunyan. 2019b. Sisiani n. adontsi anvan patmutyan tangarani ardzanagir ararkanere. *Hnagitutyun: Vem*, 11(2(66)).

Arsen Harutyunyan. 2022. Tarerayin aghetnern u yerknaayin yerevuytnere vimagrerum. mas arayin: Yerkracharzh. *Hnagitutyun: Vem*, 14(1(77)).

Arsen Harutyunyan. 2023a. Tarerayin aghetnern u yerknaayin yerevuytnere vimagrerum. mas yerrord: Yerash, morekh. *Hnagitutyun: Vem*, 15(1(81)).

Arsen Harutyunyan. 2023b. Tatevi mets anapati vimagrakan zharangutyune.

Arsen Harutyunyan and Armine Melkonyan. 2021. S. ejmiatsni tapanagreri: Ghevond vrd. pirghalemiani norahayt zhoghovatsu. *Ejmiatsin*.

S. Karapetyan. 2021a. The armenian lapidary inscriptions of boon aghvank.

Samvel Karapetyan. 1998. *Vratsi Petakan Kaghakakanutyune yev Hay Mshakuyti Hushardzannere (1988–1998), Book II*. Gitutyun, NAS RA, Yerevan.

Samvel Karapetyan. 2004. *Kovkasyan Tangarani Haykakan Havaqatsun*. HChU, Yerevan.

Samvel Karapetyan. 2014. *Akunk yev Katnaghbyur Gyugheri Patmakan Hushardzannere*. HChU, Yerevan.

Samvel Karapetyan. 2021b. Bun aghvanki hay vimagrutyunnere.

Samvel Karapetyan. 2021c. Hyusisayin artsakh. https://raa-am.org/northern-artsakh/.

Samvel Karapetyan. 2024. *Khojivank*. HChU, Yerevan.

Raffi Kortoshyan. 2013a. *Halepi Vimagrere, Prak 16*. RAA Publishing, Yerevan.

Raffi Kortoshyan. 2013b. *The Inscriptions of Aleppo*, volume 16. RAA Publishing, Yerevan.

Raffi Kortoshyan. 2024. *Hndkastani Hayahots Bnakavayreri Vimagrere*. HChU, Yerevan.

Fangchao Liu, John Hindmarch, and Mona Hess. 2023. A review of the cultural heritage linked open data ontologies and models.

Husik Melkonyan. 2024. *Tsaghats Kar Vanke (Patmahnaagitakan Usumnasirutyun)*. HAI Publishing, Yerevan.

H. A. Orbeli. 1965a. *Corpus Inscriptionum Armeniacarum*, volume 1.

Hovsep A. Orbeli. 1960. *Divan Hay Vimagrutyan, Prak II*. Academy of Sciences of the Armenian SSR, Yerevan.

Hovsep A. Orbeli. 1965b. *Divan Hay Vimagrutyan, Prak I*. Academy of Sciences of the Armenian SSR, Yerevan.

Hamlet Petrosyan. 2008. *Khachkar*. Printinfo, Yerevan.

G. G. Sargsyan, A. E. Harutyunyan, and K. T. Asatryan. 2017a. *Corpus Inscriptionum Armeniacarum*, volume 10.

G. G. Sargsyan, Arsen E. Harutyunyan, and Karen T. Asatryan. 2017b. *Divan Hay Vimagrutyan, Prak X*. Institute of Archaeology and Ethnography, NAS RA, Yerevan.

Hamest Tamrazyan, Gagik Hovhannisyan, and Arman Harutyunyan. 2026. From stone to standards: A digital heritage interoperability model for armenian epigraphy within the leiden and epidoc frameworks. *Heritage*, 9(1):27.

Hamest Tamrazyan and Gayane Hovhannisyan. 2024a. Digital guardianship: Innovative strategies in preserving armenian's epigraphic legacy. *Heritage*, 7(5):2296–2312.

Hamest Tamrazyan and Gayane Hovhannisyan. 2024b. Preserving endangered heritage: Integrating geonames/pleiades, armenian toponyms and regularization for cultural identity preservation in conflict zones. *International Journal of Humanities and Arts Computing*, 18(2):224–248.

Hamest Tamrazyan and Gayane Hovhannisyan. 2025. Cultural categorization in epigraphic heritage digitization. *Heritage*, 8(5):148.

Lori Tashjian. 2018. *Tseghaspanutyune Verapratsneri Tsndavayrere Libanani Tapanagrerum*. National Academy of Sciences of Armenia Press, Yerevan.

Armine Tigranyan. 2023. The armenian cultural heritage of artsakh. *Mechanisms for Protection In The International System For Preservation of Heritage, Vem Series*, (6).

Yingwei Yan, Kenneth Dean, Chen-Chieh Feng, Guan Thye Hue, Khee-heong Koh, Lily Kong, Chang Woei Ong, Arthur Tay, Yi-chen Wang, and Yiran Xue. 2020. Chinese temple networks in southeast asia: a webgis digital humanities platform for the collaborative study of the chinese diaspora in southeast asia. *Religions*, 11(7):334.

A Prompting

```
zero_shot_prompt = f"""
You are an expert in Armenian epigraphy.
Your task is to extract specific terms
from the provided text chunks.
For each term, you will identify its
occurrences in the text,
along with the context in which
it appears. The context should include
a few words before and after the term
to provide clarity on its usage.

Each occurrence should be documented
with the following details:
- file_name: The name of the file where
the term was found.
- page_number: The page number in the
document where the term was found.
- term: The term itself.
- context: A snippet of text
surrounding the term.

Each term can have multiple occurrences,
and each occurrence should be
recorded separately and stored in an
Occurrence object. All occurrences of
a term should be grouped under a Term
object, which includes the term, its
definition, and a list of its occurrences.

Finally, all Term objects should be
compiled into an ExtractionResult object.

The text is written in Armenian and may
require special handling. Extract all
occurrences of the following terms from
the document text.
Terms and definitions: {terms}

--- BEGIN TEXT CHUNK ---
{chunk}
--- END TEXT CHUNK ---
"""
```