

The ArchiveGene Corpus: A Synthetic Multi-Layer Benchmark for Genealogical Information Extraction from Uzbek Historical Archival Texts

Adilbek Dauletov ^{1,*} , Noila Matyakubova ¹, Sevara Allabergenova ², Nargisa Ashirmatova ², Miyassar Tillayeva ³, Sevara Yoqubova ⁴ and Ikrom Islomov ⁵

¹ Department of Digital Technologies/Language Teaching Center, Alfraganus University, Tashkent 100190, Uzbekistan; n.matyakubova@afu.uz

² Department of Natural Sciences, National Research Nuclear University Tashkent Branch of the National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Tashkent 100214, Uzbekistan; smallabergenova@mephi.ru (S.A.); nmashirmatova@mephi.ru (N.A.)

³ Department of Uzbek Linguistics, Urgench State University Named After Abu Rayhon Beruni, Urgench 220100, Uzbekistan; tillayevamiyassar@mailurdu.uz

⁴ Department of Uzbek Language and Literature, Uzbekistan National Pedagogical University Named After Nizami, Tashkent 100185, Uzbekistan; s.yoqubova@npuu.uz

⁵ Department of Philology, Renaissance University of Education, Tashkent 100071, Uzbekistan; i_islomov@renessans-edu.uz

* Correspondence: a.dauletov@afu.uz

Abstract

Automatic extraction of genealogical information from historical archival-genealogical documents in Uzbek is an understudied problem for low-resource languages. Multi-layer NLP benchmarks are not sufficient to automatically identify individuals, family relationships, dates, place names, and archival identifiers in such texts. Also, the same people are mentioned in various forms: full name, pronoun (18.8%), initial, surname-name order, indirect expression (9.4%), and title. Existing NER and relation extraction corpora are mainly focused on high-resource languages or general domain texts and do not sufficiently cover the FAMILY_ROLE signals, historical spelling variants, and fond-opis-delos identifiers specific to Uzbek archival-genealogical texts. Proposed resource: We present the ArchiveGene Corpus, a controlled, fully synthetic, and reproducible five-layer resource consisting of 1000 Uzbek archival-genealogical-style documents, divided into 700 training, 150 validation, and 150 test documents. The corpus contains 8366 named entities, 10,625 person mentions, 2000 coreference chains, and 1000 genealogical relation triples. The dataset was generated using a deterministic template-based pipeline and a lexicon of Uzbek names, and is fully reproducible. Inter-annotator agreement values were 0.847 for NER, 0.793 for coreference, and 0.821 for RE, according to Cohen's κ . Comparative results are presented with four baseline models (rule-based, BiLSTM-CRF, mBERT, and XLM-RoBERTa). The dataset is openly hosted on the Zenodo platform under the CC BY 4.0 license; concept DOI: 10.5281/zenodo.20670360, v1.1.1 version DOI: 10.5281/zenodo.21429998. Scientific significance: To the best of our knowledge, ArchiveGene is among the first openly released, controlled synthetic resources for Uzbek that integrates named-entity recognition, person-mention detection, coreference resolution, genealogical relation extraction, and final tuple generation within a single annotation framework. The baseline analysis provides three main conclusions: (1) on the clean synthetic test set, the transformer models already reach 100.00 Micro-F1 for NER and 100.00 Macro-F1 for coreference-aware relation extraction, so coreference aggregation adds little on synthetic data (+2.25 for mBERT and +0.04 for XLM-RoBERTa) but its contribution is expected to grow on real archival text;



Received: 17 June 2026

Revised: 22 July 2026

Accepted: 31 July 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

(2) the rule-based and heuristic baselines lag far behind (Macro-F1 50.28 and 70.73) and fail entirely on spouse_of, showing the limits of lexical rules; and (3) a zero-shot evaluation on a real-document pilot reduces NER Micro-F1 from 100.00 to 22.17, indicating that the synthetic corpus is trivially learnable and that real-archival validation is essential.

Dataset: Dataset concept DOI: 10.5281/zenodo.20670360 | Version v1.1.1 DOI: 10.5281/zenodo.21429998

Dataset License: CC BY 4.0

Keywords: Uzbek NLP; historical archival texts; genealogical information extraction; named entity recognition; coreference resolution; relation extraction; synthetic benchmark corpus; low-resource languages; inter-annotator agreement

1. Summary

Digital archives, historical documents and genealogical sources are important sources of information for reconstructing the biographies of individuals, establishing intergenerational relationships and preserving cultural heritage. Such documents contain personal names, dates, marriage and death records, places of residence, social status and kinship relations, but they are usually presented in free text. Therefore, the automatic identification, association and transformation of this information into a structured format is a complex NLP task.

Historical archival-genealogical texts differ significantly from general domain texts: they often contain historical spellings, formal styles of presentation, titles, kinship terms, document identifiers and context-dependent expressions. The agglutinative structure of the Uzbek language further complicates this process, since although units such as “otasi”, “onasining”, “farzandi”, “qizining” are important signals for genealogical relation extraction, which person they belong to is often determined only through coreference resolution.

The literature review shows that although existing NLP datasets cover NER, coreference resolution, and relation extraction tasks, most of them focus on English, general domain, or historical newspaper texts. Agglutinative morphology, FAMILY_ROLE signals, historical spelling variants, and fond-opis-delo identifiers specific to Uzbek archival-genealogical texts are not sufficiently covered in existing resources.

In this study, the ArchiveGene Corpus is presented as a multi-layer benchmark supporting the tasks of NER, person-mention detection, coreference resolution, genealogical relation extraction, and final tuple generation for Uzbek archival-genealogical texts. The final output of the corpus is given in the form of structural tuples in the format ⟨Person A, Relation, Person B, Date, Location, Source Document⟩. Its main novelty is that it covers domain-specific entities such as FAMILY_ROLE and DOCUMENT_ID along with common named entity labels and allows empirically assessing the impact of the coreference resolution stage on the quality of relation extraction.

1.1. Literature Review

Existing research on archival-genealogical NLP has been conducted in four areas: NER, coreference resolution, genealogical relation extraction, and multi-layer information extraction in historical texts. Table 1 compares the ArchiveGene Corpus with the most widely used datasets in this field and shows the methodological gap.

Table 1. Comparative analysis of archival-genealogical NLP datasets.

Dataset	Language	Domain	NER	Coref	RE	IAA	Open
CoNLL-2003	English	News	+	—	—	$F_1 = 0.93$	Yes
OntoNotes 5.0	English/Chinese	Multi-domain	+	+	—	$\kappa = 0.89$	Yes
ACE 2005	English	Multi-domain	+	+	+	$F_1 = 0.85$	Limited
CLEF HIPE 2020	Multi lingual	Historical newspaper	+	—	—	$\kappa = 0.78$	Yes
FamHist (Yang 2020)	English	Clinic	+	—	+	$\kappa = 0.82$	Limited
WikiCoref	English	Wikipedia	—	+	—	$\kappa = 0.85$	Yes
ArchiveGene	Uzbek	Historical archive	+	+	+	$\kappa = 0.847$	Yes

As can be seen from Table 1, most of the existing datasets focus on English language, general domain or historical newspaper texts. None of them combines NER, coreference resolution and genealogical relation extraction tasks for Uzbek archival-genealogical texts in a single multi-layer annotation scheme. In addition, the FAMILY_ROLE label, which is an important semantic signal for relation extraction in the Uzbek archival-genealogical domain, is not specifically indicated in the existing schemes.

1.2. NER in Historical and Archival Texts

Named-entity recognition in historical and archival texts is more challenging than in general domain texts due to OCR/HTR errors, archaic spellings, language evolution, domain differences, and lack of annotation. Blouin et al. show that off-the-shelf NER systems perform poorly on historical texts, while small amounts of annotation and transfer learning improve quality [1].

The lack of annotation has also been addressed by LLMs and domain-specific transformers. Zhang and Colavizza show that LLMs can be a practical solution for low-resource historical corpora with zero-shot and few-shot prompting without additional training [2]. Schweter et al. demonstrate that a pre-trained model specifically for historical texts using the hmBERT model is more effective than a general language model [3].

The association of entities with knowledge bases is also an important issue. Boros et al. studied NER and entity linking together in multilingual historical documents, showing OCR errors and the difficulty of linking names to DBpedia/Wikidata in multilingual situations [4]. Ehrmann et al. presented HIPE 2020, which is an important benchmark for historical newspapers that covers NERC and entity linking, but it does not cover genealogical relationships, coreference, and tuple generation [5]. Similarly, studies on Uzbek digital library and archive systems also show that NER, semantic enrichment, OCR, noise removal, and data standardization steps are important in the intellectual processing of archival documents [6,7]. Tual et al. studied the problem of nested NER in historical structured documents, showing the limitations of simple NER models in separating internal entities such as address, street, number, person, and profession [8].

Overall, these works show that historical NER requires a special approach due to OCR/HTR noise, archaic spellings, domain differences, multilingualism and structural complexity, but most of them are limited to recognition, classification or linking. In Uzbek archival-genealogical texts, NER is not the ultimate goal—it acts as a base layer for mention detection, coreference resolution, relation extraction and tuple generation.

1.3. Coreference Resolution and Linking Person Mentions

Coreference resolution is central to genealogical relation extraction, because relations are determined between resolved person entities, not raw-mention spans. A person is

repeatedly mentioned by full name, surname-first name order, initial, title, pronoun, kinship role or indirect expression; if these mentions are incorrectly linked, the wrong subject–object pairs are selected at the relation extraction stage.

Cruz, Rocha, and Cardoso summarize the coreference field from the perspective of end-to-end and cross-lingual systems, citing pipeline errors and lack of annotation in low-resource languages as major limitations [9]. Joshi et al. show that modeling entire spans rather than tokens in SpanBERT improves coreference quality [10]. Xu and Choi argue that complex inference mechanisms are not always superior when a strong contextual encoder is available [11]. Multilingual and low-resource coreference studies are of direct relevance for Uzbek archival texts. Results from the CRAC shared task show that mention detection, identity clustering, and zero anaphora are still difficult problems [12–14]. Although recent work has strengthened LLM and instruction-tuning approaches, the issue of competition with task-specific coreference systems in historical and low-resource languages remains open [15,16]. This situation demonstrates the need to clearly link indirect mentions such as “uning o’g’li” (his son), “marhumning turmush o’rtog’i” (the deceased’s spouse), and “mazkur shaxs” (this person) in Uzbek archival texts.

In this context, coreference resolution is not an auxiliary but a key semantic link between NER and RE. The accurate identification of person–place or person–person relations is directly related to the correct identification of mentions and their association at the entity-level [17]. Therefore, coreference resolution for Uzbek archival-genealogical texts is a necessary component for reducing mispairs and accurately reconstructing lineage relationships. Existing work on genealogical relation extraction mainly focuses on obituaries, family books, biographical or English-language texts. He et al. identified persons, attributes and family relations from online obituaries using multitask neural extraction to build a genealogical knowledge graph [18]. Parrolivelli and Stanchev propose a relation classification pipeline based on NER, coreference and LUKE [19]. Plum et al. created a semi-supervised dataset for biographical relations based on Wikipedia, Pantheon and Wikidata [20]. Boylan et al. demonstrated the possibility of flexible extraction of previously unseen relation labels with the GLiREL zero-shot model [21]. However, these works do not cover OCR/HTR noise, agglutinative morphology, FAMILY_ROLE signals, and coreference-aware genealogical RE in Uzbek archival documents in a single annotation scheme.

1.4. NER and Relation Extraction Together

In recent years, NER and relation extraction have not been considered separate steps, but rather joint information extraction that extracts entity–relation triplets together. Nasar et al. systematize rule-based, machine learning, deep learning, and joint modeling approaches, and show that hybrid and joint deep learning models have become the leading direction [22].

Joint extraction models such as CasRel, TPLinker, PRGC, OneRel, UIE, and single-table filling serve to jointly extract entity–relation triplets, reduce overlapping relations, and generalize text-to-structure generation [23–28]. However, these models have been evaluated mainly on common domain benchmarks in English or Chinese, and do not sufficiently cover the pronoun mentions, FAMILY_ROLE signals, fond–opis–delo identifiers, OCR/HTR noise, and agglutinative morphology specific to Uzbek archival-genealogical texts. Therefore, the ArchiveGene Corpus is proposed as a domain-specific synthetic multi-layer benchmark that evaluates NER, mention detection, coreference resolution, and genealogical relation extraction within the framework of Uzbek archival-genealogical texts.

1.5. Synthetic Corpus Creation and Reproducible Benchmarks

Annotated real-world data are often insufficient for low-resource languages and special domains. In this case, synthetic corpus creation is an effective approach to test the annotation scheme, generate initial benchmark results, and provide controlled evaluation [29].

However, synthetic corpora cannot fully reflect all the linguistic, historical, and technical complexities of real texts—in particular, OCR/HTR errors, archaic spelling, dialectal forms, damaged text, and contextual uncertainties are only partially modeled in template-based generation. Therefore, synthetic corpuses should not be considered as a complete replacement for real datasets, but rather as an initial reproducible benchmark, a tool for testing the annotation scheme, and evaluating the model under controlled conditions.

ArchiveGene Corpus is based on this approach: it is created using deterministic templates and a lexicon of Uzbek names, kinship terms, historical place names, and archival identifiers, and automatically and consistently generates NER, mention detection, coreference resolution, and genealogical relation extraction layers. At the same time, the synthetic-to-real difference is recognized as a specific limitation, and an extended version enriched with real archival documents is planned in the future.

2. Data Description

This section describes the general structure, annotation layers, label schemes, and final structured output format of the ArchiveGene Corpus. The corpus is based on the requirement to extract personal information from historical archival-genealogical texts in a step-by-step manner. In such texts, genealogical relationships are expressed not only between two names, but also through person mentions, pronouns, kinship roles, dates, places, and document identifiers. Therefore, the ArchiveGene Corpus was developed as a multi-layered synthetic benchmark that encompasses not only NER, but also person-mention detection, coreference resolution, genealogical relation extraction, and final tuple generation. Table 2 summarizes the annotation layers, corpus statistics, and baseline results of the ArchiveGene Corpus.

Table 2. ArchiveGene Corpus: annotation layers, statistics, and baseline results.

Indicator	Value
Tasks	NER · Mention Detection · Coreference Resolution · Genealogical RE · Tuple Generation
Language/Period	Uzbek language (agglutinative, low resource) · 1860–1970
Domain	Historical archive texts · Genealogical-biographical documents
Corpus Size	1000 documents Train 700/Val 150/Test 150
Named entities	8366 (7 label: PERSON, FAMILY_ROLE, DATE, LOCATION, TITLE, EVENT, DOCUMENT_ID)
Person mentions	10,625 (8 types, including PRONOUN 18.8% and INDIRECT_REFERENCE 9.4%)
Coreference chains	2000 (average 5.31 mentions/chain)
Relation triples	1000 (7 types of genealogical relations)
Average token/document	182.9 tokens Vocabulary: 1320 unique tokens
IAA (Cohen's κ)	NER: $\kappa = 0.847$ Coreference: $\kappa = 0.793$ RE: $\kappa = 0.821$
Output format	(Person A, Relation, Person B, Date, Location, Source Document)
NER Micro-F1 (baseline)	Rule-based: 77.30% mBERT: 100.00% XLM-RoBERTa: 100.00%
Coref CoNLL F1 (baseline)	Exact matching: 14.27% mBERT: 100.00% XLM-RoBERTa: 100.00%
RE Macro-F1 (baseline)	Rule-based: 50.28% mBERT + coref: 100.00% XLM-RoBERTa + coref: 100.00%

The corpus design is based on three main principles. First, each document has five layers of related annotation, starting with the identification of named entities and ending with the formation of a final genealogical tuple. Second, the annotation scheme is adapted to the linguistic and documentary features of the Uzbek archival domain: domain-specific labels such as FAMILY_ROLE and DOCUMENT_ID are introduced along with generic labels such as PERSON, DATE and LOCATION. Third, the output of the corpus is provided in a structured format that can be fed into genealogical knowledge graphs, semantic search or tree reconstruction algorithms. Such a structured output format is compatible with the functional architecture of electronic archival systems and metadata-driven semantic search approaches. Previous studies have shown that such systems provide fast document storage, indexing and context-aware retrieval through data storage, electronic document cataloging, full-text modules, text classification/search and semantic indexing based on Dublin Core, TF-IDF and BERT [30,31].

2.1. General Structure of the Corpus

The ArchiveGene Corpus consists of 1000 historical archival-genealogical texts in the Uzbek language, divided into train, validation, and test parts in the ratio of 700/150/150. The documents model the official-biographical and genealogical style of narration typical of the period 1860–1970. Each document was checked for the presence of at least one person, one genealogical signal, date or place element, and an identifier indicating the source. By design, each document contains exactly two coreference chains and one gold relation triple; these counts are fixed generation parameters, not empirical frequencies. The main statistical indicators of the corpus—including its size, entity and mention counts, and structural properties—are summarized in Table 3.

Table 3. Main statistical indicators of the ArchiveGene Corpus.

Metrik	Value
Dataset type	Synthetic, deterministic pipeline-based, fully reproducible
Language and period	Uzbek language; agglutinative structure; 1860–1970 style
Number of documents	1000 documents
Train/Val/Test	700/150/150
Named entities	8366 entities, 7 main label types
Person mentions	10,625 mentions, 8 mention types
Coreference chains	2000 chains; average 5.31 mentions/chain
Relation triples	1000 genealogical relation triples
Final tuples	1000 (Person A, Relation, Person B, Date, Location, Source) records
Average token/document	182.9 tokens
Vocabulary	1320 unique tokens
File format	JSON-L: BIO NER, mention spans, coreference clusters, relation triples, final tuples

The fact that the number of named entities is much higher than the number of relation triples indicates that the corpus needs to be processed on the basis of a multi-stage pipeline. The model first identifies the entities, then separates the entity mentions, joins them into coreference chains, and then defines the genealogical relationships. In separating the train, validation and test parts, a balance was maintained according to the types of label, relation and mention and the length of the document; the test part was kept separate from the model selection and parameter setting process.

2.2. Annotation Layers

The ArchiveGene Corpus consists of five interrelated annotation layers: Named entity → Person mention → Coreference → Genealogical relation → Final tuple. In this sequence, each layer forms the input for the next step.

- Named-entity annotation defines the main units;
- Person-mention annotation specifies all forms referring to a person;
- Coreference annotation unites mentions belonging to one person into a single cluster;
- Genealogical-relation annotation determines the kinship relations between resolved-person entities;
- Final-tuple annotation turns the extracted information into a structural genealogical record.

2.3. Named-Entity Annotation

The named-entity annotation layer identifies entities of genealogical and archival significance in the text. It uses the labels PERSON, FAMILY_ROLE, DATE, LOCATION, TITLE, EVENT, and DOCUMENT_ID (Table 4); tokens that do not belong to an entity are marked with the “O” marker in the BIO schema.

Table 4. Named-entity label types.

Label	Description	Example
PERSON	A person’s name, surname, nickname, or lineage form	Ali Karimov; Mahmud ibn Yusuf
FAMILY_ROLE	A kinship role or family status; a semantic signal for relation extraction	ota; onasi; o’g’li; qizi; turmush o’rtog’i
DATE	A date of birth, death, marriage, migration, or other biographical event	1910-yil; 12-mart 1924; 1936-yilda
LOCATION	A place of birth, residence, or event	Samarqand; Toshkent; Qo’qon uyezdi
TITLE	A title, position, professional, or social status	domla; mirza; qori; mudarris
EVENT	A genealogical or biographical event	tug’ilgan; vafot etgan; turmush qurgan
DOCUMENT_ID	A source identifier for a document, collection, record, case, or archive	ArchiveDoc_017; fond 12; opis 3; delo 45

The FAMILY_ROLE label is an important feature of the corpus, since units such as “otasi” (father), “onasining” (mother’s), “farzandi” (child) and “turmush o’rtog’i” (spouse) indicate genealogical connections between individuals. The boundary between PERSON and FAMILY_ROLE is consistently separated: if the kinship element is a component of the name of a person, it is included in the PERSON span; if it is an independent kinship signal, it is designated as FAMILY_ROLE. The DATE, LOCATION, TITLE, EVENT and DOCUMENT_ID labels serve to form the time, place, event and source attributes in the final tuple.

2.4. Person-Mention Annotation

Person mention annotation identifies all mentions in the text that directly or indirectly refer to a person. In historical archival-genealogical texts, a person can be mentioned again by full name, initial, surname-name order, title, kinship role, pronoun or indirect expression. Therefore, eight types of person mentions are distinguished in the corpus. Table 5 presents these mention types together with their short descriptions and examples.

Table 5. Mention types for person-mention annotation.

Mention Type	Short Description	Example
FULL_NAME	A person's full name or first and last name; usually an anchor mention	Ali Karimov
INITIAL_NAME	A name in the form of an initial or abbreviation	A. Karimov
REVERSED_NAME	A form in which the last name and first name are interchanged	Karimov Ali
TITLE_BASED	A mention with a title or status	domla Ali; qori Mahmud
FAMILY_ROLE_BASED	A mention given through a kinship role	uning otasi; Ali Karimovning qizi
PRONOUN	A mention given through a pronoun	u; uning; ular
INDIRECT_REFERENCE	A direct reference	marhum; mazkur shaxs
UNCERTAIN	A mention with an uncertain antecedent	uning otasi—antecedent noma'lum

FULL_NAME is used as the main anchor form of the coreference chain, while INITIAL_NAME and REVERSED_NAME cover variants of name spelling in archival documents. FAMILY_ROLE_BASED mentions provide an important signal for genealogical relation extraction, but they are only included in the cluster if their antecedent is clear from the context. PRONOUN mentions account for 18.8% of the corpus, and INDIRECT_REFERENCE mentions account for 9.4%; this indicates that simple approaches such as exact string matching are not sufficient. The UNCERTAIN category prevents ambiguous mentions from being forced into the cluster.

The percentage distribution of the eight mention types mentioned above across the corpus is summarized in Figure 1.

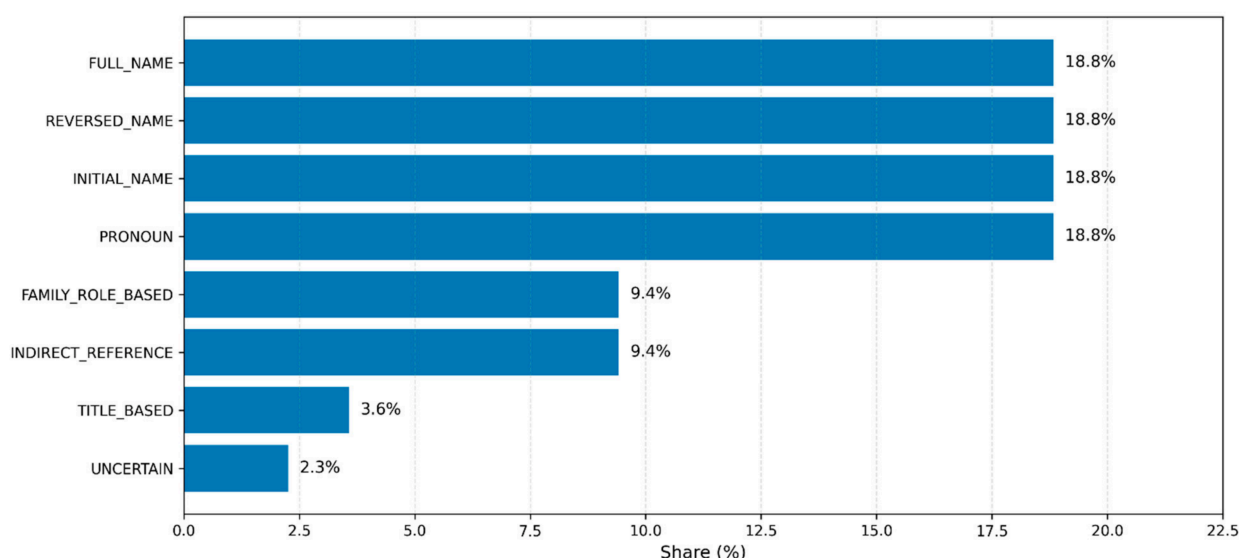


Figure 1. Template-controlled distribution of person-mention types in the ArchiveGene Corpus (v1.1.1; total 10,625 mentions). The equal proportions of FULL_NAME, REVERSED_NAME, INITIAL_NAME, and PRONOUN (18.8% each) are intentional generation parameters produced by the deterministic mention-variation assignment procedure; they are neither a plotting artifact nor empirical estimates of mention frequencies in authentic archival texts.

The distribution shows that pronoun (PRONOUN) and indirect (INDIRECT_REFERENCE) mentions occupy a significant share; this necessitates linking mentions based on context rather than relying only on surface-form matching, that is, through coreference resolution (Section 2.5).

2.5. Coreference Annotation

Coreference annotation combines different mentions of a single real person into a single cluster. Genealogical relations are not established between raw mentions, but between resolved-person entities specified by coreference. For example, if “Ali Karimov”, “A. Karimov”, “u” (he) and “mazkur shaxs” (this person) refer to the same person in context, they are included in one cluster; similar names, but other persons, are kept in separate clusters. Table 6 summarizes the coreference annotation rules applied in the corpus.

Table 6. Coreference annotation rules.

Rule	Content
Merge definite mentions	Mentions that refer to a single real person are included in a single cluster.
Define anchor mention	The full name or most specific mention is taken as the base form of the cluster.
Carefully link pronouns and indirect mentions	They are only added to a cluster if the antecedent is clear from the context.
Separate UNCERTAIN mentions	Mentions with an uncertain antecedent are not forced into a cluster.
False-merge and false-split control	Incorrect merging of different persons or excessive separation of one person are checked.

This layer is the main semantic connecting stage for relation extraction. If the mention “Uning” (His) in the sentence “Uning otasi Mahmud Karimov bo’lgan” (His father was Mahmud Karimov) is not linked to Ali Karimov, it becomes difficult to determine whether Mahmud Karimov is his father. Therefore, coreference annotation directly affects the correct selection of subject–object pairs and the quality of determining genealogical relations.

2.6. Genealogical-Relation Annotation

The genealogical-relation annotation defines the kinship relationships between resolved-person entities. The ArchiveGene Corpus uses seven positive genealogical relations and one negative class. *no_relation* is included to indicate cases where two persons appear in the same document, but no genealogical relationship is indicated between them. Table 7 lists the genealogical relation label types together with their categories and examples.

Table 7. Label types for genealogical0relation annotation.

Relation Label	Type	Example
parent_of	Directional positive relation	Mahmud Karimov parent_of Ali Karimov
child_of	Directional positive relation	Ali Karimov child_of Mahmud Karimov
spouse_of	Symmetrical positive relation	Ali Karimov spouse_of Zaynab Karimova
sibling_of	Symmetrical positive relation	Hasan Karimov sibling_of Malika Karimova
ancestor_of	Directional positive relation	Yusuf Karimov ancestor_of Ali Karimov
descendant_of	Directional positive relation	Ali Karimov descendant_of Yusuf Karimov
relative_of	The type is uncertain positive relation	Ali Karimov relative_of Hasan Karimov
no_relation	Negative class	Ali Karimov no_relation Mirza Sobirov

In directional relations, the subject–object order is strictly maintained, while in symmetric relations, the *spouse_of* and *sibling_of* entries are brought into the standard order and duplicate inverse relations are removed. *relative_of* is used for cases where the existence of a relationship is known, but its exact type is not specified. The relation annotation process checks for the presence of FAMILY_ROLE, EVENT, or contextual signals, reliance on coreference, merging of duplicate relations, and correct separation of *no_relation* cases.

2.7. Final Tuple Annotation

The final tuple annotation brings together the results of the named entity, person mention, coreference cluster, and genealogical relation identified in the previous layers into a single structural format. The final output of the dataset is given in the following format: (Person A, Relation, Person B, Date, Location, Source Document). For example, the tuple entry (Ali Karimov, parent_of, Malika Karimova, 1924, Samarkand, ArchiveDoc_017) indicates the parent_of relation between Ali Karimov and Malika Karimova, the time and place context of this relationship, and the source document. Table 8 describes the fields that constitute the final-tuple annotation.

Table 8. Final-tuple annotation fields.

Field	Description	Example
Person A	The resolved-person entity that acts as the subject of the relation	Ali Karimov
Relation	The type of genealogical relationship between Person A and Person B	parent_of
Person B	The resolved-person entity that acts as the object of the relation	Malika Karimova
Date	The date associated with the relation or genealogical event	1924
Location	The location associated with the relation or event	Samarkand
Source Document	The identifier of the document or archive from which the information was obtained	ArchiveDoc_017

Each tuple must have at least two resolved person entities and one relation label. The Date and Location fields are only populated if they are explicitly present in the text; otherwise, they are set to null. The Source Document field is always maintained, as indicating the source of genealogical information is important for subsequent verification and provenance control. Unclear or unsubstantiated relations are not included in the final tuple layer, and duplicate tuples are merged into a single standard record.

Thus, the corpus design described in Section 2 forms the methodological foundation of the ArchiveGene Corpus.

2.8. Data Records, Open Access, and File Format

The ArchiveGene Corpus (v1.1.1) is available as an open scholarly resource on the Zenodo platform under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The concept DOI of the dataset is 10.5281/zenodo.20670360, which always points to the latest version; the version-specific DOI for the current version v1.1.1 is 10.5281/zenodo.21429998. When using the dataset, reference to this article and the v1.1.1 DOI is required; the concept DOI is provided as a global reference to the most recent version of the dataset.

The corpus is stored in JSON-L format, with each document represented as a separate JSON object. This format allows for working with large NLP datasets, extracting layered annotations, and evaluating them independently for different tasks. The repository contains 700/150/150 train, validation, and test partitions, deterministic generation scripts, a label scheme, an annotation guide, a README file, and a requirements file for rebuilding the corpus. Complete baseline implementation details, including model heads, candidate construction, seed handling and evaluation scripts, are provided in Appendix B.

Each JSON-L record consists of the following main fields: doc_id is the unique identifier of the document, text is the full text of the document, tokens is the text separated into tokens, ner_annotations are the named entity annotations in BIO format, mention_annotations are the person mention type and span boundaries, coreference_clusters are the groups of mention IDs belonging to a person, relation_triples are the genealogical relationships in

the form of subject–relation–object, and final_tuples are the final structural records in the format ⟨Person A, Relation, Person B, Date, Location, Source⟩.

This multi-layered file structure allows the ArchiveGene Corpus to be used separately or integrated in various NLP tasks. Researchers can use BIO annotations for NER tasks, mention and cluster data for coreference resolution, relation_triples for genealogical relation extraction, and final_tuples for end-to-end genealogical information extraction. Therefore, the corpus is a suitable synthetic benchmark resource not only for evaluating individual models, but also for developing and comparing complete genealogical information extraction pipelines.

3. Methods

The ArchiveGene Corpus was created using a deterministic template-based pipeline to be fully reproducible. The deterministic approach is based on two factors: the quality of the annotation is ensured by full control over the source, with errors only due to the template design, not the generation process; and second, any researcher can recreate the exact same dataset with the same seed and lexicon. The pipeline consists of four stages: (1) template design, (2) lexicon preparation, (3) parametric document generation, and (4) automatic annotation and quality filtering. All scripts, templates, and lexicon files are publicly available in the Zenodo repository along with the dataset.

The four-stage deterministic process of corpus creation—template and lexicon loading, parametric document generation, automatic annotation, quality filtering, and the final 700/150/150 split—is formally summarized in Algorithm 1. Each candidate document is generated iteratively using random.seed(candidate_id); documents that do not meet the quality criteria are rejected and moved to the next candidate_id, and the process ends when valid_count reaches 1000. The pipeline stages are discussed in detail in the following subsections: template design (3.1), lexicon preparation (3.2), parametric document generation (3.3), corpus statistics and quality control (3.4–3.8). The formal pseudocode for the generation and annotation processes is presented in Algorithms 1 and 2.

Algorithm 1: Deterministic generation of documents

```

Input: templates T, lexicon L, target corpus size K
Output: annotated corpus D
D <- {}; candidate_id <- 1; valid_count <- 0
while valid_count < K do
  seed(candidate_id)
  (p1, p2) <- sample_names(L)
  rel <- sample_relation()
  meta <- sample_metadata(L)      // year, loc, src
  variants <- assign_mention_variants(p1, p2)
  text, spans <- build_text(T, variants, rel, meta)
  if quality_ok(text, spans) then // 3
    D <- D + annotate(text, spans, rel)
    valid_count <- valid_count + 1
  end if
  candidate_id <- candidate_id + 1
end while
D <- shuffle(D, fixed_seed)
return split(D, 700/150/150)

```

Algorithm 2: Greedy span alignment (mention/entity indexing)

Input: *hujjat matni s, joylanadigan spanlar M*
 Output: *span-based annotations A with start and end indices*
A <- {}; *used* <- {}
M <- sort(*M*, by length descending) // uzun avval
 for each span *m* in *M* do
 pos <- find(*s*, *m.text*, from = 0)
 while *pos* != −1 do
 (*start*, *end*) <- (*pos*, *pos* + len(*m.text*))
 if not overlaps((*start*, *end*), *used*) then
 A <- *A* + (*m*, *start*, *end*)
 used <- *used* + (*start*, *end*)
 break
 end if
 pos <- find(*s*, *m.text*, from = *pos* + 1)
 end while
 end for
 return *A*

3.1. Template Design

Archival-genealogical texts were divided into three categories and a separate template group was developed for each: (A) birth and death records—PERSON, DATE, LOCATION and FAMILY_ROLE are found in high density; (B) marriage and family status documents—there are direct lexical signals for the spouse_of and parent_of relations; (C) registration and administrative records—formal style, TITLE, DOCUMENT_ID and INDIRECT_REFERENCE are more common.

The templates were implemented in the Jinja2 template engine; its support for conditional expressions and loops allows you to create several document variants with different mention variations (FULL_NAME, PRONOUN, REVERSED_NAME, FAMILY_ROLE_BASED) from a single template. Template parameters and their entity types are listed in Table 9.

Table 9. Template parameters and entity types.

Template Parameter	Entity Type	Example
{{ subject_name }}	PERSON	Ali Karimov, Mahmud ibn Yusuf
{{ family_role }}	FAMILY_ROLE	otasi, onasi, o'g'li, qizi
{{ event_date }}	DATE	1910-yil, 12-mart 1924
{{ location }}	LOCATION	Samarqand, Toshkent, Qo'qon uyezdi
{{ doc_id }}	DOCUMENT_ID	fond 12, opis 3, delo 45
{{ title }}	TITLE	domla, mirza, qori
{{ pronoun }}	—	u, uning, ular

The templates were reviewed by three Uzbek history and archival experts according to two criteria: suitability for the archival style of the 1860s–1970s and the realistic representation of genealogical information; based on expert comments, 9 out of 14 options were included in the final set. The nine retained templates were evenly distributed across the three document groups: three templates for Group A (birth and death records), three templates for Group B (marriage and family-status records), and three templates for Group C (registration and administrative records). The template identifiers and group assignments are provided in the repository configuration file. For each template, the forms in which a

person is mentioned throughout the text are determined by a separate variation sequence (for example: 1-gap FULL_NAME → 3-gap PRONOUN → 5-gap TITLE_BASED → 7-gap INDIRECT). This sequence is directly transmitted to the mention annotation and coreference cluster generation, thus providing high-quality mention types without manual annotation during construction.

3.2. Lexicon Preparation

To fill in the template parameters, a special lexicon of 1096 entries was created, covering historical forms of names, places, and titles typical of archival documents, as presented in Table 10 below.

Table 10. Lexicon content and sources.

Lexicon Type	Quantity	Source
Uzbek male names	312	Historical archives and onomastic sources
Uzbek female names	287	Historical sources
Surnames	156	Archival documents of the Soviet period (-ov/-eva forms)
Place names	94	Historical administrative maps and archive catalogs
Titles and statuses	47	Historical and biographical dictionaries
Stock-office-case formats	200	Structure of the catalog of the National Archives of Uzbekistan

Elements were selected based on the principle of historical authenticity: names are in the forms recorded in the National Archives and onomastic sources at that time (modern, but less commonly used forms at that time were not included); surnames are forms with the suffix -ov/-ova typical of the Soviet era; place names are not modern, but official administrative names of that time (such as “Kokand uyezdi”); and fund–opis–delo identifiers are in the range of fond 1–150, opis 1–20, delo 1–500. All files are in the repository in CSV format, and future researchers can add elements appropriate to their period or region.

3.3. Parametric Document Generation

Each document was created through a deterministic process consisting of four consecutive stages. Reproducibility is ensured by the random.seed(candidate_id) mechanism: candidate_id increases sequentially starting from 1 for each attempt and continues until valid_count reaches 1000. In this process, a total of 1342 candidate documents were generated, of which 1000 that passed the quality criteria were included in the corpus; each candidate_id value returns exactly the same set of parameters. The general generation–annotation–splitting process is summarized in Algorithm 1; each of its steps is explained in detail below.

Stage 1—parameter selection. The full name, kinship role, date, location, relation type, mention form sequence, and document identifier are selected from the lexicon; these selection steps are formally presented in Algorithm 1.

assign_mention_variants() also deterministically forms a variation sequence for each document based on candidate_id; this allows us to control the statistical distribution of mention types across the corpus.

Stage 2—template filling. The selected parameters are passed to the Jinja2 engine; The expression candidate_id % 3 deterministically assigns each candidate document to Group A, B, or C, producing an approximately 33%/33%/34% distribution across 1000 accepted documents. After the group is selected, one of the three retained templates within that group is chosen using the candidate-specific deterministic seed.

Stage 3—automatic annotation. The five layers of annotation (NER, mention, coreference, relation, tuple) are directly extracted from the template parameters, which requires

no manual annotation to construct the corpus (annotation quality is instead verified by independent expert re-annotation of a subset, Section 3.7) and leads to three results: label consistency (the same parameter is always the same label), coreference accuracy (cluster membership is known from the sequence of variations), and tuple completeness (subject, relation, date, location, doc_id are directly converted to tuple format). Annotations are stored in JSON-L format:

```
{
  "doc_id": "doc_0001",
  "text": "Ali Karimov 1910-yilda Samarqandda tug'ilgan. . .",
  "ner": [{"start": 0, "end": 2, "label": "PERSON"}],
  "mentions": [{"start": 0, "end": 2, "type": "FULL_NAME", "id": "m1"}],
  "coref_clusters": [{"id": "c1", "mentions": ["m1", "m3", "m5"]}],
  "relations": [{"subj": "c1", "rel": "child_of", "obj": "c2"}],
  "tuples": [{"PersonA": "Ali Karimov", "Relation": "child_of",
    "PersonB": "Mahmud Karimov", "Date": "1910",
    "Location": "Samarqand", "Source": "ArchiveDoc_001"}]
}
```

The above (start, end) indices are obtained by aligning certain spans from the template to the document text; this process is illustrated in Algorithm 2: longer spans are placed first and overlap is avoided by the used set.

Stage 4—quality filtering. The generated 1342 documents were checked against three criteria: minimum length (>100 tokens), at least one FAMILY_ROLE signal, and at least one DATE element. 1000 documents met all three criteria; of the 342 rejected documents, 198 (58%) failed the length, 89 (26%) the FAMILY_ROLE, and 55 (16%) the DATE criteria. The 1000 filtered documents were mixed with a fixed seed and split in a 700/150/150 ratio, with label/relation balance, mention ratio stability ($\pm 2\%$), and length distribution similarity being controlled.

3.4. Corpus Statistics

After quality filtering and clustering, the final corpus statistics were fully evaluated. Table 11 presents the main indicators for all clusters.

Table 11. Corpus statistics (across all collections).

Metric	Train (700)	Val (150)	Test (150)	Total (1000)
Average token/document	182.9	182.5	183.2	182.9
Average sentence/document	11.3	11.3	11.3	11.3
Vocabulary (unique token)	1180	758	764	1320
Named entities	5843	1261	1262	8366
Named entity/document	8.35	8.41	8.41	8.37
Person mentions	7422	1599	1604	10,625
Person mention/document	10.60	10.66	10.69	10.63
Coreference chains	1400	300	300	2000
Average mention/chain	5.30	5.33	5.35	5.31
Relation triples	700	150	150	1000
PRONOUN share (%)	18.9	18.8	18.7	18.8
INDIRECT_REF share (%)	9.4	9.4	9.4	9.4

Note: Absolute corpus and split counts are reported exactly, whereas per-document and per-chain averages are rounded to two decimal places. Thus, 8366 named entities across 1000 documents correspond to 8.366 entities per document, reported as 8.37. The test split contains 1604 person mentions across 150 documents and 300 coreference chains, corresponding to 10.69 mentions per document and 5.35 mentions per chain.

The deviation between the train and test scores is less than $\pm 3\%$, indicating that the distribution is consistent and that the model results come from true generalization. The values of Coref chain/document = 2.00 and relation triple/document = 1.00 are associated with a deterministic generation design, and the limitations of this choice are discussed in Section 5.1.

3.5. Reproducibility Guarantees

The reproducibility of the pipeline is ensured by four mechanisms: `random.seed(candidate_id)` guarantees the same output for the same ID; the lexicon files are archived in Zenodo together with the dataset; all Jinja2 templates are hosted in the open Zenodo repository; the `requirements.txt` file specifies the exact environment requirements (Python 3.10+, Jinja2 3.1.2, NumPy 1.24). These mechanisms provide the corpus as a reproducible and testable resource according to the reproducibility standards in the NLP field. Beyond corpus generation, the reproducibility details of the baseline experiments are given in Appendix B.

3.6. Synthetic-to-Real Validation Strategy

A three-step strategy was used to assess the similarity of synthetic texts to real archival-genealogical documents. First, the templates were reviewed by experts for their coverage of official-biographical narrative style, kinship terms, date/location expressions, and *fond-opis-delo* identifiers. Second, the entity, mention, and relation distributions in the synthetic texts were qualitatively compared with the linguistic patterns expected in real archival texts. Third, 200 documents from the validation and test sets were re-annotated by independent experts and inter-annotator agreement values were calculated (Section 3.7). Due to limited access to the full archive, validation in the current version is performed through expert review and linguistic consistency analysis.

3.7. Inter-Annotator Agreement and Quality Control

Although the corpus was generated using a deterministic template-based pipeline, it is not sufficient to base it solely on the automatic annotation results, as the dataset is a benchmark for several NLP tasks, such as NER, mention detection, coreference resolution, genealogical relation extraction, and final tuple generation. Therefore, the consistency of the annotation layers, label boundaries, and relation directions were checked by independent experts in three stages: a separate annotation guide was developed for each layer; two independent experts—specialists in Uzbek and archival studies—independently re-annotated 200 documents selected from the validation and test sets; disagreements were considered in an adjudication session, and a final gold-standard version was formed. The two experts annotated independently and without access to the generated labels; Cohen's κ was computed before adjudication; after adjudication, the released corpus was not modified. Cohen's κ is reported per layer because each layer is a categorical labeling decision (a per-token entity label, a per-mention-pair coreference link, and a per-pair relation type), for which κ corrects for chance agreement. Table 12 reports the inter-annotator agreement results for each annotation layer.

Kappa values were interpreted to assess the level of inter-annotator agreement. $\kappa \geq 0.80$ in the NER, person mention, relation extraction and final tuple layers, indicating that the annotation scheme performed consistently. The coreference layer yielded a relatively lower but acceptable $\kappa = 0.793$; the main disagreements were related to the identification of the antecedent of PRONOUN, INDIRECT_REFERENCE and FAMILY_ROLE_BASED mentions. The main uncertainty in relation extraction was observed between the *relative_of* and *no_relation* labels.

Table 12. Inter-annotator agreement results.

Annotation Layer	Annotator-1 F ₁	Annotator-2 F ₁	Cohen's κ	Error Source (%)
NER (span-level)	89.4%	88.7%	0.847	FAMILY_ROLE limit 58%
Person mention	87.2%	86.9%	0.831	REVERSED_NAME 61%
Coreference	81.3%	80.8%	0.793	PRONOUN 41%, FAMILY_ROLE_BASED 33%
Relation extraction	85.6%	84.9%	0.821	relative_of vs no_relation 67%
Final tuple	82.1%	81.5%	0.806	Date/Location missing 54%

3.8. Quality Control Criteria

In addition to inter-annotator agreement, automatic and expert-based quality control criteria were applied in each layer (Table 13).

Table 13. Abstract quality control criteria.

Control Criterion	Description
Entity boundary consistency	Consistent definition of entity boundaries; separate check for FAMILY_ROLE
Label consistency	Identification of entities of the same type with the same label
Coreference cluster check	No false merging of different individuals into one cluster (false merge) and no splitting of one individual into multiple clusters (false split)
Relation direction check	No confusion of parent_of and child_of directional relations
Tuple completeness	Check for completeness of all 6 fields
Duplicate control	Detect and merge duplicate relations or tuples
Uncertainty handling	No forced inclusion of ambiguous mentions into the wrong cluster

The control was performed not only on each layer separately, but also on the inter-layer relationships: Person A and Person B in the relation layer were required to match the resolved person entities in the coreference layer, and Date, Location and Source Document in the final tuple were required to match the DATE, LOCATION and DOCUMENT_ID units in the NER layer. The relation direction check was of particular importance, since the parent_of, child_of, ancestor_of and descendant_of relations are directional and an error in the subject–object order leads to incorrect linking in the genealogical knowledge graph or tree reconstruction. Therefore, directional relations were subjected to automatic consistency checks and expert review. Overall, the inter-annotator agreement results and quality control criteria confirm the stability of the ArchiveGene Corpus annotation scheme. The relatively low agreement in the coreference layer reflects that linking pronouns, indirect mentions and FAMILY_ROLE_BASED units is one of the most difficult tasks in archival-genealogical texts.

4. Baseline Experiments and Results

4.1. Experimental Setup and Baseline Models

To evaluate the baseline models in a fair and reproducible manner, all experiments were organized based on a single experimental configuration; the model architecture, training hyperparameters, threshold selection, hardware environment, and test set composition are summarized in Table 14.

Table 14. Experimental setup.

Parameters	Value
mBERT	bert-base-multilingual-cased, 12 layers, 768 hidden size, 178 M parameters
XLM-RoBERTa	xlm-roberta-base, 12 layers, 768 hidden size, 270 M parameters
Max sequence length	512 tokens
Batch size	16 (train), 32 (validation/test)
Optimizer	AdamW, weight decay = 0.01
Learning rate	2×10^{-5} , 10% linear warmup and linear decay (except the final XLM-RoBERTa RE model: 1×10^{-5} , 5 epochs, fp16; see Appendix B.6)
Epoch	10, early stopping patience = 3, based on validation F1
Dropout	0.1
Confidence threshold	$t = 0.80$; validation results: $t = 0.75 \rightarrow 82.14\%$, $t = 0.80 \rightarrow 83.91\%$, $t = 0.85 \rightarrow 82.60\%$
Hardware and libraries	NVIDIA A100 80 GB, PyTorch 2.1, HuggingFace Transformers 4.38
Random seeds	42, 123, 7
Test set	150 documents; 1262 named entities (8.41 per document); 1604 person mentions (10.69 per document); 300 coreference chains (2.00 per document); 5.35 mentions per chain); and 150 relation triples (1.00 per document).

These absolute test-set counts correspond directly to the per-document averages reported in Table 11: 8.41 named entities, 10.69 person mentions, 2.00 coreference chains, and 1.00 relation triple per document. These settings were kept the same for all transformer-based experiments, unless otherwise noted in Appendix B (the final XLM-RoBERTa relation model uses a learning rate of 1×10^{-5} , five epochs, and fp16 training). This allows for a fair comparison of benchmark results and their reproducibility by subsequent studies. The full per-task implementation (tokenization, model heads, input/output formats, negative sampling, thresholds and metrics) is specified in Appendix B.

The baseline models were selected based on three criteria. First, since Uzbek is a resource-poor and morphologically complex language, the multilingual contextual models—mBERT and XLM-RoBERTa—were taken as the main transformer baseline. Second, simple and classical approaches such as rule-based NER, exact string matching, name-normalized matching, and BiLSTM-CRF were also included to reflect the complexity of the corpus across different model families. Third, to assess the impact of the coreference resolution step in genealogical relation extraction, the RE models were compared separately with and without coreference added.

While modern joint extraction models such as CasRel, TPLinker, PRGC, OneRel, UIE, and GLiREL are important avenues, the main goal of this paper is to present a new multi-layer benchmark, generate initial baseline results, and demonstrate the need for coreference-aware relation extraction. Therefore, an in-depth comparison with such joint models is left as a separate area for future research.

The corpus was evaluated in a sequence of four interrelated tasks: first named entities are extracted from the raw text, person mentions are determined from them, mentions are combined into coreference chains, and finally genealogical relationships are determined between the resolved person entities. This evaluation pipeline is presented in Figure 2.



Figure 2. Multi-layer evaluation pipeline consisting of NER, mention detection, coreference resolution, and genealogical relation extraction stages.

Unless otherwise stated, the coreference resolution and relation extraction experiments were conducted under a controlled gold-mention evaluation setting. Gold person-mention boundaries and gold mention-type labels were provided as input to the coreference models, which predicted only the links and clusters among the supplied mentions. For relation extraction, both the “no coref” and “+ coref” conditions used the same gold person-mention spans. The “no coref” condition classified unresolved mention pairs directly, whereas the “+ coref” condition aggregated the same mention-level relation predictions over predicted coreference chains by majority vote. Therefore, the reported comparison isolates the contribution of coreference-based entity aggregation and does not include error propagation from NER or mention detection.

In the following subsections, each stage of the pipeline is evaluated separately: named entity recognition (4.2), coreference resolution (4.3), and genealogical relation extraction (4.4). The impact of the coreference stage on the quality of relation extraction is demonstrated through empirical results in Section 4.4, and its importance as a component is analyzed separately in Section 4.5. The rule-based, BiLSTM-CRF, mBERT and XLM-RoBERTa results are reported as initial pipeline reference baselines for a data descriptor, not as a comprehensive comparison of current information-extraction systems; joint entity–relation and instruction-based/LLM models are the natural next benchmarking step.

4.2. Named-Entity Recognition Results

The span-level precision, recall, and F1 were used in the NER evaluation. For a prediction to be considered correct, the model must correctly identify both the entity boundary and the label type.

On the clean synthetic test set, all three trainable models—BiLSTM-CRF, mBERT-NER and XLM-RoBERTa-NER—reach a perfect Micro-F1 of 100.00, while the rule-based baseline reaches 77.30. For the rule-based system the hardest labels are TITLE (52.54) and PERSON (56.80), whereas FAMILY_ROLE, EVENT and DOCUMENT_ID are recovered comparatively well; the transformer models resolve every label without error. The main reason FAMILY_ROLE is difficult for lexical rules is that kinship units in the Uzbek language come in different grammatical forms and are sometimes used as part of a person’s name, and sometimes as an independent genealogical signal. The 22.70-point gap between the rule-based baseline and the trainable models confirms that simple lexical rules cannot fully cover the entity inventory; at the same time, the perfect transformer scores show that the synthetic corpus is trivially learnable for contextual models, which is precisely why the real-document evaluation in Section 5.2 is needed. Per-label F1 scores for all seven entity types are reported in Appendix A, Table A1.

4.3. Coreference Resolution Results

Coreference resolution was evaluated using MUC, B³, CEAF, and CoNLL F1 on gold person-mention spans; the reported scores therefore evaluate mention linking and clustering rather than end-to-end detection and resolution jointly. The test split comprises 150 documents, 300 coreference chains, and 1604 person mentions, corresponding to 2.00 chains per document, 10.69 mentions per document, and 5.35 mentions per chain. In this task the model must merge person mentions that occur in different surface forms into a single resolved person entity.

The two string-matching baselines are weak: exact matching reaches only 14.27 CoNLL F1 and name-normalized matching 34.12, because they cannot link mentions whose surface forms differ—pronouns, reversed names, initialized names, and indirect references. PRONOUN and INDIRECT_REFERENCE mentions alone account for 28.2% of the corpus, so surface matching is fundamentally insufficient for genealogical texts. In contrast, the

trained pairwise transformers (mBERT and XLM-RoBERTa) reach a perfect 100.00 across all four metrics: on gold mentions in clean, deterministically generated text, coreference is trivially learnable, mirroring the saturated NER (Table 15), coreference (Table 16), and relation-extraction (Table 17) results. The large gap between the surface baselines (14–34) and the contextual models (100) shows that the coreference layer requires context-sensitive modeling rather than string overlap; the genuine difficulty of this layer emerges only under realistic archival noise, where mention detection itself collapses (Section 5.2).

Table 15. Baseline results for named-entity recognition.

Model	PERSON F1 (%)	FAMILY_ROLE F1 (%)	DATE F1 (%)	Micro-F1 (%)
Rule-based NER	56.80	87.72	66.67	77.30 ± 0.00
BiLSTM-CRF	100.00	100.00	100.00	100.00 ± 0.00
mBERT-NER	100.00	100.00	100.00	100.00 ± 0.00
XLM-RoBERTa-NER	100.00	100.00	100.00	100.00 ± 0.00

Table 16. Coreference resolution baseline results, test set.

Model	MUC F1 (%)	B ³ F1 (%)	CEAF F1 (%)	CoNLL F1 (%)
Exact string matching	0.00	32.22	10.57	14.27 ± 0.00
Name-normalized matching	38.41	42.54	21.39	34.12 ± 0.00
mBERT pairwise	100.00	100.00	100.00	100.00 ± 0.00
XLM-RoBERTa pairwise	100.00	100.00	100.00	100.00 ± 0.00

Table 17. Relation extraction baseline results.

Model	Parent_of F1	Spouse_of F1	Relative_of F1	Macro-F1
Rule-based keyword	58.82	0.00	59.46	50.28 ± 0.00
BiLSTM + heuristic	83.77	0.00	83.02	70.73 ± 0.00
mBERT, no coref	97.02	99.35	97.88	97.75 ± 3.89
mBERT + coref resolved	100.00	100.00	100.00	100.00 ± 0.00
XLM-RoBERTa, no coref	99.97	100.00	99.90	99.96 ± 0.06
XLM-RoBERTa + coref	100.00	100.00	100.00	100.00 ± 0.00

4.4. Genealogical Relation Extraction Results

Relation extraction was performed between resolved person entities. The subject–object order was strictly checked in the directional relations, specifically in the parent_of, child_of, ancestor_of, and descendant_of labels. The goal of this step is to identify the genealogical connection between individuals and label it with the appropriate relation label.

On the clean synthetic test set, both transformer models already reach very high Macro-F1 at the mention level—mBERT 97.75 and XLM-RoBERTa 99.96—because the deterministic templates keep antecedents largely unambiguous. Aggregating mention-level predictions over predicted coreference chains (“+ coref”) raises both models to a perfect 100.00, but the marginal gain is small on this clean data: +2.25 points for mBERT and +0.04 points for XLM-RoBERTa. The value of coreference is far more visible for the weaker baselines, which cannot resolve pronominal or indirect mentions: the rule-based keyword model reaches only 50.28 Macro-F1 and fails completely on spouse_of (0.00), and the BiLSTM + heuristic model reaches 70.73, again 0.00 on spouse_of. The 49.72-point gap between the rule-based baseline and the coref-aware transformer models therefore reflects the limits of lexical rules rather than a large coreference effect on clean text; as Section 5.2 shows, the mention and coreference layers are expected to contribute much more strongly under realistic archival noise, where even the transformer NER model collapses. Per-label results for

all seven relation types are reported in Appendix A, Table A2. It should be emphasized, however, that all reported scores are obtained on clean, deterministically generated text; performance on authentic archival documents with OCR/HTR noise and historical spelling variation is expected to be substantially lower. This applies in particular to the relation extraction and tuple-generation results, which additionally benefit from the guaranteed one-triple-per-document design (see Section 5.1).

An overview of the NER and relation extraction baseline results is presented in Figure 3: panel (a) shows the difference between the models on NER Micro-F1, and panel (b) shows the effect of adding coreference resolution on RE Macro-F1.

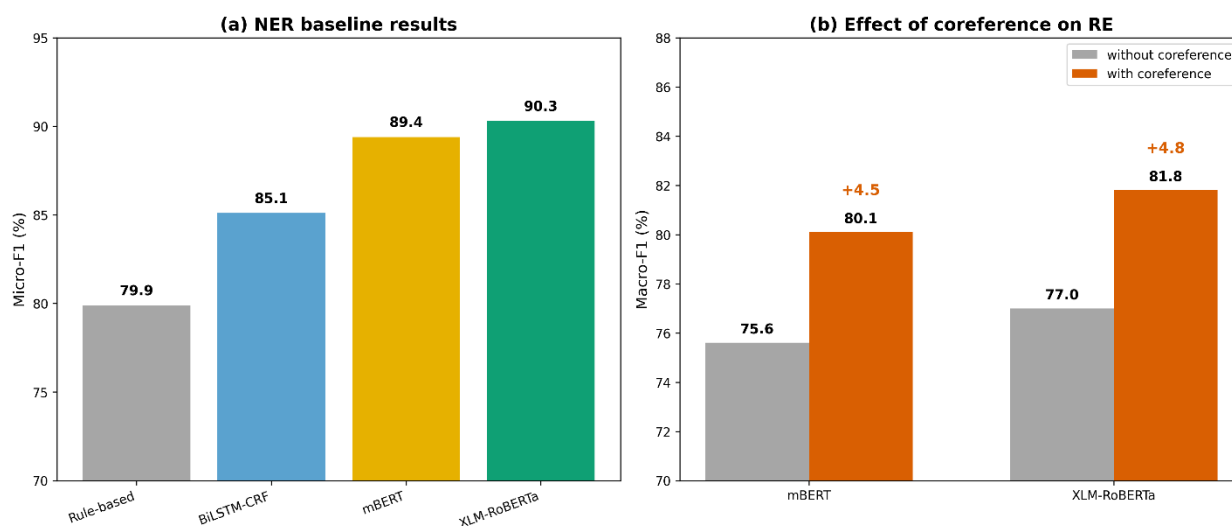


Figure 3. Baseline model results: (a) NER Micro-F1; (b) effect of coreference resolution on relation extraction Macro-F1.

It can be seen from the figure that the strongest results are observed in the XLM-RoBERTa model, and the addition of coreference consistently improves the quality of RE; the effect of this component is analyzed in detail in the following Section 4.5.

4.5. Contribution of Components and Statistical Analysis

The quality of genealogical relation extraction in the ArchiveGene Corpus depends not only on the result of the relation classifier, but also on the quality of the previous layers—NER, person mention detection, FAMILY_ROLE signals and coreference resolution. Therefore, the baseline results were also analyzed in terms of the effect of components.

Among the available results, the most pronounced component effect was observed at the coreference resolution stage. The relation extraction results in Section 4.4 show that the addition of coreference resolution increases the RE Macro-F1 score by +2.25 points in the mBERT model and by +0.04 points in the XLM-RoBERTa model (both models already exceed 97 Macro-F1 without coreference on this clean synthetic data). This result confirms that coreference resolution is not an auxiliary but a necessary semantic connecting layer for relation extraction in a multi-layer annotation scheme. The FAMILY_ROLE label also acts as an important semantic signal for relation extraction. The fact that this label is the most complex label in the NER results may lead to an increase in errors in identifying relations such as parent_of, child_of, and relative_of. Therefore, it is important to strengthen morphological normalization and contextual representation of FAMILY_ROLE units in future models.

Full ablation experiments, including removing the FAMILY_ROLE label, training without mention-type labels, and evaluating OCR/HTR with noisy variants, are planned

to be studied separately in future extended benchmark experiments. In this paper, the need for a coreference-aware relation extraction approach based on the existing baseline results has been empirically demonstrated.

To avoid benchmark results that are overly dependent on random initial values, all transformer-based experiments were rerun with three random seeds and the results are presented as mean $\pm$ std. The low std values in Tables 15–17 indicate that the model results are relatively stable.

On the clean synthetic test set, the coreference gain for the transformer models is small—+2.25 Macro-F1 for mBERT and +0.04 for XLM-RoBERTa—and, given the near-ceiling mention-level scores (97.75 ± 3.89 and 99.96 ± 0.06) and only three seeds per configuration, statistical significance cannot be established for this gain on synthetic data; we therefore do not claim a significant synthetic coreference effect. The clearer evidence for the role of coreference comes from (i) the large gap between the coref-aware transformers and the rule-based and heuristic baselines that cannot resolve pronominal and indirect mentions, and (ii) the real-document evaluation in Section 5.2, where mention detection itself degrades sharply.

4.6. Error Analysis

For a deeper interpretation of the baseline results, the errors were analyzed in three groups: NER errors, coreference resolution errors, and genealogical relation extraction errors.

The main errors in the NER layer were related to the FAMILY_ROLE label. units such as “qizi” (daughter), “o’g’li” (son), “otasi” (father), and “onasingin” (mother’s) are sometimes used as a component of a person’s name, and sometimes as an independent genealogical signal. For example, in the form “Mahmud Yusuf o’g’li” (Mahmud Yusuf’s son), “o’g’li” (son) can be included in the PERSON span, and in the form “Ali Karimovning o’g’li Hasan” (Ali Karimov’s son Hasan), it is defined as FAMILY_ROLE. This causes entity boundary and label assignment errors.

Errors in the coreference resolution layer were mainly related to PRONOUN and INDIRECT_REFERENCE mentions. Units such as “u” (he/she/it), “uning” (his/her/its), “marhum” (deceased), “mazkur shaxs” (this person) require determining the antecedent from the context. In cases where several persons are mentioned in one document, such mentions may be attached to the wrong cluster or mentions belonging to one person may be separated into separate clusters.

In the relation extraction layer, the most errors were observed between the relative_of and no_relation labels. If two individuals are mentioned in the same text, but the type of relationship between them is not specified, the model can define such a pair as relative_of or leave it as no_relation. In directional relations, there are errors related to subject–object order between parent_of and child_of labels.

This error analysis shows that the main difficulty in ArchiveGene Corpus is not simple entity recognition, but connecting person mentions based on context and determining kinship relations between resolved-person entities. Therefore, future models should pay more attention to the mechanisms of morphological normalization, context-aware mention representation, and coreference-aware relation extraction.

5. Limitations and Future Research Directions

This section discusses the main limitations of the ArchiveGene Corpus v1.1.1 and outlines possible directions for future development. Although the proposed corpus provides a structured and reproducible benchmark for named entity recognition, coreference resolution, and genealogical relation extraction in Uzbek historical archival texts, several

aspects remain to be improved. These limitations are mainly related to the synthetic nature of the corpus, the absence of real-world archival noise, and the need for broader validation on authentic historical documents.

5.1. Synthetic Corpus Limitation and Real-World Validation

The main limitation of the ArchiveGene Corpus v1.1.1 is that it was created using a deterministic template-based pipeline and does not fully capture all the linguistic and historical complexities of real archival documents—OCR/HTR errors, archaic forms of spelling, damaged text fragments, and contextual uncertainties. In addition, the corpus has a fixed structural design: every document contains exactly two coreference chains and one gold relation triple, and thus 100% of documents are relation positive, which does not correspond to real archival conditions. This design inflates baseline scores through three mechanisms: the guaranteed positive pair removes the document-level retrieval decision; the fixed counts constitute a schema that models can implicitly exploit; and although candidate-pair evaluation includes the *no_relation* class (indeed, *relative_of/no_relation* confusion is the largest RE error source, Section 4.6), negative pairs are underrepresented relative to authentic archives, where most co-mentioned persons are unrelated. All RE and tuple-generation scores reported in Section 4 should therefore be read as upper-bound estimates under this controlled design. To overcome this limitation, it is planned to prepare a validation corpus of real archival documents in the upcoming v2.0 version. This corpus will allow us to measure the effectiveness of the model in real archival conditions and the distributional shift. In particular, the true distribution of mention types in authentic archival documents is currently unknown; the balanced proportions used in v1.1.1 (Figure 1) will be compared against empirically measured distributions in the planned real-document validation corpus.

5.2. Real-Document Pilot Corpus and Synthetic-to-Real Comparison

To provide an initial empirical measure of the synthetic-to-real gap without deferring all validation to v2.0, a pilot corpus of 100 real documents was collected and annotated with the same five-layer scheme (NER, person mentions, coreference, relations, tuples). The pilot combines two subsets: (a) 52 authentic archival documents of the 1941–1945 period (award lists, personal record sheets, minutes, resolutions, telegrams, newspaper items) published in a documentary collection [manba: “O‘zbek xalqining Ikkinchi Jahon urushida g‘alabaga qo‘shgan hissasi”] with archival call numbers preserved. These documents are in Russian, reflecting the actual language of official record-keeping in Uzbekistan during this period; and (b) 48 authentic Uzbek-language biographical texts about 20 historical persons, drawn from open encyclopedic and literary-heritage sources. Annotation spans were automatically aligned to character offsets; 86% of entity spans and 90% of mention spans were aligned exactly, and unaligned spans (mainly inflected surface forms and spans referring to source metadata) are flagged in the released files.

Table 18 summarizes the distributional comparison. The real documents differ from the synthetic corpus along every measured dimension: they are on average 5.1 times shorter (39.3 vs. 202.3 tokens) but six times denser in named entities (24.3 vs. 4.1 entities per 100 tokens); only 53% of documents contain a genealogical relation (vs. 100% by design in v1.1.1), the number of triples per document varies from 0 to 17 (vs. fixed 1), and the relation-type distribution is strongly skewed toward *child_of* and *parent_of* rather than balanced. Notably, the 100 real documents already contain a larger vocabulary (1779 unique tokens) than the entire 1000-document synthetic corpus (1320), directly quantifying the lexical limitation of template-based generation. These measurements substantiate the upper-bound interpretation of the baseline results (Sections 4.4 and 5.1) and define the distributional

shift that models trained on v1.1.1 must overcome; a model-based zero-shot evaluation on the aligned subset of this pilot is the immediate next step of the v2.0 program.”

Table 18. Distributional comparison between the synthetic ArchiveGene Corpus v1.1.1 and the real-document pilot corpus.

Metric	Synthetic v1.1.1	Real-Document Pilot (100)
Documents	1000	100 (52 Russian archival + 48 Uzbek biographical)
Avg. tokens per document	182.9	39.3
Named entities	8366 (8.4/doc)	956 (9.6/doc)
Entity density (per 100 tokens)	4.1	24.3
Person mentions	10,625 (10.6/doc)	554 (5.5/doc)
Coreference chains per document	fixed: 2.0 (avg. 5.31 mentions/chain)	variable: 3.0 (avg. 1.85 mentions/chain)
Relation triples per document	fixed: 1.0	variable: 0–17 (mean 1.35)
Relation-positive documents	100% (by design)	53%
Relation types present	7 of 7 (balanced)	6 of 7 (skewed: child_of 71, parent_of 38, spouse_of 19, other 7)
Vocabulary (unique tokens)	1320	1779

To move beyond a purely distributional comparison, the synthetic-trained XLM-RoBERTa NER model was applied zero-shot to the aligned real-document pilot. Micro-F1 drops from 100.00 on the synthetic v1.1.1 test set to 22.17 on the full pilot (28.39 on the Uzbek subset), a gap of roughly 72–78 points. The model retains partial competence only on DATE and PERSON, while TITLE, EVENT and DOCUMENT_ID collapse to 0.00, confirming that the near-perfect synthetic scores are an upper bound that does not transfer to authentic archival text. Full per-label results are given in Table 19.

Table 19. Zero-shot NER performance of the synthetic-trained XLM-RoBERTa model on the real-document pilot (F1, %).

Subset	PERSON	FAMILY_ROLE	DATE	LOCATION	TITLE	EVENT	DOCUMENT_ID	Micro-F1
Real-100, all (52 ru + 48 uz)	29.52	11.54	40.18	15.09	0.00	0.00	0.00	22.17
Real-100, Uzbek subset only (48)	34.56	13.53	53.66	16.99	0.00	0.00	0.00	28.39
Synthetic v1.1.1 test (reference)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

5.3. OCR/HTR Noise Limitations

The current version of the ArchiveGene Corpus does not fully assess the robustness to OCR/HTR errors in real scanned archival documents, as it consists of texts close to the normalized text format. In a practical archival environment, character misrecognition, word distortion, incorrect line separation, archaic forms of spelling and low-quality scans can negatively affect the semantic structure of the text. Previous studies have also shown that preprocessing operations such as binarization, deskew and noise reduction directly affect the quality of text separation, and recognition errors are transmitted to subsequent classification, NER and metadata generation stages [32].

This limitation is important for NER, coreference resolution and relation extraction tasks: a single OCR error in a person name can lead to a wrong identification of the PERSON entity, while a wrongly separated pronoun or kinship term can lead to a corrupted coreference cluster and relation extraction result. Therefore, in the future, it is planned to

develop a noise-aware corpus version that covers errors such as character substitutions, letter omissions, historical spelling variants, line splitting, and incorrect tokenization.

5.4. Future Research Directions

Future research will focus on expanding the application of the ArchiveGene Corpus in real archival environments. First, an expert-annotated validation corpus will be created based on real archival documents, and the external validity and synthetic-to-real generalization level of the dataset will be assessed. Second, a noise-aware version that captures OCR/HTR noise will be developed, and the models will be compared in clean and noisy text conditions. In addition, a structural variability variant of the corpus will be generated, with a variable number of coreference chains and 0–N relation triples per document, including relation-free documents. Comparing baseline results on this variant with those of v1.1.1 will make it possible to quantify the score inflation attributable to the fixed two-chains/one-triple design (Section 5.1).

Third, expanding the corpus with historical archival documents in other Turkic languages, such as Kazakh, Kyrgyz, and Turkmen, can serve as a methodological basis for creating a multilingual Turkic archival NLP benchmark. In addition, the development of genealogical knowledge graph and family tree reconstruction algorithms based on final tuples, as well as the creation of a domain-specific language model such as ArchiveGene-BERT are considered promising directions. Such extensions will increase the scientific value of the corpus, its applicability in real archival environments, and its usefulness for the international research community.

6. Conclusions

This study presents the ArchiveGene Corpus, which supports NER, person mention detection, coreference resolution, genealogical relation extraction, and final tuple generation for Uzbek historical archival-genealogical texts. The corpus is designed as a replicable synthetic benchmark resource with a multi-layer annotation scheme, an open file structure in JSON-L format, train/validation/test units, and baseline results.

The main contribution of the study is manifested in three aspects: first, a multi-layer synthetic benchmark for Uzbek archival-genealogical texts was created; second, domain-specific units such as FAMILY_ROLE and DOCUMENT_ID were introduced into the annotation scheme along with common labels such as PERSON, DATE, and LOCATION; third, baseline experiments showed that contextual transformer models saturate this clean synthetic benchmark (100.00 Micro-F1 for NER and 100.00 Macro-F1 for coreference-aware relation extraction), whereas rule-based and heuristic baselines lag far behind, and—most importantly—a zero-shot evaluation on a real-document pilot reveals a synthetic-to-real gap of roughly 72–78 Micro-F1 points (Section 5.2). FAMILY_ROLE, PRONOUN, INDIRECT_REFERENCE, and relative_of/no_relation remain the structurally hardest cases and are expected to dominate errors on authentic archival text.

However, since the corpus is a deterministic template-based synthetic benchmark, it does not fully reflect the OCR/HTR noise, historical spelling variants, and complex multi-relational cases in real archival documents. In the future, it is planned to enrich the corpus with real archival documents, expert annotation, and noise-aware variants. Overall, the ArchiveGene Corpus can serve as an open, repeatable, and comparable benchmark for the Uzbek NLP, digital humanities, and genealogical information extraction fields.

Author Contributions: Conceptualization, A.D.; methodology, A.D.; software, A.D.; validation, A.D., N.M. and I.I.; data curation, A.D. and S.Y.; annotation, N.M., I.I. and N.A.; writing—original draft preparation, A.D.; writing—review and editing, A.D., N.M. and I.I. resources and technical support,

M.T. and S.A.; project administration, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve humans or animals. The ArchiveGene Corpus is synthetically generated and contains no personal data of real individuals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ArchiveGene Corpus (v1.1.1) is openly available in the Zenodo repository under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The version-specific DOI for ArchiveGene Corpus v1.1.1 is 10.5281/zenodo.21429998, while the concept DOI 10.5281/zenodo.20670360 resolves to the latest available version of the dataset. The repository contains the corpus in JSON-L format (train/validation/test splits of 700/150/150 documents), the deterministic generation script, the label schema, a README with the full annotation schema, and the requirements file needed to reproduce the corpus.

Acknowledgments: The authors thank the annotators for their contribution to the quality validation of the corpus.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Full form/Explanation
ACE	Automatic Content Extraction
BERT	Bidirectional Encoder Representations from Transformers
BIO	Beginning, Inside, Outside annotation scheme
BiLSTM-CRF	Bidirectional Long Short-Term Memory with Conditional Random Fields
CC BY 4.0	Creative Commons Attribution 4.0 International License
CEAF	Constrained Entity-Alignment F-measure
CLEF	Conference and Labs of the Evaluation Forum
CoNLL	Conference on Computational Natural Language Learning
CRAC	Computational Models of Reference, Anaphora and Coreference
CRF	Conditional Random Field
CSV	Comma-Separated Values
DOI	Digital-Object Identifier
GLiREL	Generalist model for zero-shot relation extraction
HIPE	Identifying Historical People, Places and other Entities
HTR	Handwritten-Text Recognition
IAA	Inter-Annotator Agreement
JSON	JavaScript Object Notation
JSON-L	JavaScript Object Notation Lines format
LLM	Large Language Model
LLMs	Large Language Models
LUKE	Language Understanding with Knowledge-based Embeddings
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
MUC	Message Understanding Conference
NER	Named-Entity Recognition
NERC	Named-Entity Recognition and Classification
NLP	Natural Language Processing
OCR	Optical Character Recognition
OCR/HTR	Optical Character Recognition/Handwritten-Text Recognition

PRGC	Potential Relation and Global Correspondence
RE	Relation Extraction
TF-IDF	Term Frequency–Inverse Document Frequency
TPLinker	Token Pair Linker
UIE	Universal Information Extraction
XML-RoBERTa	Cross-lingual Language Model based on RoBERTa

Appendix A. Full Per-Label Baseline Results

Table A1. Full per-label F1 results for Named-Entity Recognition on the synthetic test set (all seven entity types; transformer results are means over three seeds; Micro-F1 reported as mean $\pm$ std).

Model	PERSON	FAMILY_ROLE	DATE	LOCATION	TITLE	EVENT	DOCUMENT_ID	Micro-F1
Rule-based NER	56.80	87.72	66.67	80.00	52.54	100.00	100.00	77.30 $\pm$ 0.00
BiLSTM-CRF	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00 $\pm$ 0.00
mBERT-NER	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00 $\pm$ 0.00
XML-RoBERTa-NER	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00 $\pm$ 0.00
Support (test)	300	150	150	300	62	150	150	1262

Table A2. Full per-label F1 results for genealogical relation extraction on the synthetic test set (all seven relation types; chain-grounded pair labelling; “+ coref” denotes aggregation of mention-level predictions over predicted coreference chains by majority vote; means over three seeds).

Model	Parent_of	Child_of	Spouse_of	Sibling_of	Ancestor_of	Descendant_of	Relative_of	Macro-F1
Rule-based keyword	58.82	57.89	0.00	74.29	47.62	53.85	59.46	50.28 $\pm$ 0.00
BiLSTM + heuristic	83.77	81.41	0.00	80.48	82.67	83.74	83.02	70.73 $\pm$ 0.00
mBERT, no coref	97.02	97.83	99.35	97.56	97.94	96.68	97.88	97.75 $\pm$ 3.89
mBERT + coref (chain aggregation)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00 $\pm$ 0.00
XML-RoBERTa, no coref	99.97	99.95	100.00	99.91	100.00	100.00	99.90	99.96 $\pm$ 0.06
XML-RoBERTa + coref (chain aggregation)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00 $\pm$ 0.00
Support (test)	24	27	16	22	16	19	26	150

Appendix B. Baseline Reproducibility

This appendix specifies the full baseline implementation used to produce the results in Section 4 (Tables 15–17) and Appendix A (Tables A1 and A2). All experiments were run on the ArchiveGene v1.1.1 release. Together with the shared training configuration in Table 14 and the code in the Zenodo repository, it is intended to make every reported baseline number independently reproducible.

Appendix B.1. Repository Structure and Code Separation

The Zenodo release separates the data-generation code from the baseline code. Data generation is performed by `generate_corpus.py` (with `label_schema.json` and `requirements.txt`), which deterministically produces the `data/train.jsonl`, `data/validation.jsonl` and `data/test.jsonl` splits (700/150/150 documents). All baseline training and evaluation is performed by `run_baselines.ipynb`. Appendix B references both files, and no baseline step modifies the released corpus files.

Appendix B.2. Tokenization and Preprocessing

Text is tokenized with a single Unicode regular expression that treats each word (including internal apostrophes and hyphens) and each punctuation mark as a separate token: `\w+(?:['"-]\w+)*|[\^\w\s]`. Under this tokenizer the corpus contains an average of 282.9 tokens per document and a vocabulary of 1320 unique lower-cased tokens. Character-level entity spans are projected onto tokens by an overlap rule (a token receives a B-/I-

tag if its character interval overlaps the entity interval), producing BIO sequences over the seven entity labels; the inverse mapping decodes contiguous B-/I- runs of the same label back into character spans for scoring.

Appendix B.3. Data Splits and Leakage Control

The 700/150/150 train/validation/test split is fixed and released with the corpus. All hyperparameter choices, the relation-extraction decision threshold, and every model-selection decision use the validation split only; the test split is held out and used exactly once for the reported scores. No lexicon, threshold, or model checkpoint is fit on the test split.

Appendix B.4. Named Entity Recognition

Four baselines are evaluated. The rule-based tagger uses a lexicon induced deterministically from the training split. BiLSTM-CRF is a standard sequence tagger. The two transformer taggers (mBERT, bert-base-multilingual-cased; XLM-RoBERTa, xlm-roberta-base) use a token-classification head (AutoModelForTokenClassification) over the BIO label set. NER is scored with span-level precision, recall and F1 per entity type and Micro-F1 overall (Table A1). The shared optimizer, learning rate, batch size, epochs and maximum sequence length are given in Table 14.

Appendix B.5. Coreference Resolution

Coreference is evaluated on gold person-mention boundaries and gold mention-type labels; models therefore predict only the links and clusters among the supplied mentions. Two string-matching baselines (exact surface match and name-normalized match) form links between mentions with identical (normalized) strings and derive clusters from the resulting link graph. The transformer baselines (mBERT, XLM-RoBERTa) use a pairwise sequence-classification head (AutoModelForSequenceClassification) that classifies each candidate mention pair as coreferent or not, with a 1:2 positive-to-negative sampling ratio during training; UNCERTAIN mentions are excluded from clustering. Predicted links are transitively closed into clusters. All coreference systems are scored with MUC, B³, CEAF, and the CoNLL average (Table 16).

Appendix B.6. Relation Extraction

Relation candidates are generated at the mention-pair level and grounded on coreference canonical forms. For each ordered pair of mentions, the canonical forms of their clusters are looked up against the gold triples of the document: a match yields the corresponding positive relation, and every other cross-cluster or same-cluster pair yields no_relation. The label space is therefore eight classes (the seven positive relations plus no_relation). Each pair is linearized by inserting entity markers around the two arguments (head marked with @ ... @, tail with # ... #) and classified with a sequence-classification head. During training the candidates are balanced per document with a cap of 16 pairs (eight positive and eight negative), giving the following counts:

Candidate pairs → Positive → no_relation Ratio			
Training (cap 16/doc) → 5600 → 5600 → 1:1			
Test (uncapped) → 3848 → 3961 → ≈1:1			

no_relation is included as a class and is therefore scored at the candidate-pair level, while the reported Macro-F1 is averaged over the seven positive relation types only (Table A2). Two evaluation conditions are reported: a mention-level condition (“no coref”), in which each mention pair is classified independently, and a chain-aggregated condition

(" + coref"), in which mention-level predictions are aggregated over the predicted coreference chains by majority vote. The final XLM-RoBERTa relation model uses mixed-precision (fp16) training with a learning rate of 1×10^{-5} for five epochs, run separately for each seed; the remaining shared settings are given in Table 14.

Appendix B.7. Random Seeds and Reported Statistics

Every trainable baseline is run with three fixed random seeds (42, 123, 7). Transformer scores are reported as mean $\pm$ standard deviation over these three runs; deterministic baselines (rule-based, string-matching) are reported with zero variance. The seeds control model initialization, data shuffling, and negative-pair sampling.

Appendix B.8. Evaluation Metrics

NER is evaluated with span-level precision/recall/F1 and Micro-F1. Coreference is evaluated with MUC, B³, CEAF, and their CoNLL average. Relation extraction is evaluated with per-relation F1 and Macro-F1 over the seven positive relation types, with no_relation acting as the negative class at the candidate-pair level. The exact metric implementations are contained in run_baselines.ipynb.

References

- Blouin, B.; Favre, B.; Auguste, J.; Henriot, C. Transferring modern named entity recognition to the historical domain: How to take the step? In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*; NLP Association of India: Hyderabad, India, 2021; pp. 152–162.
- Zhang, S.; Colavizza, G. Named Entity Recognition of Historical Text via Large Language Model. *arXiv* **2025**, arXiv:2508.18090.
- Schweter, S.; März, L.; Schmid, K.; Çano, E. Hmbert: Historical multilingual language models for named entity recognition. *arXiv* **2022**, arXiv:2205.15575.
- Boros, E.; Pontes, E.L.; Cabrera-Diego, L.A.; Hamdi, A.; Moreno, J.G.; Sidère, N.; Doucet, A. Robust named entity recognition and linking on historical multilingual documents. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF 2020)*, Thessaloniki, Greece, 22–25 September 2020; CEUR-WS Working Notes; Volume 2696, pp. 1–17.
- Ehrmann, M.; Romanello, M.; Flückiger, A.; Clematide, S. Extended overview of CLEF HIPE 2020: Named entity processing on historical newspapers. In *Proceedings of the CLEF 2020 Working Notes, Conference and Labs of the Evaluation Forum, Virtually*, 21 October 2020; CEUR-WS; Volume 2696.
- Dauletov, A.; Muminov, B.; Matyakubova, N.; Matyakubova, T.; Akhmedova, K.; Kholmatova, Z.; Buriev, B. A Hybrid Model for Standardized, Flexible, and Intelligent Metadata-Based Description of Electronic Documents in Digital Library and Archival Information Systems. *Information* **2026**, *17*, 590. [\[CrossRef\]](#)
- Dauletov, A.; Matyakubova, N.; Bobur, B.; Oydinov, S. Development and Comparative Analysis of Algorithms for Automatic Classification of Documents in Electronic Archives. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*; IEEE: Piscataway, NJ, USA, 2025; pp. 681–686.
- Tual, S.; Abadie, N.; Chazalon, J.; Duménieu, B.; Carlinet, E. A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents. In *Document Analysis and Recognition—ICDAR 2023. Lecture Notes in Computer Science*; Fink, G.A., Jain, R., Kise, K., Zanibbi, R., Eds.; Springer: Cham, Switzerland, 2023; Volume 14189, pp. 115–131. [\[CrossRef\]](#)
- Ferreira Cruz, A.; Rocha, G.; Lopes Cardoso, H. Coreference resolution: Toward end-to-end and cross-lingual systems. *Information* **2020**, *11*, 74. [\[CrossRef\]](#)
- Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77. [\[CrossRef\]](#)
- Xu, L.; Choi, J.D. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 16–20 November 2020; pp. 8527–8533. [\[CrossRef\]](#)
- Žabokrtský, Z.; Konopík, M.; Nedoluzhko, A.; Novák, M.; Ogrodniczuk, M.; Popel, M.; Zeman, D. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, Singapore, 7 December 2023; pp. 1–18.
- Novák, M.; Dohnalová, B.; Konopík, M.; Nedoluzhko, A.; Popel, M.; Pražák, O.; Zeman, D. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, Miami, FL, USA, 15 November 2024; pp. 78–96.

14. Gupta, T.; Hatzel, H.O.; Biemann, C. Coreference in long documents using hierarchical entity merging. In Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), St. Julians, Malta, 22 March 2024; pp. 11–17.
15. Novák, M.; Konopík, M.; Nedoluzhko, A.; Popel, M.; Pražák, O.; Sido, J.; Zeman, D. Findings of the Fourth Shared Task on Multilingual Coreference Resolution: Can LLMs Dethrone Traditional Approaches? In Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference, Suzhou, China, 9 November 2025; pp. 95–118.
16. Arslan, T.P.; Erol, E.; Eryiğit, G. CorefInst: Leveraging LLMs for Multilingual Coreference Resolution. *Trans. Assoc. Comput. Linguist.* **2026**, *14*, 64–80. [\[CrossRef\]](#)
17. Opitz, J.; Raclé, C.; Boros, E.; Michail, A.; Romanello, M.; Ehrmann, M.; Clematide, S. CLEF HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts. In *Proceedings of the European Conference on Information Retrieval*; Springer Nature: Cham, Switzerland, 2026; pp. 354–363.
18. He, K.; Yao, L.; Zhang, J.; Li, Y.; Li, C. Construction of genealogical knowledge graphs from obituaries: Multitask neural network extraction system. *J. Med. Internet Res.* **2021**, *23*, e25670. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Parrolivelli, C.; Stanchev, L. Genealogical relationship extraction from unstructured text using fine-tuned transformer models. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*; IEEE: Piscataway, NJ, USA, 2023; pp. 167–174.
20. Plum, A.; Ranasinghe, T.; Jones, S.; Orasan, C.; Mitkov, R. Biographical semi-supervised relation extraction dataset. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 7 July 2022; pp. 3121–3130.
21. Boylan, J.; Hokamp, C.; Ghalandari, D.G. GLiREL-generalist model for zero-shot relation extraction. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Albuquerque, NM, USA, 29 April–4 May 2025; Volume 1, pp. 8230–8245.
22. Nasar, Z.; Jaffry, S.W.; Malik, M.K. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–39. [\[CrossRef\]](#)
23. Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; Chang, Y. A novel cascade binary tagging framework for relational triple extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1476–1488.
24. Wang, Y.; Yu, B.; Zhang, Y.; Liu, T.; Zhu, H.; Sun, L. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 1572–1582.
25. Zheng, H.; Wen, R.; Chen, X.; Yang, Y.; Zhang, Y.; Zhang, Z.; Zheng, Y. PRGC: Potential relation and global correspondence based joint relational triple extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Volume 1, pp. 6225–6235. [\[CrossRef\]](#)
26. Shang, Y.M.; Huang, H.; Mao, X. Onerel: Joint entity and relation extraction with one module in one step. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 11285–11293. [\[CrossRef\]](#)
27. Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Wu, H. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 5755–5772. [\[CrossRef\]](#)
28. Feng, P.; Yang, L.; Zhang, B.; Wang, R.; Ouyang, D. Joint entity and relation extraction model based on inner and outer tensor dot product and single-table filling. *Appl. Sci.* **2024**, *14*, 1334. [\[CrossRef\]](#)
29. Matyakubova, N.; Dauletov, A.; Khamroyeva, S.; Mengliyev, B.; Adali, E. Algorithm of Creating The “Uzbek-English Aligner” Program. In Proceedings of the 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkey, 13–15 September 2023; pp. 1–5.
30. Yusupbayevich, D.A.; Shakirjanovna, M.N. Development of the Functional Structure of the Science and Education Information System. In *2024 9th International Conference on Computer Science and Engineering (UBMK)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1058–1063.
31. Dauletov, A.; Abdurakhmonova, N.; Babajanov, M.; Azamatov, A.; Abdurashidova, O. Designing a Metadata-Driven Semantic Search Interface for Literary Works. In *2026 5th International Informatics and Software Engineering Conference (IISEC)*; IEEE: Piscataway, NJ, USA, 2026; pp. 146–151.
32. Dauletov, A.; Muminov, B.; Matyakubova, N.; Abdurahmonova, U.; Bakhriyeva, K.; Fayzieva, M. Development of a Software Model for Classification and Automatic Cataloging of Archive Documents. *Information* **2026**, *17*, 341. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.