

DOI: 10.15514/ISPRAS-2026-38(6)-11



Automatic Restoration of Birchbark Manuscripts Using Masked Language Modeling

¹ M.A. Eremeev, ORCID: 0009-0009-0546-5134 <maxim.a.eremeev@gmail.com>^{2,3} I.S. Humonen, ORCID: 0000-0003-2621-507X <humonen@axxx.tech>^{2,3} M.Yu. Golyadkin, ORCID: 0000-0002-0679-6981 <mgolyadkin@hse.ru>¹ A.A. Fitiskina, ORCID: 0000-0002-5482-1416 <annaftiskina@gmail.com>^{2,3} I.A. Makarov, ORCID: 0000-0002-3308-8825 <iamakarov@hse.ru>¹ National Research University Higher School of Economics,
20, Myasnitskaya St., Moscow, 101000, Russia.² AXXX, 32k1 Kutuzovsky Ave., Moscow, 121170, Russia.³ Russia Trusted AI Center, RAS,
25, Aleksandra Solzhenitsyna St., Moscow, 109004, Russia.

Abstract. This work addresses the automatic restoration of lacunae in Old Novgorodian birchbark manuscripts, a challenging task due to the small available corpus and the distinctive dialectal features of the target variety. We systematically compare three BERT-like encoders – mBERT, BERTslav, and ModernBERT – in character-level and token-level prediction modes, zero-shot and after fine-tuning, on a purpose-built corpus of Old Russian and Old Church Slavonic texts. After domain fine-tuning, ModernBERT achieves the strongest restoration, reaching token-level top-1 accuracy of 30.72% on real editorial lacunae and 91.40% on artificially masked text. At the character level, however, the fine-tuned encoders are matched or surpassed by a simple character n-gram baseline, revealing a structural mismatch between subword pre-training and single-character prediction. We further probe the frozen embeddings for document genre and date: they carry useful signal for dating, whereas genre proves largely a surface-orthographic property that a character TF-IDF baseline captures as well or better.

Keywords: birchbark manuscripts; lacuna restoration; old Novgorodian dialect; old Russian; Church Slavonic; masked language modeling; low-resource languages.

For citation: Eremeev M.A., Humonen I.S., Golyadkin M.Yu., Fitiskina A.A., Makarov I.A. Automatic Restoration of Birchbark Manuscripts Using Masked Language Modeling. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 6, part 1, pp. 177-192. DOI: 10.15514/ISPRAS-2026-38(6)-11.

Acknowledgements. The authors would like to express sincere gratitude to I. Filippov, D. Krylov, Y. Mikhalevskyi, D. Sitchinava, and O. Syrov for productive discussions and invaluable assistance. The research conducted by A. Fitiskina was supported by the Russian Science Foundation, grant no. 25-18-00862.

Автоматическое восстановление берестяных грамот методами маскированного языкового моделирования

¹ М.А. Еремеев, ORCID: 0009-0009-0546-5134 <maxim.a.eremeev@gmail.com>^{2,3} И.С. Хумонен, ORCID: 0000-0003-2621-507X <humonen@axxx.tech>^{2,3} М.Ю. Голядкин, ORCID: 0000-0002-0679-6981 <mgolyadkin@hse.ru>¹ А.А. Фитискина, ORCID: 0000-0002-5482-1416 <annaftiskina@gmail.com>^{2,3} И.А. Макаров, ORCID: 0000-0002-3308-8825 <iamakarov@hse.ru>¹ Национальный исследовательский университет «Высшая школа экономики»,
Россия, 101000, г. Москва, ул. Мясницкая, д. 20.² АXXX, Россия, 121170, г. Москва, Кутузовский пр-т, д. 32к1.³ Исследовательский центр доверенного искусственного интеллекта ИСП РАН,
Россия, 109004, г. Москва, ул. Александра Солженицына, д. 25.

Аннотация. Настоящая работа посвящена задаче автоматического восстановления лакун в новгородских берестяных грамотах, представляющей особую сложность из-за небольшого объёма доступного корпуса и отличительных черт древненовгородского диалекта. Мы представляем систематическое сравнение трёх BERT-подобных энкодеров – mBERT, BERTslav и ModernBERT – на уровне символов и токенов, в режимах без дообучения и после дообучения, на специально подготовленном корпусе древнерусских и церковнославянских текстов. После доменного дообучения ModernBERT демонстрирует лучший результат реконструкции, достигая точности по первому кандидату 30.72% на реальных лакунах и 91.40% на искусственно замаскированном тексте при предсказании по токенам. Однако на уровне символов дообученные энкодеры не превосходят простую символьную n-граммную модель, что указывает на структурное несоответствие между субсловным предобучением и посимвольным предсказанием. Также мы обучили линейные классификаторы на замороженных векторных представлениях для предсказания жанра и периода создания документа: полученные представления несут полезный сигнал для датировки, тогда как жанр оказывается преимущественно поверхностным орфографическим свойством, которое не хуже, а порой и лучше улавливает символьная модель на признаках TF-IDF.

Ключевые слова: берестяные грамоты; восстановление лакун; древненовгородский диалект; древнерусский; церковнославянский; маскированное языковое моделирование; малоресурсные языки.

Для цитирования: Еремеев М.А., Хумонен И.С., Голядкин М.Ю., Фитискина А.А., Макаров И.А. Автоматическое восстановление берестяных грамот методами маскированного языкового моделирования. Труды ИСП РАН, 2026, том 38, вып. 6, часть 1, с. 177-192 (на английском языке). DOI: 10.15514/ISPRAS-2026-38(6)-11.

Благодарности. Авторы выражают искреннюю благодарность И.И. Филиппову, Д.С. Крылову, Ю.П. Михальчевскому, Д.В. Сичинаве и О.А. Сырову за продуктивные обсуждения и неоценимую помощь. Исследование А.А. Фитискиной выполнено при поддержке Российского научного фонда, проект № 25-18-00862.

1. Introduction

Advances in computational methods have opened new possibilities for the study of historical texts. The present work focuses on the restoration of birchbark manuscripts (BBMs) – documents inscribed on pieces of birchbark that have been discovered, and continue to be discovered, in the vicinity of Veliky Novgorod and Staraya Russa, as well as at a number of other locations across former Rus¹.

Computational approaches to lacuna restoration have demonstrated notable results for ancient Greek and Latin epigraphy [1-3], texts in Akkadian [4], Old Chinese [5], Sanskrit [6], and other ancient languages [7-9], but Slavic historical manuscripts have not, to our knowledge, been addressed in this line of work.

The majority of BBMs are written in the Old Novgorodian dialect (OND), which differs substantially from standard Old Russian. The corpus of direct OND sources is considerably small. It includes, in addition to the BBMs themselves, a number of epigraphic inscriptions on objects, some marginal notes in ecclesiastical books, and a portion of Novgorodian parchment and paper charters in which dialectal features predominate over supradialectal norms. Indirect sources (oriented primarily towards Church Slavonic norms but containing Novgorodian features) encompass the remaining Novgorodian charters, the Novgorodian chronicles, and ecclesiastical books [10].

Many BBMs have survived only in fragmentary condition, making their restoration and interpretation a considerable challenge. Computational models can serve as valuable assistants in this process, providing additional evidence to support and complement the work of human scholars. The aim of the present work is to investigate computational approaches to lacuna restoration in BBMs, and to develop auxiliary tools for their temporal dating and categorical attribution. The small size of the corpus and the distinctive features of the dialect make this task particularly difficult. In a longer-term perspective, the text restoration module described here could be integrated into a broader manuscript processing pipeline, also comprising handwritten text recognition (HTR) and the segmentation of continuous text into words.

The main contributions of this work are: (1) a systematic comparison of multilingual encoders for BBMs' lacuna restoration, evaluated at both the character and token levels in zero-shot and fine-tuned settings; (2) auxiliary classifiers for genre and date prediction trained on frozen model embeddings; and (3) a publicly available code repository to support reproducibility and further research [11].

2. Related Work

2.1 Birchbark Manuscripts and Computational Approaches

Computational approaches to handwritten historical documents have grown in recent years, covering optical character recognition, script identification, handwriting analysis, and authorship attribution. Studies addressing Cyrillic sources, however, remain comparatively rare and tend to focus on documents from the 18th century onwards [12].

Nevertheless, there are some works on BBMs. Studies [13, 14] focus on the visual modality of documents. The former combines classical appearance modeling with convolutional networks and an ensemble of support vector machines for automatic dating, achieving a mean absolute error of 18.9 years. The latter addresses character recognition, training YOLOv10 on synthetic data with simulated birchbark noise and reaching a mean average precision of 0.89.

Study [15] analyzes chronological variation in the BBM corpus, relying on textual content. Texts are represented as sequences of 3-character shingles, and a FastText model is trained on this representation. According to the authors, given the non-standard orthography, dialectal variation, and fragmentary nature of the BBMs, classical character-level approaches tend to be competitive with embedding-based methods.

Study [16] shows that a domain-adapted BERT can identify the specific manuscript, century, and region of pre-modern Slavic texts. The model's vocabulary is extended with domain-specific words, and it is fine-tuned on a historical corpus using masked language modeling (MLM). The results indicate that domain adaptation improves accuracy even on relatively small historical Slavic datasets.

2.2 Computational Restoration of Ancient Texts

A number of works propose approaches to restoring ancient texts. In practice, such approaches can be used in a human-in-the-loop setting, where a scholar selects among a model's top candidates, which is why ranking metrics such as $\text{Hit}@k$, $k>1$ are meaningful.

Character n-gram models are used in several works on restoration as a non-neural baseline, as in [7] for Babylonian and [8] for Mycenaean Linear B.

Within the neural paradigm, a productive observation is that the MLM objective of BERT [17] is structurally identical to lacuna restoration: in both, the missing part is predicted from its surrounding context. This has motivated a body of work adapting BERT-like models to ancient and medieval corpora [4-6, 9].

Among the first, study [4] restored lacunae in transliterated Akkadian. As Akkadian signs may map to several subword tokens, the authors adapted a k-beam algorithm for multi-subtoken prediction, and a fine-tuned multilingual mBERT reached a $\text{Hit}@5$ of 89% – outperforming a model pre-trained from scratch on Akkadian. This suggests that broad multilingual pre-training can compensate for the absence of domain-specific pre-training. Similar framing was applied to Sanskrit [6], with a custom BPE tokenizer for its complex morphology ($\text{Hit}@10$ of 70%), and to classical Chinese [5], where a phrasal embedding was proposed for aggregating features from neighboring tokens to compensate for the absence of inter-word spaces ($\text{Hit}@5$ of about 76% on masking rate of 10%).

A complementary line trains task-specific models from scratch on epigraphic corpora, beginning with *Pythia* [1] for ancient Greek and continuing with *Ithaca* [2] and *Aeneas* [3] for Greek and Latin respectively, which integrate geographical and chronological attribution alongside restoration. Ithaca exploits its non-autoregressive nature with an easy-first decoding strategy: each step fills the highest-confidence position first, updating the context as positions are resolved. Aeneas, in turn, adopts a generative decoder that is able to restore gaps of unknown length. Both studies [2, 3] report that combining model predictions with expert judgment outperforms either alone.

The present work focuses on pre-trained bidirectional encoders as a natural starting point. Training a task-specific model from scratch on the assembled corpus is left for future work.

3. Data

3.1 Dataset Sources

The dataset used for model training comprises nearly 3.8M word tokens and was assembled from a variety of sources, subsequently filtered and normalized (see Section 3.3). It can broadly be divided into two parts: texts related to the Old Novgorodian dialect to varying degrees, and all remaining texts.

The open corpus of BBMs, hosted at the portal gramoty.ru [18] (accessed on 07.04.2026), currently contains 1,257 birchbark and 2 lead documents totalling around 24,000 words, with a vocabulary exceeding 4,100 lexical items. The BBMs are short texts, with a median length of 72 characters or 16 words per document. In the dataset, each document is provided in two forms – in the original writing without spaces (*scriptio continua*) and with word boundaries marked by spaces. The texts employ a system of editorial conventions: scribal errors are enclosed in curly braces {}, partial reconstructions of letters based on surviving traces are marked with square brackets [], full reconstructions supported only by contextual inference are indicated by round brackets (), a single hyphen denotes an unrestorable gap of one character, and an ellipsis denotes an unrestorable gap of unknown length.

The majority of BBMs are damaged to some degree: more than two thirds of the corpus (879 documents) contain gaps that are either irrecoverable or not yet reconstructed. Only 258 manuscripts, totalling over 5,000 words, are either fully intact or carry only partial reconstructions. The majority of documents originate from Novgorod (1,145 items) and the nearby town of Staraya Russa (58), with the remainder distributed across Torzhok, Smolensk, Pskov, Tver, Moscow, and several other locations.

As additional sources of the OND, we use Novgorod (34,300 words) and Pskov (1,300) parchment charters and the First Novgorodian Chronicle (nearly 34,000 words) from NKRYA [19]. These documents differ stylistically from most BBMs yet retain a number of OND features. The Old

Russian epigraphic dataset hosted at epigraphica.ru [20] (accessed on 07.04.2026) is also a structurally and partly thematically comparable source, as its editorial conventions follow the same system as the birchbark corpus.

Given the limited amount of data in the OND subset, we compiled a broader dataset drawing on documents in other Old Russian dialects, Church Slavonic, and a number of closely related languages, using both existing datasets (DIACU [21], TOROT [22]), and a biblical corpus of the Ponomar Project [23]) and independently crawled materials. A full breakdown of the sources and their word token counts after preprocessing (Section 3.3) is given in Table 1.

Table 1. Dataset sources, languages, and word token counts (after preprocessing).

Source	Language	Word tokens
Birchbark manuscripts [17]	Old Novgorodian (mostly)	19,045
Epigraphy [19]	Old Church Slavonic (mostly)	7,095
DIACU [20]	Old Church Slavonic; Church Slavonic (Old Russian, Middle Bulgarian, Serbian, Resava recensions); Middle Russian	1,588,323
TOROT [21]	Old Russian; Church Slavonic	603,047
Bible (Ponomar) [22]	Church Slavonic	682,430
Byliny [23]	Old Russian (11th–17th c.)	42,412
Pushkin House [24]	Old Russian	430,103
Military Statute (Part 2) [25]	Old Russian	49,787
NKRYA (historical) [18]	Old Russian (11th–18th c.), Old Novgorodian	327,315
Total		3,749,557

3.2 Birchbark Manuscripts’ Metadata

Each BBM in the dataset is accompanied by metadata. We use the conventional date (*uslovnaya data*) and the genre category as target variables for the auxiliary classifiers of Section 4.2. The conventional date, as defined by [27], is assigned by mapping each manuscript onto a segment of a chronological scale ranging from the second quarter of the 11th century to the mid-15th century. Each date therefore takes the form of an interval, reflecting the uncertainty of archaeological dating. Both the dating and category distributions are highly asymmetric: most manuscripts cluster around the 12th and 14th centuries, and the dominant category is private letters. For dating, we divide the timeline into nine 50-year bins spanning 1050–1499. For category, we merge the original fine-grained categories into four broader classes: *letters* (private letters), *records* (business records and official documents), *religious* (religious texts), and *other* (educational texts, folklore, varia, and uncertain content); manuscripts with several categories (17) are left without label. Even after consolidation, the distribution remains imbalanced: letters (696), records (330), other (171), religious (45) – this is addressed during training via weighted cross-entropy loss.

3.3 Text Preprocessing and Normalization

For BBMs and epigraphic texts, curly brackets {} are removed together with their content to prevent training on ungrammatical forms, while square [] and round brackets () are deleted but their content is kept as context. Unrestorable gaps are replaced with [GAP]. Line-break markers are removed, and hyphenated word fragments split across lines are rejoined. All text is lowercased. Of the available punctuation, three symbols are retained: the middle dot, the colon, and the cross – the last one used in certain periods to mark the beginning of a document. Some rare characters are deleted and some mapped to their invariants. For models that require it, an additional normalization step ensures compatibility with the model’s built-in tokenizer (Section 4.1).

Moreover, as very short documents tend to carry too little context for restoration, we retain only birchbark documents exceeding 25 characters, yielding 958 manuscripts with over 19,000 words. The same filter reduced the epigraphic corpus from 10,200 to about 7,100 words. Overlapping portions are removed to prevent leakage – for instance, the BBM corpus was excluded entirely from TOROT – and a substantial number of DIACU documents are discarded due to poor HTR quality in the original dataset, reducing it from about 2.1M to 1.59M tokens.

3.4 Masking Scheme and Set Construction

Data are split into training, validation, and test sets in an approximately 90/5/5 ratio. The test set is organized into two tracks: *Test A* for artificially generated gaps, and *Test B* for real lacunae based on editorial reconstructions. Test B consists of all birchbark and epigraphic documents with at least one editorial reconstruction; these are removed from the rest of the corpus before the split. The remaining deduplicated corpus is split randomly into training, validation, and Test A. Source-level sampling weights (Section 3.5) are then applied only to the training set.

Lacunae in training, validation and Test A are generated by artificially masking known characters. In Test B, by contrast, gaps are the concealed editorial reconstructions (in square or round brackets), and the model is evaluated on its ability to recover them. The two tracks distinguish the model’s general restoration capacity (Test A) from its practical utility for manuscript scholarship (Test B).

The 8% masking probability matches the empirically observed reconstruction density of the BBM corpus: in Test B, 7,130 of 96,606 characters (7.38%) are subject to editorial reconstruction, and including the gap tokens already present raises the combined ratio of masked and unreadable material to 8.54%. An 8% rate therefore approximates the proportion of damaged content in the target domain. Damage is simulated by masking contiguous spans. Span lengths follow a geometric distribution ($p=0.35$), yielding mostly short gaps with a few longer ones. Each position in a span is masked individually, so the number of mask tokens equals the gap length. The gap token, which marks genuinely unreadable positions, is never a mask target: the model observes it as context but is never trained to predict it.

3.5 Source Balancing

The assembled corpus is thematically skewed: religious texts, drawn largely from the substantial Bible corpus in Church Slavonic, are overrepresented relative to the target domain. Since the BBMs consist predominantly of everyday correspondence and legal records, training on the raw distribution would bias the models away from the registers most relevant to restoration. To mitigate this, all documents were divided into six thematic groups with context tags: DAILY, LEGAL, CHURCH, LIT, EPIC, and SCIENCE. This grouping is not intended to be exhaustive or precise – assignments follow the original characterization of each source, and multi-register sources are assigned to their most representative group; the tags serve only to balance the sampling distribution and do not constitute a linguistic classification.

Using these tags, source-level sampling weights were applied to the training split: documents from underrepresented categories are sampled multiple times per epoch, while overrepresented sources are downweighted (see Figure 1). The weights approximate the ordering $DAILY \approx LEGAL > CHURCH > LIT > EPIC > SCIENCE$, reflecting the thematic distribution of the target BBM domain.

4. Methods

We study lacuna restoration at two levels of prediction granularity. In token-level mode, the model predicts a missing position over its native subword vocabulary. In character-level mode, its output is constrained to single characters, aligning with the editorial markup of lacunae. Comparing the two lets us ask whether constraining a subword-pre-trained model to character-level output is

worthwhile, or whether it merely discards the model’s strengths. The alignment matters because WordPiece tokens need not respect lacuna boundaries. When a lacuna boundary falls inside a subword, the whole token must be masked, so the model loses access to characters outside the lacuna but within the same token – a subword boundary mismatch.

All evaluated architectures use bidirectional attention, so each masked position can be filled using context on both sides. Restoration is therefore performed by independent per-position prediction: each masked position is predicted in a single forward pass, without left-to-right generation and without conditioning on the model’s own predictions at other positions. This applies in both modes.

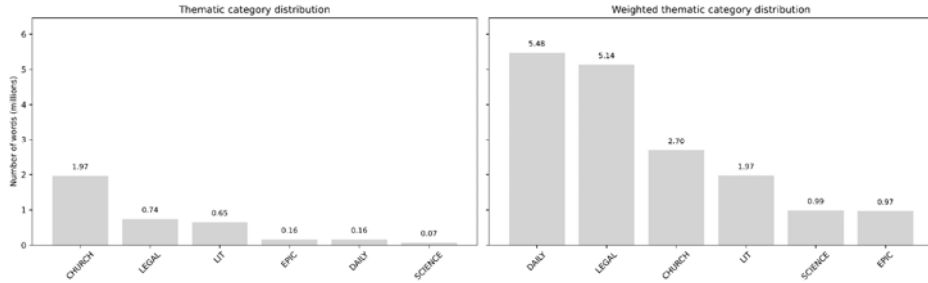


Fig. 1. Distribution of thematic categories in the dataset.

4.1 Text Restoration Models

Three pre-trained transformer encoders are selected for fine-tuning:

- mBERT [17] – a multilingual BERT model trained on 104 languages using a shared WordPiece vocabulary.
- BERTislav [28] – a ruBERT-based model pre-trained on a 10M-token corpus of Old Church Slavonic, Later Church Slavonic, Old Russian, Middle Russian, and Medieval Serbian texts.
- ModernBERT [29] – a bidirectional encoder with an extended context window and rotary positional embeddings.

mBERT is included because of its promising results on Akkadian lacuna restoration [4]. BERTislav is chosen because its pre-training domain partially overlaps with the target variety. ModernBERT represents recent architectural improvements to the bidirectional encoder paradigm and serves as a more capable general-purpose baseline. All three are evaluated both zero-shot and after fine-tuning on the assembled corpus with the MLM objective; their architectural characteristics are summarized in Table 2.

Table 2. Architectural comparison of the three fine-tuned models.

Model	Lay.	Heads	Params	Vocab	Tokeniz.	Max len.	Emb.
mBERT	12	12	~179M	119K	WordPiece	512	APE
BERTislav	12	12	178M	120K	WordPiece	512	APE
ModernBERT	22	12	149M	50K	byte-level BPE	8192	RoPE

Fine-tuning uses the standard MLM objective with a masking probability of 8%, a learning rate of 5×10^{-5} , and a maximum of 30 epochs. After each epoch we evaluate Hit@1 on the validation split; training stops early when this metric does not improve for three consecutive epochs (patience 3), and the checkpoint with the highest validation Hit@1 is retained. We train the two granularities as separate runs: the token-level run selects on validation token Hit@1, while the character-level run selects on character Hit@1, obtained by constraining the output to single-character tokens. The

character-level mode is thus not a separate training objective but the same subword MLM decoded under a single-character constraint.

Even after the preprocessing of Section 3.3, a number of letterforms outside the standard Unicode Cyrillic block fall outside the vocabularies of BERTislav and mBERT. For BERTislav, we apply the normalization script provided by its authors. For mBERT, we construct a custom normalizer that maps characters absent from the tokenizer’s vocabulary to their closest standard equivalents. ModernBERT, whose byte-level BPE tokenizer has no out-of-vocabulary characters, requires no additional normalization.

Decoding in character-level mode is constrained to tokens corresponding to one character of the target output space – Cyrillic letters, the three punctuation marks retained by normalization, and the inter-word space. However, this space is not equally reachable even within a model – WordPiece cannot emit spaces, ModernBERT cannot emit historical graphemes split into byte tokens; we note these structural limitations in Section 6. In token-level mode, each masked position is predicted over the full subword vocabulary without constraints.

For comparison, we additionally evaluate a bidirectional character n-gram language model with interpolated absolute discounting [30] as a baseline. A forward model estimates $P(c | \text{left context})$ and a backward model $P(c | \text{right context})$, using n-grams up to order five (four characters of context on each side), and each candidate character is scored by the product of the two, mirroring the bidirectional setting of the encoders. The model is trained on the same train split and evaluated with the identical character output space and metrics. We also report a unigram baseline relying only on global character frequency.

4.2 Auxiliary Classification

Building on the metadata of Section 3.2, we train two auxiliary classifiers on top of the frozen embeddings produced by each model, one for category and one for date. For each model and task we train a linear probe – a single linear layer preceded by layer normalization and dropout, applied to the [CLS] representation. Two inputs are evaluated: embeddings from the restored text (target) and from the masked text with lacunae replaced by [GAP] tokens (masked). Both probes are optimized with Adam ($\text{lr}=10^{-3}$) for 50 epochs, with the best checkpoint selected on validation by macro-F1 for category and by lowest mean distance for date.

Category classification is a four-class problem (letters, records, religious, other) trained with weighted cross-entropy loss to address class imbalance. Date classification follows the Ithaca approach [2]: rather than predicting a single bin, the probe outputs a distribution over nine 50-year bins. For each manuscript, the ground-truth target is constructed by distributing probability proportionally to each bin’s overlap with the date range – a manuscript dated 1200–1300, for instance, has equal mass on the 1200–1249 and 1250–1299 bins and zero elsewhere.

As a non-neural baseline, we use character n-gram TF-IDF features ($n = 2-5$, computed within word boundaries) with a linear classifier, on the same stratified split. For category, the features feed a logistic-regression classifier with balanced class weights. For date, a nine-class logistic-regression classifier is trained on the dominant date bin of each manuscript; at inference its predicted class distribution is taken as Q(1) and scored with the same distance-to-interval metric as the probe (see Section 4.3). The logistic-regression regularization strength is selected on the validation split, just as the probes select their checkpoint on validation. As trivial lower bounds we also report a majority-class baseline for category and a constant-mean predictor for date.

4.3 Evaluation Metrics

All models are evaluated on Test A and Test B with complementary metrics.

Hit@k measures whether the correct answer appears among the top-k predictions ($k \in \{1, 5, 20\}$), computed per character or per token:

$$\text{Hit}@k = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{\text{rank}_i \leq k\} \quad (1)$$

where N is the number of masked instances and rank_i is the rank of the original masked character or token.

Character error rate (CER) is the normalized Levenshtein distance between predicted and target strings; for token-level models we report the analogous *token error rate (TER)*.

Perplexity (PPL) measures overall language modeling quality; following standard practice for MLMs, we report pseudo-perplexity, the exponentiated mean cross-entropy over masked tokens.

For category classification, we report accuracy and macro-averaged F1 across four classes on a stratified 70/15/15 split.

For date prediction, the metric is the distance in years between the predicted mean year and the target interval:

$$\text{dist}(\hat{y}, [a, b]) = \max(0, a - \hat{y}, \hat{y} - b) \quad (2)$$

where $\hat{y} = \sum_{l=1}^9 m_l \cdot Q(l)$ is the predicted mean year, m_l is the midpoint of bin l , $Q(l)$ is the predicted probability for bin l , and $[a, b]$ is the ground-truth interval. The distance is zero when the predicted mean falls within the interval.

5. Results

5.1 Zero-shot Performance

All three models are first evaluated without fine-tuning. In character-level mode (Table 3), performance is uniformly weak, as expected when models pre-trained on subword tokens are constrained to single-character output. Scores are slightly higher on Test B than Test A, as the former has fewer inter-word spaces, which WordPiece models cannot recover, and more punctuation, which becomes a popular prediction under the character constraint. ModernBERT leads on Test B due to its better ability to predict reduced jers, which are more frequent in Test B compared to Test A (see Section 5.4).

Table 3. Zero-shot performance (characters), %.

Model	Test A			Test B		
	Hit@1	Hit@5	CER	Hit@1	Hit@5	CER
mBERT	0.82	4.06	96.20	4.98	22.14	94.76
BERTis Slav	1.16	5.04	93.98	4.74	23.92	94.15
ModernBERT	0.90	2.98	93.04	7.38	24.84	91.12

Token-level results (Table 4) are an order of magnitude stronger on Test A than in character-level mode, reflecting the alignment between the models' pre-trained prediction space and this setting. BERTis Slav leads on Test A showing the benefit of its domain pre-training. ModernBERT is the weakest on Test A by a wide margin (Hit@1 17.81%), though it matches the other models on Test B. On Test B scores drop for all models, more sharply for the WordPiece ones, consistent with the subword boundary mismatch (Section 4). The drop affects even ModernBERT, however, pointing to the intrinsic difficulty of real material: nearly half of the Test B lacunae (45%) sit adjacent to unrecoverable gaps in the surrounding text, and historical graphemes are far more frequent than in the artificially masked Test A.

5.2 Fine-tuned Performance

After domain fine-tuning, character-level models slightly improve but remain weak. ModernBERT is the strongest encoder, reaching Hit@1 of 3.00% on Test A and 16.54% on Test B. The non-neural baselines expose how poorly the encoders fit the character-level task: a bidirectional character n-gram model and even the context-free unigram substantially outperform the fine-tuned encoders on

Test A, and only ModernBERT surpasses them on Test B Hit@1, seemingly, as in the zero-shot setting, due to its stronger recovery of the reduced jers (see Table 5).

Table 4. Zero-shot performance (tokens), %.

Model	Test A			Test B		
	Hit@1	Hit@5	TER	Hit@1	Hit@5	TER
mBERT	42.23	59.07	52.04	12.19	21.77	84.55
BERTis Slav	45.67	64.34	51.44	13.75	24.97	84.87
ModernBERT	17.81	32.30	89.16	12.52	25.58	84.95

Table 5. Fine-tuned and baseline performance (characters), %.

Model	PPL	Test A			Test B		
		Hit@1	Hit@5	CER	Hit@1	Hit@5	CER
mBERT	4.05	1.10	4.80	93.16	6.40	27.62	91.71
BERTis Slav	10.05	1.36	5.18	90.37	4.20	21.26	94.89
ModernBERT	2.28	3.00	6.70	80.57	16.54	38.44	80.07
Char n-gram (n=5)		19.55	46.07	76.44	15.01	41.13	79.55
Unigram		17.70	44.09	83.56	6.73	34.36	96.04

In token-level mode (Table 6), ModernBERT leads on Test A (Hit@1 91.40%), with mBERT close behind (86.65%) and BERTis Slav lagging (75.61%). Hit@1 drops from over 75% on Test A to 20-31% on Test B for all models. On Test B the ranking compresses: mBERT and ModernBERT reach nearly identical Hit@1 (30.08% and 30.72%) and close Hit@5, but this parity masks different strengths: mBERT stronger on mid-length subwords, ModernBERT on short and long ones.

BERTis Slav, despite its domain-relevant pre-training, underperforms mBERT on Test B both in chars and tokens. A plausible explanation is that its pre-training on standard Old Russian and Church Slavonic biases it toward supradialectal orthographic norms, from which the OND diverges, whereas mBERT's more diffuse multilingual representations adapt more flexibly to the target variety.

Fig. 2 shows validation Hit@1 curves during fine-tuning. ModernBERT outperforms the other two from the first epoch in both modes; in character-level mode the models converge quickly and training stops early (BERTis Slav already at epoch 1, mBERT and ModernBERT at epoch 5), whereas token-level training is more stable and requires more epochs to converge.

Table 6. Fine-tuned performance (tokens), %.

Model	PPL	Test A			Test B		
		Hit@1	Hit@5	TER	Hit@1	Hit@5	TER
mBERT	2.96	86.65	94.80	12.80	30.08	48.22	67.95
BERTis Slav	5.84	75.61	85.58	20.89	20.22	33.27	77.74
ModernBERT	1.90	91.40	97.66	8.70	30.72	52.72	66.06

5.3 Auxiliary Classification

Tables 7-8 report probing results for genre category and conventional date, evaluated on both restored (target) and damaged (masked) text.

Token-level embeddings seem to yield better results than character-level ones. For category, BERTis Slav token achieves the highest macro-F1 on target (67.9%), with mBERT character close behind. For date, BERTis Slav token leads on mean distance both for target and masked text (40.1 and 45.3 years), while mBERT token achieves the best median (25.0 and 30.6 respectively). All models perform worse on masked than on target text, with the largest gap for ModernBERT (6.3% F1 drop

and 7.4-year increase in mean distance), indicating greater sensitivity to degradation despite its stronger restoration.

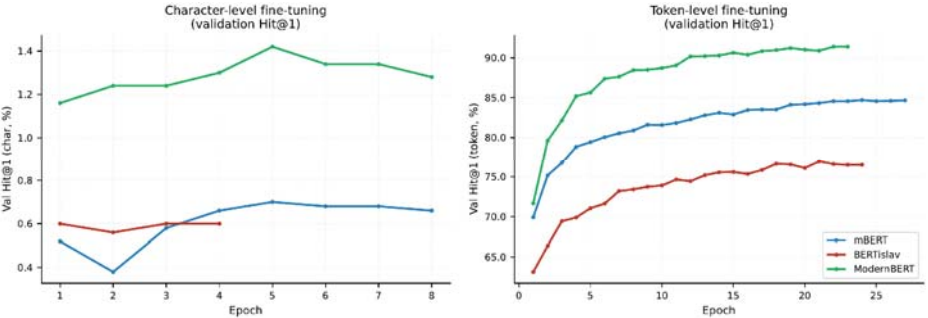


Fig. 2. Validation Hit@1 curves during fine-tuning in character-level (left) and token-level (right) modes.

Table 7. Category prediction performance, %.

Model	Mode	Target		Masked	
		Acc	F1	Acc	F1
BERTislav	char	76.5	67.5	74.3	64.0
	tok	78.6	67.9	74.9	62.1
mBERT	char	77.0	66.1	72.7	63.3
	tok	73.3	57.5	74.3	58.6
ModernBERT	char	71.1	55.1	65.8	52.1
	tok	78.6	62.5	72.7	56.2
TF-IDF + LogReg		76.5	61.7	78.6	64.3
Majority		55.6	17.9	55.6	17.9

Table 8. Date prediction performance (distance in years).

Model	Mode	Target		Masked	
		Mean	Median	Mean	Median
BERTislav	char	41.3	27.4	46.6	32.2
	tok	40.1	26.5	45.3	33.5
mBERT	char	45.6	34.3	47.7	35.6
	tok	40.4	25.0	45.4	30.6
ModernBERT	char	55.5	40.0	57.4	50.7
	tok	48.2	37.2	55.6	45.4
TF-IDF + LogReg		46.5	38.6	49.4	40.6
Constant (mean)		81.5	72.4	81.5	72.4

The character n-gram TF-IDF baseline is remarkably strong: it is competitive with the probes on clean target text (macro-F1 61.7%) and best on masked text, exceeding every encoder (64.3%). Genre is thus largely a surface-orthographic property that simple character features capture as well as – or better than – fine-tuned embeddings, which add little. For date, by contrast, the baseline trails the probes clearly (46.5 vs. 40.1 years), showing that conventional dating relies on more distributed contextual signal. Both tasks lie well above the trivial floors (category macro-F1 17.9%; date 81.5 years).

5.4 Error Analysis

A position-level analysis of the Test B predictions reveals different failure modes. The WordPiece encoders overwhelmingly default to punctuation – typically one-character tokens – accounting for 45.1% of BERTislav’s and 33.4% of mBERT’s character-level predictions, against 10.1% for ModernBERT and 1.1% for the n-gram. ModernBERT alone recovers targets the others miss – inter-word spaces (25.9%, versus 0% for both WordPiece models and 8.1% for the n-gram) and jers (36.9%, up from 11.6% zero-shot, versus at most 5% elsewhere). The jer result, however, reflects over-prediction of ѣ rather than a grasp of reduction: in 28% of the cases where ModernBERT outputs ѣ, the target is a consonant, space, or punctuation – positions that cannot correspond to a reduced vowel. The n-gram, by contrast, does well on frequent vowels and consonants but poorly on the jers (5.3%), which stay hard to predict from local context.

At the token level, mBERT’s accuracy stays roughly constant across target lengths, whereas BERTislav degrades on longer subwords (28% on single characters to 16% on five or more) and leaves more than half its targets outside the top-20, against 36.9% for mBERT and 25.4% for ModernBERT – a concrete account of its weaker ranking. Here ѣ appears as a frequent default for the WordPiece models, standing in for a wide range of true subwords.

6. Limitations

We note several limitations.

Source proportion. The birchbark and epigraphic components of the training corpus are small (approximately 26,000 word tokens combined), and the models rely heavily on related but stylistically distinct sources.

Boundary effects in token-level evaluation. When a lacuna boundary falls inside a subword, the whole token must be masked on Test B, so the model loses access to characters outside the bracketed span but within the same token. This systematically widens the lacunae and places token-level prediction of WordPiece models at a mild disadvantage.

Structural unreachability in character-level decoding. The character-level output space is not equally reachable across tokenizers. WordPiece models (mBERT, BERTislav) have no single token decoding to a space, so inter-word spaces are structurally unpredictable for them – their space-level accuracy on Test B is 0.00, although spaces make up about 10% of their masked positions. ModernBERT has the complementary blind spot: historical Cyrillic graphemes absent from its pre-training are split into multiple byte tokens, none decoding to a single character, and are therefore unreachable – though such characters account only for about 4% for Test A and 6.5% for Test B.

7. Conclusion

This work presents a comparison of fine-tuned multilingual encoders for automatic lacuna restoration in Old Novgorodian birchbark manuscripts, evaluated at both the character and token levels in zero-shot and fine-tuned settings on a purpose-built corpus of Old Russian and Old Church Slavonic texts. In general, token-level prediction yields substantially higher absolute accuracy than character-level prediction, since it matches the units the models were pre-trained on.

This advantage narrows on real lacunae (Test B), because lacuna boundaries fall inside subword tokens. However, real material is in general harder to process: it carries rare historical graphemes, and unrecoverable gaps in the surrounding context.

After domain fine-tuning, ModernBERT achieves the strongest restoration, reaching token-level Hit@1 of 30.72% on real editorial lacunae and 91.40% on artificially masked text, with the lowest perplexity in both modes. The character level tells a different story: constraining a subword model to single-character output does not exploit its alignment with the editorial markup but simply weakens the model. The encoders gain little from fine-tuning and are matched or outperformed by

a character n-gram baseline. This, together with the inability to emit certain characters, shows that the character constraint discards the models' subword strengths without a compensating gain.

Regarding the auxiliary classification tasks, a character n-gram TF-IDF baseline outperforms every encoder on damaged text, indicating that genre is largely a surface-orthographic property for which the embeddings add little. In date prediction, however, the baseline trails the probes clearly, showing that conventional dating draws on deeper contextual cues that the embeddings capture but surface statistics do not. Among the encoders, BERTslav attains the best genre macro-F1 and mean date error.

The lacuna restoration module described here is one component of a broader manuscript processing pipeline that could be extended to other historical Cyrillic corpora. Future directions include further hyperparameter search, varying the masking rate and span-length distribution, incorporating document-level thematic tags as conditioning tokens, and expanding the training corpus.

A more impactful direction is to train a task-specific model from scratch, following [2] and [3]: a character-level encoder with non-sequential easy-first decoding, or a decoder with rotary position embeddings and a length-prediction head that would enable restoration in the more realistic setting where lacuna boundaries are uncertain. Such a model would also avoid the structural limitations discussed in Section 6.

References

- [1]. Assael Y., Sommerschild T., Prag J. Restoring ancient text using deep learning: a case study on Greek epigraphy. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 6368-6375. DOI: 10.18653/v1/D19-1668.
- [2]. Assael Y., Sommerschild T., Shillingford B., et al. Restoring and attributing ancient texts using deep neural networks. *Nature*, 2022, vol. 603, pp. 280-283. DOI: 10.1038/s41586-022-04448-z.
- [3]. Assael Y., Sommerschild T., Cooley A., et al. Contextualizing ancient texts with generative neural networks. *Nature*, 2025, vol. 645, pp. 141-147. DOI: 10.1038/s41586-025-09292-5.
- [4]. Lazar K., Saret B., Yehudai A., Horowitz W., Wasserman N., Stanovsky G. Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach. In Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing, ACL, 2021, pp. 4682-4691. DOI: 10.18653/v1/2021.emnlp-main.384.
- [5]. Gu Z., Guan Y., Zhang S. A text restoration model for ancient texts based on pre-trained language model RoBERTa. In Proc. of the 2024 4th International Conference on Internet of Things and Machine Learning (IoTML '24), ACM, 2024, pp. 96-102. DOI: 10.1145/3697467.3697610.
- [6]. Baidya A., Rizk R., Santosh KC. Masked Language Model for Ancient Sanskrit Text Restoration. In *Computational Intelligence in Pattern Recognition. CIPR 2024. Lecture Notes in Networks and Systems*, vol. 1153, Springer, Singapore, 2025, pp. 641-655. DOI: 10.1007/978-981-97-8093-8_43.
- [7]. Fetaya E., Lifshitz Y., Aaron E., Gordin S. Restoration of fragmentary Babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 2020, vol. 117, no. 37, pp. 22743-22751. DOI: 10.1073/pnas.2003794117.
- [8]. Papavassileiou K., Kosmopoulos D.I., Owens G. A Generative Model for the Mycenaean Linear B Script and Its Application in Infilling Text from Ancient Tablets. *Journal on Computing and Cultural Heritage*, 2023, vol. 16, no. 3. DOI: 10.1145/3593431.
- [9]. Fono N., Moshayof H., Karol E., Assraf I., Last M. Embible: Reconstruction of Ancient Hebrew and Aramaic Texts Using Transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, St. Julian's, Malta, ACL, 2024, pp. 846-852. DOI: 10.18653/v1/2024.findings-eacl.56.
- [10]. Зализняк А.А. Древненовгородский диалект. 2-е изд., Москва, Языки славянской культуры, 2004, 872 с. / Zaliznyak A.A. Old Novgorodian dialect. 2nd ed., Moscow, Yazyki slavyanskoy kultury, 2004, 872 p. (in Russian).
- [11]. Birchbark Restoration code repository. Available at: <https://github.com/Maximkou1/birchbark-restoration/tree/v1.0-paper>.
- [12]. Diao X., Bo R., Xiao Y., Shi L., Zhou Z., Xu H., Li C., Tang X., Poesio M., John C.M., Shi D. Ancient script image recognition and processing: A review. 2025. DOI: 10.48550/arXiv.2506.19208.

- [13]. Sidorov K.A. Estimating the Age of Birch Bark Manuscripts using Computational Paleography. *Programming and Computer Software*, 2018, vol. 44, no. 4, pp. 286-294. DOI: 10.1134/S0361768818040114.
- [14]. Malashin I.P., Tynchenko V.S., Gantimurov A.P., Nelyub V.A., Borodulin A.S. Recognition of Handwritten Characters in Birch-Bark Manuscripts via Object Detection. *IEEE Access*, 2025, vol. 13, pp. 120563-120577. DOI: 10.1109/ACCESS.2025.3586910.
- [15]. Afanasev I., Lyashevskaya O. The application of corpus-based language distance measurement to the diatopic variation study (on the material of the Old Novgorodian birchbark letters). In Proc. of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025), Tallinn, Estonia, 2025, pp. 153-164.
- [16]. Lendvai P., Reichel U., Jouravel A., Rabus A., Renje E. Domain-Adapting BERT for Attributing Manuscript, Century and Region in Pre-Modern Slavic Texts. In Proc. of the 4th Workshop on Computational Approaches to Historical Language Change, Singapore, ACL, 2023, pp. 15-21. DOI: 10.18653/v1/2023.lchange-1.2.
- [17]. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. DOI: 10.48550/arXiv.1810.04805.
- [18]. Gramoty.ru: Corpus of Old Russian Birchbark Manuscripts. Available at: <https://gramoty.ru/birchbark/>, accessed 07.04.2026.
- [19]. NKRYA: Russian National Corpus (Historical Subcorpus). Available at: <https://ruscorpora.ru>, accessed 07.04.2026.
- [20]. Epigraphica.ru: Corpus of Old Russian Epigraphy. Available at: <https://epigraphica.ru/epigraphy/>, accessed 07.04.2026.
- [21]. Cassese M., Puccetti G., Napolitano M., Esuli A. DIACU: a dataset for the diachronic analysis of Church Slavonic. In Proc. of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), Vienna, Austria, ACL, 2025, pp. 101-107. DOI: 10.18653/v1/2025.bsnlp-1.12.
- [22]. Eckhoff H.M., Berdicevskis A. TOROT: Tromso Old Russian and Old Church Slavonic Treebank. Available at: <https://torotreebank.github.io>, accessed 07.04.2026.
- [23]. Andreev A., et al. Ponomar Project: Church Slavonic Bible (Elizabeth Recension). Available at: <https://github.com/typiconman/ponomar>, accessed 07.04.2026.
- [24]. Тихонравов Н.С., Миллер В.Ф. (ред.) Русские былины старой и новой записи. Москва, 1894. / Tikhonravov N.S., Miller V.F. (eds.) Russian byliny of old and new records. Moscow, 1894 (in Russian).
- [25]. Pushkin House: Manuscripts of Old Russian Literature. Institute of Russian Literature (Pushkin House), RAS. Available at: <http://lib.pushkinskijdom.ru>, accessed 07.04.2026.
- [26]. Радишевский О.М. Устав ратных, пушечных и других дел, касающихся до воинской науки. Часть 2. Санкт-Петербург, 1781. / Radishevsky O.M. Statute of military, artillery and other matters. Part 2. Saint Petersburg, 1781 (in Russian).
- [27]. Янин В.Л., Зализняк А.А. Новгородские грамоты на бересте (из раскопок 1990-1996 гг.). Палеография берестяных грамот и их внестратиграфическое датирование. Москва: Русские словари, 2000, 430 с. / Yanin V.L., Zaliznyak A.A. Novgorod letters on birch bark (from the 1990-1996 excavations). Paleography of birchbark documents and their extra-stratigraphic dating. Moscow: Russkiye slovari, 2000, 430 p. (in Russian).
- [28]. Pedrazzini N. BERTslav. Available at: <https://huggingface.co/npedrazzini/BERTslav>, accessed 07.04.2026.
- [29]. Warner B., Chaffin A., Clavie B., et al. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria, ACL, 2024, pp. 2526-2547. DOI: 10.48550/arXiv.2412.13663.
- [30]. Ney H., Essen U., Kneser R. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 1994, vol. 8, no. 1, pp. 1-38. DOI: 10.1006/csla.1994.1001.

Информация об авторах / Information about authors

Максим Алексеевич ЕРЕМЕЕВ – бакалавр компьютерной лингвистики НИУ ВШЭ. Сфера научных интересов: обработка естественного языка, digital humanities, малоресурсные языки.

Maxim Alekseevich EREMEEV – bachelor's degree in computational linguistics from HSE University (National Research University "Higher School of Economics"), Russia. Research interests: natural language processing, digital humanities, low-resource languages.

Иннокентий Сергеевич ХУМОНЕН – магистр компьютерной лингвистики НИУ ВШЭ. Сфера научных интересов: обработка естественного языка, digital humanities, малоресурсные и древние языки.

Innokentiy Sergeevich HUMONEN – master's degree in computational linguistics from HSE University, Moscow. Research interests: natural language processing and digital humanities, with a focus on low-resource and ancient languages.

Максим Юрьевич ГОЛЯДКИН – магистр прикладной математики и информатики НИУ ВШЭ, где в настоящее время работает над диссертацией по направлению «компьютерные науки». Сфера научных интересов: применение нейронных сетей в промышленном ИИ.

Maksim Yurievich GOLYADKIN – master's degree in applied mathematics and informatics from HSE University, Moscow, where he is currently pursuing the Ph.D. degree in computer science. Research interests: neural network applications to industrial AI.

Анна Александровна ФИТИСКИНА – аспирантка Школы лингвистики НИУ ВШЭ, младший научный сотрудник ИРЯ РАН, Москва. Сфера научных интересов: исторический синтаксис русского языка, древнерусские берестяные грамоты, древнерусская эпиграфика.

Anna Aleksandrovna FITISKINA – postgraduate student at the School of Linguistics, HSE University, and Junior Research Fellow at the Vinogradov Russian Language Institute of the RAS, Moscow. Research interests: Russian historical syntax, Old Russian birchbark manuscripts, Old Russian epigraphy.

Илья Андреевич МАКАРОВ – кандидат компьютерных наук (Университет Любляны), специалист по математике (МГУ имени М.В. Ломоносова). Сфера научных интересов: машинное обучение, представление графов, компьютерное зрение.

Ilya Andreevich MAKAROV – Cand. Sci. (Comp/ Sci.) from the University of Ljubljana. Specialist degree in mathematics from Lomonosov Moscow State University. Research interests: machine learning, graph representation learning, computer vision.