UDC 004.912:81'322

A MATHEMATICAL MODEL FOR THE AUTOMATIC EXTRACTION OF IDIOMS IN TURKIC LANGUAGES

[1]***Nazirova E.Sh.***, [2*]***Madetbayeva N.B.***, [1]***Mirzayeva F.Sh.***, [3]***Islomova M.B.***
[*]madetbaeva@gmail.com
[1]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi,
108, Amir Temur Avenue, Tashkent, 100084 Uzbekistan;
[2]Digital Technologies and Artificial Intelligence Development Research Institute,
17A, Buz-2, Tashkent, 100125 Uzbekistan;
[3]Inha University in Tashkent,
9, Ziyolilar Street, Tashkent, 100170 Uzbekistan

This paper presents a complete mathematical model for the automatic extraction of phraseological units (idioms) from texts in Turkic languages, with a focus on Uzbek. The task is formulated as a binary classification problem and is solved by a three-stage pipeline. At the first stage, a set of idiom candidates is generated by a linguistic filter based on verb-headed grammatical templates. At the second stage, each candidate is mapped to a feature vector that combines statistical association (pointwise mutual information), syntactic fixedness (a flexibility index), semantic non-compositionality, and template-lexical indicators. At the third stage, a decision function is trained using four classification algorithms, namely Naive Bayes, logistic regression, support vector machines (SVM), and gradient boosting. The model was evaluated on a combined Uzbek corpus of 1.07 million lemma-tokens, supported by a phraseological ontology of 1,362 units and a manually annotated, stratified gold-standard set of 400 sentences. The best overall result was achieved by the SVM ($F_1 = 0.795$; recall $= 0.906$), followed closely by gradient boosting ($F_1 = 0.791$), whereas the lexical string-matching baseline proved nearly ineffective ($F_1 = 0.100$) and ontology-based matching provided the highest precision ($P = 0.854$). Stratified analysis showed that the main advantage of the model lies in recognizing morphologically transformed forms of idioms, where the learning-based model reached a recall of 0.919 against 0.041 for the lexical baseline. The role of each method within the processing pipeline and its most effective area of application are discussed.

Keywords: phraseological unit, idiom detection, binary classification, pointwise mutual information, syntactic fixedness, Uzbek language

Citation: Nazirova E.Sh., Madetbayeva N.B., Mirzayeva F.Sh., Islomova M.B. 2026. A mathematical model for the automatic extraction of idioms in Turkic languages. *Problems of Computational and Applied Mathematics.* 4(74): 152-170.

DOI: 10.71310/pcam.4_74.2026.10

1 Introduction

Bringing the Uzbek language fully into the world of digital technologies has become one of the priorities of the national language policy. Decree No. PF-5850 of the President of the Republic of Uzbekistan "On measures to radically increase the prestige and status of the Uzbek language as the state language", adopted on October 21, 2019, set the task of integrating the state language with modern technologies. A year later, Decree No. PF-6084 of October 20, 2020 approved the concept for improving the language policy for 2020–2030. Implementing these decisions requires expanding the corpus resources of the Uzbek and Turkish languages and building, on their basis, translation tools that

convey meaning in full. The most difficult test on this path is posed by phraseological units (idioms), since the quality of translation largely comes down to whether idioms are correctly recognized in the text.

The choice of the Uzbek–Turkish language pair is not accidental. The two kindred languages are close in vocabulary and grammatical structure, and the literary and cultural ties between the two nations keep expanding: fiction, scholarly and journalistic texts, and television series are translated on a regular basis. It is in this flow that phraseology suffers the greatest losses, because an idiom is a condensed expression of national thought and culture and does not yield to word-for-word rendering. Modern lexicography and language teaching likewise rely not on bare lists of idioms but on their use in living text. A tool that detects idioms automatically therefore supplies Uzbek–Turkish phraseological dictionaries with authentic examples, shows learners an idiom in its natural context, and steadily enriches the phraseological layer of corpora.

A phraseological unit (idiom) is a fixed expression whose meaning does not follow from the sum of the meanings of its component words; it is precisely this property that makes idioms one of the most difficult objects in automatic text processing [1–4]. For example, the idiom *bosh qotirmoq* means ‘to think long and hard’, and a system that translates it component by component produces a meaningless rendering along the lines of “to freeze the head”. By contrast, *ovqat pishirmoq* ‘to cook food’ is a free phrase whose meaning is assembled directly from its components. Similarly, *ko‘z yummoq* ‘to turn a blind eye’ is an idiom, whereas *eshik yopmoq* ‘to close the door’ is a free phrase. All four expressions share the same grammatical pattern, a NOUN + VERB sequence. Hence an idiom cannot be told apart from a free phrase by surface grammatical cues alone; one has to measure the statistical and semantic properties of each expression and assign it to the class “idiom” or “free phrase”.

Several lines of research have taken shape in the world literature. The statistical line relies on association measures that quantify the cohesion of components, above all pointwise mutual information [5, 6]; comprehensive analyses of such measures are given in [7, 8], and the C-value measure was proposed for multi-component units [9]. The syntactic line established the resistance of idioms to transformations, that is, their syntactic fixedness, as an informative feature [10], and dedicated evaluation datasets were built on this basis [11]. The semantic line measures non-compositionality of meaning directly in distributional vector spaces [12–14]; word- and sentence-level embeddings [15–17], including the multilingual LaBSE model [18], serve as key instruments here. Substantial results have recently been obtained with neural models for idiomaticity detection [19, 20], and the international PARSEME initiative ran shared tasks on identifying verbal multiword expressions in more than a dozen languages [21, 22], with Turkish material also represented [23]. At the same time, these approaches were developed mainly on inflectional and analytic languages. The agglutinative morphology of Turkic languages [24, 25] and the computational resources being created for them [26–28] call for rebuilding the existing methods around a morphological analysis stage and national phraseological resources. The theoretical foundations of the typology of phraseological units are laid in [29, 30], and the scholarly description of Uzbek phraseology is given in [31, 32].

Corpora of millions of words cannot be inspected manually, so entrusting this work to machine learning and deep learning technologies has become a demand of the time. Yet any trainable model works reliably only on a solid mathematical foundation: the problem must be stated formally, the essential properties of idioms must be expressed as

measurable quantities, and the results must be assessed by quantitative criteria. Only once this fundamental basis is in place do learning algorithms deliver the expected effect.

The aim of this paper is to give a rigorous mathematical statement of the idiom extraction problem for Turkic languages, primarily Uzbek, to combine statistical measures and machine learning methods proven in computational linguistics into a single three-stage pipeline adapted to Uzbek phraseological material, and to evaluate the resulting model quantitatively on a real corpus. The problem is formulated as binary classification; each stage is expressed by separate formulas whose constituent parts are explained in detail; a system of verb-headed grammatical templates and four groups of features are substantiated for Uzbek idioms. Four classification algorithms are compared with each other and with two baselines on a combined corpus of more than 1.07 million words and on a manually verified gold-standard set of 400 sentences, and the most effective place of each method in the pipeline is identified.

The paper is organized as follows. Section 2 gives the mathematical statement of the problem, Section 3 describes the three stages of the model, Section 4 presents the evaluation methodology, Section 5 works through a numerical example, Section 6 describes the experimental data and the overall flowchart of the model, Section 7 discusses the results and the role of each method, and Section 8 concludes.

2 Mathematical statement of the problem

The source text D is viewed as a sequence of tokens. From it, the set C of all candidates is formed out of contiguous words. Candidates are word combinations that could potentially be idioms. The task is to assign every candidate to the class "idiom" or "free phrase". Formally, this is written as the mapping

$$F: \; c_i \mapsto y_i \in \{0, 1\}, \qquad i = 1, 2, \ldots, m, \tag{1}$$

where:

- c_i is the ith element of the candidate set $C = \{c_1, c_2, \ldots, c_m\}$ extracted from the text, that is, a combination of 2–5 contiguous words;
- m is the total number of candidates;
- y_i is the class label of the candidate c_i: $y_i = 1$ means "idiom" and $y_i = 0$ means "free phrase";
- F is the mapping that assigns each candidate its class, that is, the final objective of the whole model.

Here the set C contains all candidates that can be extracted from the text. Uzbek idioms mostly consist of two to four words; in our data such idioms account for 95 % of the total, while the longest ones reach five words, for example *olami fanodan olami baqoga o'tmoq* 'to pass from the mortal world to eternity'.

The mapping F is not constructed directly. It consists of three consecutive stages: a linguistic filter → vectorization → a classifier. First the set of idiom candidates is formed, then feature values are computed for every candidate, and finally the classifier is trained and applied.

3 The idiom extraction model

3.1 Stage one: forming the candidate set

First, only those candidates that can be idioms on grammatical grounds are retained in C. To this end a filter function $f(c)$ is introduced, and the filtered set is defined as

$$C_{\text{cand}} = \{\, c_i \in C \mid f(c_i) = 1 \,\}, \tag{2}$$

where:

- C_{cand} is the set of candidates that passed the filter;
- $f(c_i)$ is the value of the filter function for the candidate c_i (defined in Eq. (3));
- the symbol "|" reads "such that";
- $C_{\text{cand}} \subseteq C$: the filter adds no new elements, it only removes non-matching ones.

The value of the filter function is determined by whether the part-of-speech sequence of the candidate matches the idiom templates:

$$f(c_i) = \begin{cases} 1, & \text{POS}(c_i) \in T_{\text{id}}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where:

- $\text{POS}(c_i)$ is the part-of-speech sequence of the candidate c_i (for instance, NOUN + VERB for *bosh qotirmoq*);
- T_{id} is the set of grammatical templates characteristic of idioms (Table 1);
- $f(c_i) = 1$ means that a template matched and the candidate is retained;
- $f(c_i) = 0$ means that no template matched and the candidate is discarded.

The principal templates for Uzbek idioms, generalized from the material of the phraseological dictionary [32], are given in Table 1.

Table 1 Principal grammatical templates of Uzbek idioms.

№	Template (pattern)	Examples
1	NOUN + VERB	*bosh qotirmoq*, *til topishmoq*, *ko'z yummoq*
2	NOUN (possessive) + VERB	*boshi qotmoq*, *ko'ngli cho'kmoq*, *kapalagi uchmoq*
3	NOUN (case-marked) + VERB	*ko'zga ilmoq*, *esdan chiqmoq*, *qo'ldan bermoq*
4	NOUN + NOUN + VERB	*og'ziga tolqon solmoq*, *ko'ziga cho'p solmoq*, *boshiga ish tushmoq*
5	NOUN (possessive) + ADJ	*qo'li ochiq*, *tili uzun*, *peshonasi sho'r*
6	ADJ + NOUN or NOUN + NOUN	*shirin so'z*, *ochiq ko'ngil*, *ko'ngil ko'zi*
7	Fixed adverbial patterns	*bir og'izdan*, *ko'z ochib yumguncha*, *ipidan ignasigacha*
8	Sentence-shaped idioms	*og'zi qulog'ida*, *ishtahasi karnay*

The templates are written not as bare part-of-speech sequences but together with morphological markers, namely possessive and case affixes, because in many idioms these affixes have become fossilized. For instance, *boshi qotdi* is used precisely with the third-person possessive affix, and *esdan chiqmoq* precisely with the ablative case. A template is a necessary but not a sufficient condition, since numerous free phrases also fit the NOUN + VERB pattern (*ovqat pishirmoq*, *eshik yopmoq*). The final decision is therefore deferred to the subsequent stages.

3.2 Stage two: mapping a candidate to a feature vector

Every candidate $c \in C_{\text{cand}}$ that passed the filter is mapped to a numerical feature vector, so the vector is a function of c, $\mathbf{x} = \mathbf{x}(c)$:

$$\mathbf{x}(c) = \big(x_1(c),\, x_2(c),\, \ldots,\, x_n(c)\big)^{\top} \in \mathbb{R}^n, \tag{4}$$

where:

- c is a candidate that passed the filter;
- $\mathbf{x}(c)$ is the feature vector collecting all measurable properties of c;
- $x_j(c)$ is the numerical value of the jth feature computed for c;
- n is the number of features (the dimension of the vector);
- $(\cdot)^{\top}$ is the transposition sign, that is, $\mathbf{x}(c)$ is written as a column vector;
- $\mathbb{R}^n$ is the n-dimensional real space; every candidate becomes a point of this space.

The concrete components of the vector are two binary features, $x_{\text{templ}}(c)$ (the template type) and $x_{\text{som}}(c)$ (the presence of a body-part term), and three continuous features, $x_{\text{PMI}}(c)$, $x_{\text{flex}}(c)$, and $x_{\text{comp}}(c)$. Thus

$$\mathbf{x}(c) = \big(x_{\text{templ}}(c),\, x_{\text{som}}(c),\, x_{\text{PMI}}(c),\, x_{\text{flex}}(c),\, x_{\text{comp}}(c)\big)^{\top}.$$

The binary features are determined not by formulas but by a logical condition, whereas the three continuous features are computed by Eqs. (5)–(7). Table 2 summarizes the feature space.

Table 2 Composition of the feature space.

Feature	Type	What it measures	Signal in favour of an idiom	Defined by
x_{templ}	binary	grammatical pattern type (Table 1)	verb-headed, fossilized patterns	Eq. (3)
x_{som}	binary	presence of a body-part term	value equal to 1	Sect. 3.2
x_{PMI}	continuous	statistical cohesion of components	high positive value	Eq. (5)
x_{flex}	continuous, $[0;1]$	syntactic flexibility	value close to zero	Eq. (6)
x_{comp}	continuous, $[-1;1]$	compositionality of meaning	low value	Eq. (7)

Template and lexical features. The feature x_{templ} encodes which specific pattern of Table 1 the candidate matches; a separate binary feature is allocated to each pattern, because the patterns are not equally likely to yield idioms. The feature x_{som} is a binary indicator of whether the candidate contains a body-part term (*bosh* 'head', *ko'z* 'eye', *qo'l* 'hand', *til* 'tongue', *yurak* 'heart', *ko'ngil* 'soul', *og'iz* 'mouth', *oyoq* 'foot', *quloq* 'ear'). Somatic idioms form one of the largest groups of Uzbek phraseology [32], so this simple indicator provides the model with a strong signal.

Statistical cohesion. The cohesion of the components of a two-component candidate $c = (u, v)$ is measured by pointwise mutual information [5]:

$$x_{\text{PMI}}(c) = \log_2 \frac{P(u,v)}{P(u)\,P(v)}, \qquad c = (u,v), \tag{5}$$

where:

- $c = (u, v)$ is a two-component candidate; u and v are its component words (for example, *boshi* and *qotdi*);
- $P(u, v)$ is the probability that the candidate c, that is, the words u and v, co-occur side by side in the corpus;
- $P(u)$ and $P(v)$ are the probabilities of each component occurring in the corpus on its own;
- $\log_2$ is the binary logarithm; for independent words $x_{\mathrm{PMI}}(c) \approx 0$, whereas for fixed expressions $x_{\mathrm{PMI}}(c) > 0$ and large.

For candidates with more than two components, a generalized variant of PMI [7] or a C-value-type measure [9] is applied, since these also account for the nesting of an expression inside larger expressions.

Syntactic fixedness. An idiom preserves its form rigidly and resists transformations [10]. The degree to which a candidate preserves its form is measured by the flexibility index

$$x_{\mathrm{flex}}(c) = 1 - \frac{N_{\mathrm{can}}(c)}{N(c)}, \tag{6}$$

where:

- c is the candidate under consideration;
- $N(c)$ is the total number of occurrences of the candidate in the corpus;
- $N_{\mathrm{can}}(c)$ is the number of those occurrences that appear in the canonical (dictionary) form;
- $N_{\mathrm{can}}(c)/N(c)$ is the share of fixed occurrences;
- $x_{\mathrm{flex}}(c)$ is its complement to one, that is, flexibility: $x_{\mathrm{flex}}(c) \approx 0$ indicates a fixed form (an idiom), while $x_{\mathrm{flex}}(c) \to 1$ indicates a freely varying form.

This feature also reveals the internal gradation of the idiom class. The average flexibility measured on a 4,408-row syntactic fixedness corpus built within this study confirms the gradient of Vinogradov's typology [29]: phraseological fusions (0.202) < phraseological unities (0.246) < phraseological combinations (0.290).

Semantic non-compositionality. The failure of the meaning to be assembled from the components is measured in a vector space [13]. The vector of the whole expression is compared with the mean of the component vectors:

$$x_{\mathrm{comp}}(c) = \cos\big(\mathbf{e}(c),\, \bar{\mathbf{e}}\big) = \frac{\langle \mathbf{e}(c),\, \bar{\mathbf{e}} \rangle}{\|\mathbf{e}(c)\| \cdot \|\bar{\mathbf{e}}\|}, \tag{7}$$

where:

- $c = (t_1, t_2, \ldots, t_r)$ is a candidate with r components, t_j being its jth component word;
- $\mathbf{e}(c)$ is the embedding of the whole expression obtained with a multilingual transformer model (LaBSE [18]);
- $\bar{\mathbf{e}} = \dfrac{1}{r} \sum_{j=1}^{r} \mathbf{e}(t_j)$ is the mean of the component word embeddings;
- $\langle \cdot, \cdot \rangle$ is the inner product and $\| \cdot \|$ is the vector norm;
- cos is the cosine similarity between the two vectors: $x_{\mathrm{comp}}(c) \to 1$ indicates a compositional meaning (a free phrase), whereas a low $x_{\mathrm{comp}}(c)$ indicates a non-compositional meaning (an idiom).

For example, for *ovqat pishirmoq* the holistic vector lies very close to the mean of the components, whereas for *bosh qotirmoq* the holistic meaning ('to think') moves away from the components and the cosine drops sharply. In this way an abstract notion such as idiomaticity turns into a directly measurable geometric quantity.

3.3 Stage three: training and applying the classifier

The probability $P(y = 1 \mid \mathbf{x})$ that a candidate is an idiom is now estimated from the feature vector $\mathbf{x}$. For this purpose a classifier is trained on a labelled training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$. Four standard algorithms of statistical learning [33–36] are presented in turn below.

Naive Bayes. Naive Bayes is a generative probabilistic classifier built on Bayes' theorem and on the assumption of conditional independence of the features within a class. It estimates the posterior probability of each class and selects the more probable one. The posterior probability is computed as

$$P(y = 1 \mid \mathbf{x}) = \frac{P(y = 1) \prod_{j=1}^{n} P(x_j \mid y = 1)}{P(\mathbf{x})}, \tag{8}$$

where:

- $P(y = 1 \mid \mathbf{x})$ is the posterior probability that the candidate is an idiom given $\mathbf{x}$ (the sought answer);
- $P(y = 1)$ is the prior probability, the overall share of idioms in the training set;
- $P(x_j \mid y = 1)$ is the probability of the jth feature taking its value in the class "idiom";
- $\prod$ denotes the product over the n features; it is here that the features are assumed mutually independent;
- $P(\mathbf{x})$ is a normalizing factor that cancels out when the two classes are compared.

The decision rule is: if $P(y = 1 \mid \mathbf{x}) > P(y = 0 \mid \mathbf{x})$, the candidate is recognized as an idiom. Naive Bayes is a fast and simple baseline, but the independence assumption (PMI and syntactic fixedness, for instance, are correlated) does not hold in reality, so it may accept strong collocations as idioms and yield low precision.

Logistic regression. Logistic regression is a discriminative linear classifier that converts a linear combination of the features into a probability in $[0; 1]$ through the sigmoid function. The probability of a candidate being an idiom is modelled as

$$P(y = 1 \mid \mathbf{x}) = \sigma(z) = \frac{1}{1 + e^{-z}}, \qquad z = \mathbf{w}^{\top}\mathbf{x} + b, \tag{9}$$

where:

- z is the linear combination of the features, the overall score of the candidate;
- $\mathbf{w} = (w_1, w_2, \ldots, w_n)^{\top}$ is the weight vector; w_j is the weight of the jth feature;
- $\mathbf{w}^{\top}\mathbf{x} = w_1x_1 + w_2x_2 + \cdots + w_nx_n$ is the weighted sum;
- b is the bias (intercept), which shifts the decision boundary away from the origin;
- $\sigma(z)$ is the sigmoid function mapping any z into the interval $(0; 1)$: $\sigma(0) = 0.5$, $\sigma \to 1$ as $z \to +\infty$, and $\sigma \to 0$ as $z \to -\infty$;
- $e \approx 2.718$ is the base of the natural logarithm.

The advantage of logistic regression lies in its interpretability. For example, the weight of the flexibility feature is expected to be large and negative, since the more freely an expression varies, the lower the probability that it is an idiom.

Support vector machine (SVM). The support vector machine is a discriminative classifier that constructs the optimal separating hyperplane with the maximum margin between the two classes in the feature space [33]. For nonlinear dependencies the method uses the kernel trick, and the decision function $g(\mathbf{x})$ for a new candidate is written as

$$g(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{l} \alpha_i \, y_i \, \varkappa(\mathbf{x}_i, \mathbf{x}) + b\right), \tag{10}$$

where:

- $g(\mathbf{x})$ is the SVM decision function (it implements the classification stage of the overall mapping F of Eq. (1));
- l is the number of training examples;
- $\mathbf{x}_i$ is the feature vector of the ith example;
- $y_i \in \{-1, +1\}$ is the class of the example; here the $\{0, 1\}$ labels of Eq. (1) are recoded to $\{-1, +1\}$ solely by the SVM convention (idiom $\to +1$, free phrase $\to -1$);
- α_i are the Lagrange multipliers; $\alpha_i > 0$ only for the support vectors (the examples closest to the boundary), and $\alpha_i = 0$ otherwise;
- $\varkappa(\mathbf{x}_i, \mathbf{x})$ is the kernel function; the radial basis kernel is $\varkappa(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2)$ with the kernel parameter $\gamma > 0$;
- b is the intercept of the hyperplane;
- $\operatorname{sign}(\cdot)$ returns "idiom" $(+1)$ if the sum is positive and "free phrase" (-1) if it is negative.

Owing to the kernel trick the method separates even classes that are not linearly separable by mapping them into a higher-dimensional space, which makes it effective in the borderline, difficult cases of distinguishing idioms from free phrases.

Gradient boosting. Gradient boosting is an ensemble method that adds weak base models, usually decision trees, one after another, each new model being trained to reduce the error of the ensemble built so far [34]. The final prediction is formed as the sum of all trees:

$$\widehat{y} = F_K(\mathbf{x}) = \sum_{k=1}^{K} \gamma_k \, h_k(\mathbf{x}), \tag{11}$$

where:

- $\widehat{y}$ is the final prediction of the model;
- K is the number of decision trees (weak models);
- $h_k(\mathbf{x})$ is the kth decision tree (a weak model on its own);
- γ_k is the contribution (weight) of the kth tree to the final answer;
- $F_K(\mathbf{x})$ is the full ensemble formed by the sum of all K trees.

Each successive tree is trained to correct the error of the previous ensemble. At step k, the residual (antigradient) is computed for every example:

$$r_{ik} = -\left[\frac{\partial L\big(y_i, \, F_{k-1}(\mathbf{x}_i)\big)}{\partial F_{k-1}(\mathbf{x}_i)}\right], \tag{12}$$

where:

- r_{ik} is the residual of the ith example at step k; the new tree h_k is trained to predict precisely these residuals;

- $L(y_i, \cdot)$ is the loss function measuring the error between the true label y_i and the ensemble prediction;
- $F_{k-1}(\mathbf{x}_i)$ is the prediction of the ensemble of the previous $(k-1)$ trees for the ith example;
- $\partial/\partial F_{k-1}$ is the partial derivative with respect to the prediction; the minus sign turns it into the direction of the fastest decrease of the loss (the antigradient).

Features that are weak on their own may indicate an idiom clearly in combination; gradient boosting and the SVM are designed to discover precisely such hidden, nonlinear dependencies.

4 Evaluation metrics

The quality of the model is assessed against the gold standard on a test set. First the four elements of the confusion matrix are determined: TP is the number of correctly detected idioms, FP is the number of free phrases wrongly labelled as idioms, FN is the number of missed true idioms, and TN is the number of correctly rejected free phrases. The metrics are then computed from them:

$$P = \frac{TP}{TP + FP}, \tag{13}$$

where:

- TP is the number of correctly detected idioms;
- FP is the number of free phrases wrongly labelled as idioms;
- P (precision) shows what share of the answers "idiom" produced by the model are indeed idioms;

$$R = \frac{TP}{TP + FN}, \tag{14}$$

where:

- FN is the number of true idioms missed by the model;
- R (recall) shows what share of the true idioms in the text the model has managed to find.

Recall is particularly important for idioms, because in texts they occur in morphologically transformed shapes (*boshim qotayapti*, *boshing qotmasin*) that simple methods fail to recognize. To balance the two metrics, their harmonic mean is taken:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{15}$$

where:

- F_1 is the harmonic mean of precision and recall (the principal quality criterion);
- in the harmonic mean the smaller value dominates: if either P or R is low, F_1 drops sharply as well. For instance, with $P = 1.0$ and $R = 0.1$ the arithmetic mean would misleadingly show 0.55, whereas the harmonic mean is about 0.18.

The threshold-independent discriminative ability of a classifier is examined by ROC analysis [37]. Two rates are computed for this purpose:

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}, \tag{16}$$

where:

- TN is the number of correctly rejected free phrases;
- TPR is the true positive rate; it coincides exactly with the recall R;
- FPR is the false positive rate, that is, the share of free phrases wrongly labelled as idioms;
- the ROC curve is built from the pairs (FPR, TPR); $AUC = 0.5$ corresponds to random guessing and $AUC \to 1$ to ideal separation.

5 A worked example of candidate evaluation

We illustrate the computation of the features on two candidates: *boshi qotdi* 'he was puzzled' (an idiom) and *ovqat pishirildi* 'the food was cooked' (a free phrase). Both match the same grammatical pattern, a possessive-marked NOUN + VERB sequence, and therefore pass the filter of Eq. (3). The final decision is made by the remaining features. The frequencies are taken from a conventional mini-corpus of $N = 100{,}000$ tokens. In this illustration the semantic non-compositionality feature $x_{\text{comp}}(c)$ is not computed explicitly, since it requires an embedding model (see Section 6.2); the decision is therefore demonstrated on the four features x_{templ}, x_{som}, x_{PMI}, and x_{flex}.

Cohesion. Suppose that for *boshi qotdi* the word *boshi* occurs 125 times, the word *qotdi* occurs 64 times, and the pair occurs side by side 40 times. Then, by Eq. (5),

$$P(u,v) = \frac{40}{100{,}000} = 4.0 \times 10^{-4}, \quad P(u) = 1.25 \times 10^{-3}, \quad P(v) = 6.4 \times 10^{-4},$$

$$x_{\text{PMI}} = \log_2 \frac{4.0 \times 10^{-4}}{1.25 \times 10^{-3} \cdot 6.4 \times 10^{-4}} = \log_2 500 \approx 8.97.$$

Suppose that for *ovqat pishirildi* the word *ovqat* occurs 800 times, *pishirildi* occurs 250 times, and the pair occurs 16 times. Then, by Eq. (5),

$$x_{\text{PMI}} = \log_2 \frac{1.6 \times 10^{-4}}{8.0 \times 10^{-3} \cdot 2.5 \times 10^{-3}} = \log_2 8 = 3.00.$$

Both values are positive, but the difference is large (8.97 against 3.00): PMI separates the idiom from the free phrase reliably.

Flexibility. If *boshi qotdi* occurs $N(c) = 50$ times in the mini-corpus, of which 41 are in the canonical form, then, by Eq. (6),

$$x_{\text{flex}} = 1 - \frac{41}{50} = 0.18.$$

If *ovqat pishirildi* preserves its pattern in only 12 of $N(c) = 30$ occurrences, then

$$x_{\text{flex}} = 1 - \frac{12}{30} = 0.60.$$

Classifier decision (illustrated with logistic regression). Suppose training has selected the weights $w_{\text{templ}} = 0.4$, $w_{\text{som}} = 1.1$, $w_{\text{PMI}} = 0.35$, $w_{\text{flex}} = -2.5$, and $b = -3.0$. For *boshi qotdi* we have $x_{\text{templ}} = 1$, $x_{\text{som}} = 1$, $x_{\text{PMI}} = 8.97$, and $x_{\text{flex}} = 0.18$, so, by Eq. (9),

$$z = 0.4 \cdot 1 + 1.1 \cdot 1 + 0.35 \cdot 8.97 - 2.5 \cdot 0.18 - 3.0 = 1.19,$$

$$P(y = 1 \mid \mathbf{x}) = \sigma(1.19) = \frac{1}{1 + e^{-1.19}} \approx 0.77.$$

For *ovqat pishirildi* we have $x_{\text{som}} = 0$, $x_{\text{PMI}} = 3.00$, and $x_{\text{flex}} = 0.60$, so

$$z = 0.4 \cdot 1 + 1.1 \cdot 0 + 0.35 \cdot 3.00 - 2.5 \cdot 0.60 - 3.0 = -3.05,$$

$$P(y = 1 \mid \mathbf{x}) = \sigma(-3.05) \approx 0.045.$$

With the decision threshold at 0.5, the model correctly classifies *boshi qotdi* as an idiom ($0.77 > 0.5$) and *ovqat pishirildi* as a free phrase ($0.045 < 0.5$). The final description of the two candidates is summarized in Table 3.

Table 3 Numerical description of the two candidates by features.

Feature	*boshi qotdi* (idiom)	*ovqat pishirildi* (free phrase)
x_{templ}	1	1
x_{som}	1 (*bosh*)	0
x_{PMI}	8.97	3.00
x_{flex}	0.18 (fixed)	0.60 (free)
x_{comp}	low (non-compositional)	high (compositional)
$P(y = 1 \mid \mathbf{x})$	**0.77 — "idiom"**	**0.045 — "free phrase"**

Both candidates pass the grammatical filter ($x_{\text{templ}} = 1$), and the final decision is made by the combination of the remaining features. This is the practical embodiment of the main thesis of Section 3.1: the template delimits the pool of candidates, while the final verdict is delivered by statistics and semantics.

6 Experiment

In practice the three-stage model described above operates in the following sequence. Candidates are taken one at a time; a candidate is first passed through the grammatical template filter of Eq. (3), then its feature vector is computed by Eqs. (5)–(7), and the classifier of Eqs. (8)–(11) labels it as an idiom or a free phrase. The process is repeated until all candidates have been examined. The complete operating procedure of the model is summarized in the flowchart of Fig. 1 at the end of this section.

6.1 Experimental data

The experiment relied on three principal resources. The corpus is a combined Uzbek corpus comprising eight literary works (by A. Qodiriy, Cho'lpon, Oybek, G'. G'ulom, O'. Hoshimov, and other authors) and a sample of an open corpus, totalling 154,502 sentences and about 1.07 million lemma-tokens. The text was lemmatized with the Uz-MorphAnalyser tool [27]. Lemmatization resolves the rich morphology of Uzbek and reduces the various grammatical shapes of an idiom to a single lemma pattern.

As the lexical resource, 1,362 units of the Uzbek phraseological ontology developed by the authors were used, each supplied with component patterns. Every idiom is given a canonical form and a lemma pattern; for instance, the idiom *xotiridan chiqmoq* 'to slip one's mind' corresponds to the pattern "xotir chiq".

As the gold standard, a manually verified, stratified test set of 400 sentences was compiled. Stratum A consists of 150 sentences containing idioms mostly in canonical form, stratum B of 150 sentences with morphologically and syntactically transformed shapes, and stratum C of 100 predominantly idiom-free, distractor sentences. The set contains 224 positive and 176 negative sentences in total. The final evaluation was carried out on this set only, at the sentence level.

6.2 Implementation details

The protocol reproduces the pipeline of Section 3 exactly. Lemma-level pairs served as candidates. These include pairs whose second component belongs to the verb lexicon (275 verb lemmas extracted automatically from the ontology) with a gap of at most 2 tokens, as well as full matches of the ontology patterns.

The features used were x_{som}; x_{PMI} computed from co-occurrences within a window; x_{flex} estimated as the share of strictly adjacent occurrences relative to windowed occurrences; and the frequency feature $\log_{10}(1 + N_{\text{win}})$. The semantic non-compositionality feature x_{comp} was not included in this test, since it requires an embedding model, and was left for future work.

The positive class of the training set was formed by the 634 idioms found in the corpus, that is, 46.8 % of the 1,362 ontology patterns occurred in the corpus at least once. The negative class comprised 1,268 frequent verb-final pairs that do not match any ontology pattern. The gold set took no part in training. The classifiers were implemented with the scikit-learn library [38].

As an internal check, five-fold cross-validation on the training set gave the following F_1 scores: Naive Bayes 0.551 ± 0.063, logistic regression 0.550 ± 0.074, SVM 0.699 ± 0.033, and gradient boosting 0.726 ± 0.029. Two baselines were taken for comparison: lexical matching (searching for the canonical form of an idiom verbatim in a sentence) and ontology-based matching (lemmatizing the sentence and searching for an ontology lemma pattern with a gap of at most 2 tokens).

The complete operating procedure of the model described above, from taking a candidate through filtering, vectorization, and classification to the iteration over all candidates, is summarized in the flowchart of Fig. 1.

7 Results and discussion

7.1 Overall results

The results of all methods on the gold-standard set are given in Table 4, and their graphical comparison in Fig. 2.

Table 4 Results of the methods on the gold-standard set (400 sentences).

Model	Precision	Recall	F_1
Lexical matching (baseline)	0.706	0.054	0.100
Ontology-based matching (lemma patterns)	**0.854**	0.496	0.627
Naive Bayes	0.691	0.679	0.685
Logistic regression	0.701	0.754	0.727
Support vector machine (SVM)	0.707	**0.906**	**0.795**
Gradient boosting	0.702	**0.906**	0.791

The results lead to several conclusions. The lexical baseline barely works ($F_1 = 0.100$; $R = 0.054$), because idioms almost never occur in literary text in their canonical dictionary form; they appear transformed by possessive, tense, and person affixes (*boshini yerdan uzmadi*, *zavqi toshib ketdi*). Combining morphology with the ontology gives a substantial improvement ($F_1 = 0.627$; $P = 0.854$), but recall stops at 0.496, since the method can only find idioms present in the ontology.

The machine learning models raise recall dramatically. The best result is shown by the SVM ($F_1 = 0.795$; $R = 0.906$), with gradient boosting a very close second ($F_1 = 0.791$).

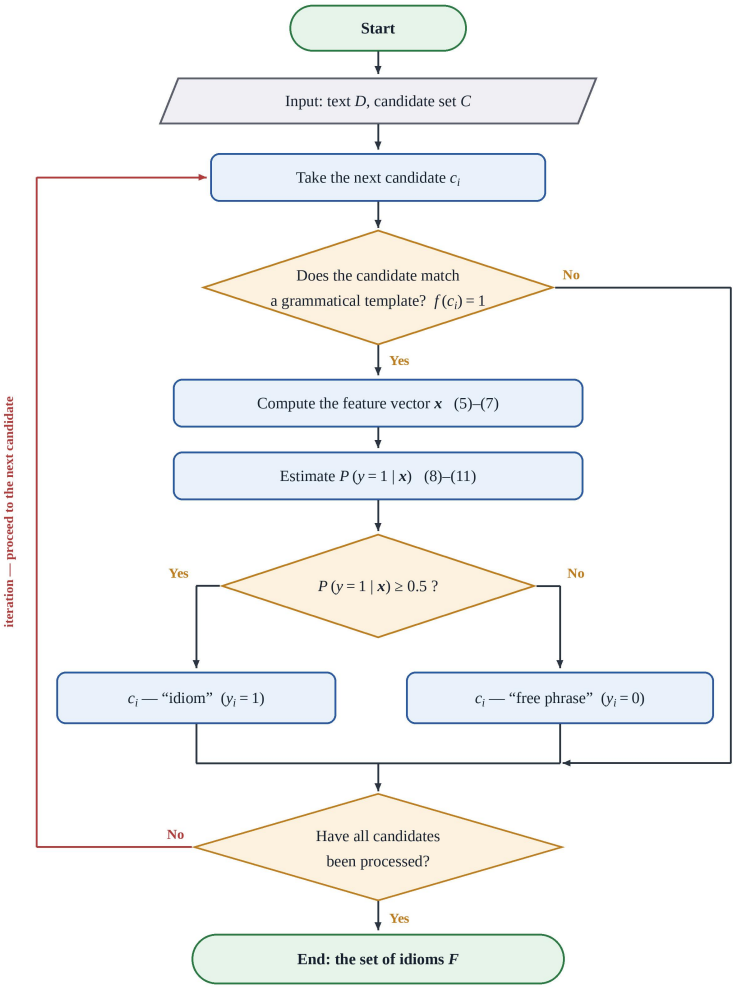

Figure 1 Flowchart of the automatic idiom extraction model.

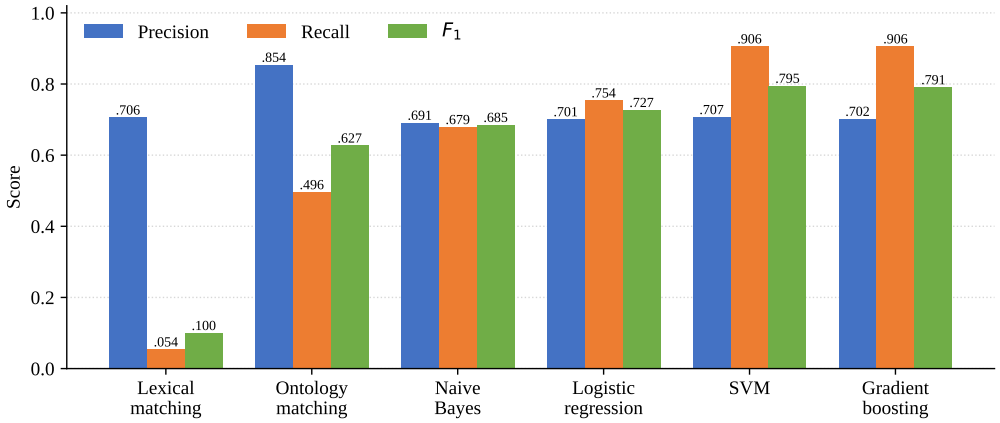

Figure 2 Comparison of the methods by precision, recall, and F_1 (gold set, 400 sentences).

They detect idioms absent from the ontology through the combination of corpus-statistical cues such as high cohesion, low flexibility, and the somatic indicator. The precision of all learning models clusters around 0.70; the main sources of false positives are stable

collocations that look like idioms and the noise of weak labelling. The next step towards higher precision is adding the semantic non-compositionality feature.

ROC analysis demonstrates the stable, threshold-independent discriminative ability of the learning models (*AUC* values from 0.70 to 0.73), as reflected in Fig. 3.

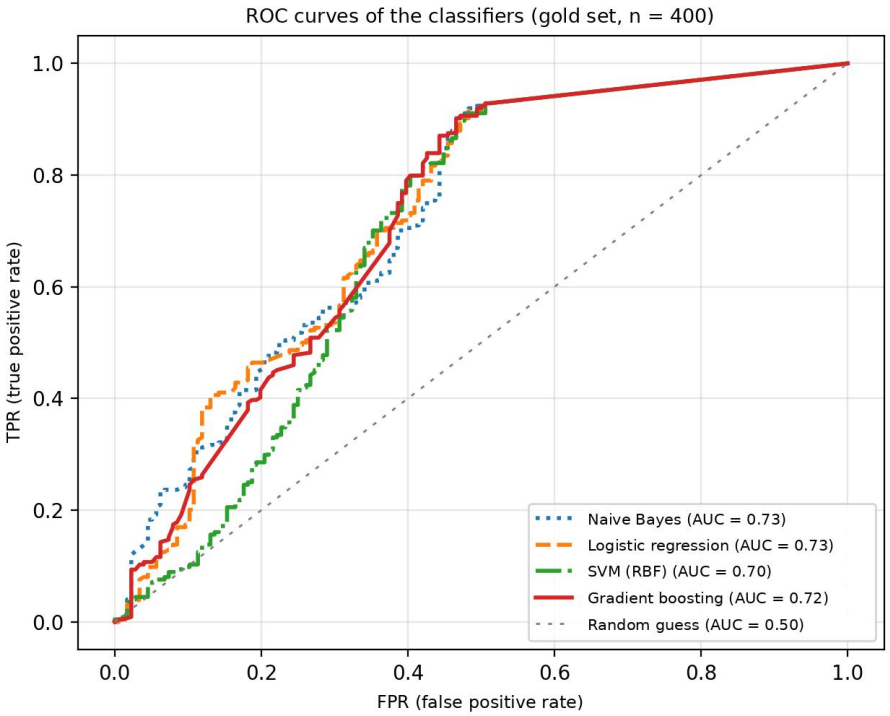

Figure 3 ROC curves of the classifiers (gold set, 400 sentences).

7.2 Stratified analysis

The stratification of the gold set makes it possible to identify the situations in which the advantage of the models manifests itself. Table 5 gives the recall of gradient boosting and of the lexical baseline by stratum.

Table 5 Recall by stratum.

Stratum	Description	Gradient boosting	Lexical matching
A	idioms in canonical form (150 sentences)	0.964	0.065
B	morphologically and syntactically transformed shapes (150 sentences)	0.919	0.041
C	rare and distractor cases (100 sentences)	low	low

The advantage of the model manifests itself precisely in detecting morphologically transformed shapes of idioms. In stratum B the learning model delivers a recall nearly 22 times that of the lexical method. Rare idioms of stratum C remain an open problem for all methods.

7.3 The role of each method in the pipeline

The experiment made it possible to establish quantitatively at which stage and under which conditions each method is most effective. These findings are summarized in Table 6.

Table 6 The most effective place of application of each method.

Method or tool	Place in the pipeline	Most effective use	Evidence from the experiment
Grammatical templates, Eqs. (2) and (3)	stage one, screening	initial narrowing of the candidate space	drastically reduces computation; necessary but not sufficient condition
Morphological lemmatization [27]	stage one, text preparation	reducing transformed shapes to a pattern	raises F_1 from 0.100 to 0.627
Ontology-based matching	stage one, high precision	reliable labelling of known dictionary idioms	highest precision, $P = 0.854$
PMI, Eq. (5)	stage two, feature	measuring the cohesion of two-component candidates	separates an idiom from a free phrase (8.97 against 3.00)
C-value-type measures [9]	stage two, feature	candidates with three or more components	accounts for nested expressions
Flexibility index, Eq. (6)	stage two, feature	measuring the degree of idiomaticity	confirms Vinogradov's gradient ($0.202 < 0.246 < 0.290$)
Semantic non-compositionality, Eq. (7)	stage two, feature	separating idioms from stable collocations	the tool for overcoming the 0.70 precision barrier (planned)
Naive Bayes, Eq. (8)	stage three	fast baseline, initial calibration	$F_1 = 0.685$; precision limited by the violated independence assumption
Logistic regression, Eq. (9)	stage three	interpreting the influence of the features	weight analysis matches theoretical expectations ($w_{\text{flex}} < 0$)
SVM, Eq. (10)	stage three, main classifier	separating borderline, difficult cases	best overall result, $F_1 = 0.795$, $R = 0.906$
Gradient boosting, Eqs. (11) and (12)	stage three, main classifier	discovering nonlinear feature combinations	$F_1 = 0.791$; most stable in internal validation (0.726 ± 0.029)

The table shows that the methods complement rather than replace one another. Grammatical templates and morphological lemmatization prepare the candidate space; ontology-based matching covers known idioms with high precision; PMI and the flexibility index express the degree of idiomaticity numerically; and the SVM and gradient

boosting, building on this combination of features, also detect idioms beyond the ontology and push recall up to 0.906. In practice the choice depends on the demands of the task: ontology-based matching suits applications that require high precision, such as dictionary linking, whereas the SVM or gradient boosting serve applications that require high recall, such as pre-marking idioms before machine translation.

8 Conclusion

The experimental test logically confirmed the central idea of the model. No single method can reliably separate an idiom from a free phrase, while each method in the pipeline contributes its own strength and pays off in its own place.

The results by method make this evident. The lexical method that searches for the canonical form verbatim produced almost no result, remaining at $F_1 = 0.100$. Ontology-based matching supported by morphological lemmatization turned out to be the most precise method, with $P = 0.854$ and $F_1 = 0.627$, which makes it appropriate for labelling known dictionary idioms. The learning models raised recall sharply: Naive Bayes reached $F_1 = 0.685$ and logistic regression $F_1 = 0.727$. The support vector machine achieved the best overall quality with a recall of $R = 0.906$ and $F_1 = 0.795$, while gradient boosting came very close with $F_1 = 0.791$ and proved the most stable method in internal validation (0.726 ± 0.029). The stratified analysis located the source of the advantage: on the stratum of morphologically transformed idiom shapes the learning model attained a recall of 0.919 against 0.041 for the lexical method, so the main strength of the model lies precisely in recognizing idioms transformed in text.

In the future we will present a study of individual models in greater depth alongside their hybrid combinations. This line of research will lay the foundation for a new family of models and algorithms serving the high-quality translation of idioms in Uzbek–Turkish machine translation.

Declarations

Conflicts of interest. The authors declare no conflict of interest.

References

[1] Sag I. A., Baldwin T., Bond F., Copestake A., Flickinger D. 2002. Multiword expressions: A pain in the neck for NLP. *Computational Linguistics and Intelligent Text Processing (CICLing 2002)*. LNCS, Vol. 2276. Berlin: Springer. 1–15.

[2] Baldwin T., Kim S. N. 2010. Multiword expressions. *Handbook of Natural Language Processing*. 2nd ed. Boca Raton: CRC Press. 267–292.

[3] Ramisch C. 2015. *Multiword expressions acquisition: A generic and open framework*. Cham: Springer. 230 p.

[4] Constant M., Eryiğit G., Monti J., van der Plas L., Ramisch C., Rosner M., Todirascu A. 2017. Multiword expression processing: A survey. *Computational Linguistics*. 43(4): 837–892.

[5] Church K. W., Hanks P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*. 16(1): 22–29.

[6] Manning C. D., Schütze H. 1999. *Foundations of statistical natural language processing*. Cambridge: MIT Press. 680 p.

[7] Evert S. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. PhD dissertation. Stuttgart: Universität Stuttgart. 353 p.

[8] Pecina P. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*. 44(1–2): 137–158.

[9] Frantzi K., Ananiadou S., Mima H. 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*. 3(2): 115–130.

[10] Fazly A., Cook P., Stevenson S. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*. 35(1): 61–103.

[11] Cook P., Fazly A., Stevenson S. 2008. The VNC-Tokens dataset. *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*. Marrakech. 19–22.

[12] Katz G., Giesbrecht E. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. *Proceedings of the Workshop on Multiword Expressions*. Sydney: ACL. 12–19.

[13] Salehi B., Cook P., Baldwin T. 2015. A word embedding approach to predicting the compositionality of multiword expressions. *Proceedings of NAACL-HLT 2015*. Denver: ACL. 977–983.

[14] Cordeiro S., Villavicencio A., Idiart M., Ramisch C. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*. 45(1): 1–57.

[15] Mikolov T., Chen K., Corrado G., Dean J. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[16] Devlin J., Chang M.-W., Lee K., Toutanova K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*. Minneapolis: ACL. 4171–4186.

[17] Reimers N., Gurevych I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP 2019*. Hong Kong: ACL. 3982–3992.

[18] Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. 2022. Language-agnostic BERT sentence embedding. *Proceedings of the 60th Annual Meeting of the ACL*. Dublin: ACL. 878–891.

[19] Zeng Z., Bhat S. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*. 9: 1546–1562.

[20] Tayyar Madabushi H., Gow-Smith E., Garcia M., Scarton C., Idiart M., Villavicencio A. 2022. SemEval-2022 Task 2: Multilingual idiomaticity detection and sentence embedding. *Proceedings of SemEval-2022*. Seattle: ACL. 107–121.

[21] Savary A., Ramisch C., Cordeiro S., Sangati F., Vincze V., QasemiZadeh B., Candito M., Cap F., Giouli V., Stoyanova I., Doucet A. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia: ACL. 31–47.

[22] Ramisch C., Cordeiro S., Savary A. [et al.] 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. *Proceedings of LAW-MWE-CxG-2018*. Santa Fe: ACL. 222–240.

[23] Berk G., Erden B., Güngör T. 2018. Deep-BGT at PARSEME shared task 2018: Bidirectional LSTM-CRF model for verbal multiword expression identification. *Proceedings of LAW-MWE-CxG-2018*. Santa Fe: ACL. 248–253.

[24] Oflazer K. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*. 9(2): 137–148.

[25] Oflazer K., Saraçlar M. (eds.) 2018. *Turkish natural language processing*. Cham: Springer. 367 p.

[26] Kuriyozov E., Doval Y., Gómez-Rodríguez C. 2020. Cross-lingual word embeddings for Turkic languages. *Proceedings of LREC 2020*. Marseille: ELRA. 4054–4062.

[27] Salaev U. 2024. UzMorphAnalyser: A morphological analysis model for the Uzbek language using inflectional endings. *AIP Conference Proceedings*. 3244: Art. 030058. doi: `https://doi.org/10.1063/5.0241461`.

[28] Salaev U., Kuriyozov E., Gómez-Rodríguez C. 2022. SimRelUz: Similarity and relatedness scores as a semantic evaluation dataset for Uzbek language. *Proceedings of SIGUL 2022*. Marseille: ELRA. 199–206.

[29] Vinogradov V. V. 1977. Ob osnovnykh tipakh frazeologicheskikh edinits v russkom yazyke [On the main types of phraseological units in the Russian language]. *Vinogradov V. V. Izbrannye trudy. Leksikologiya i leksikografiya [Selected works. Lexicology and lexicography]*. Moscow: Nauka. 140–161. (In Russian)

[30] Shanskiy N. M. 1985. *Frazeologiya sovremennogo russkogo yazyka [Phraseology of the modern Russian language]*. 3rd ed. Moscow: Vysshaya shkola. 160 p. (In Russian)

[31] Rahmatullayev Sh. 1966. *O'zbek frazeologiyasining ba'zi masalalari [Some issues of Uzbek phraseology]*. Tashkent: Fan. (In Uzbek)

[32] Rahmatullayev Sh. 1978. *O'zbek tilining izohli frazeologik lug'ati [Explanatory phraseological dictionary of the Uzbek language]*. Tashkent: O'qituvchi. (In Uzbek)

[33] Cortes C., Vapnik V. 1995. Support-vector networks. *Machine Learning*. 20(3): 273–297.

[34] Friedman J. H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. 29(5): 1189–1232.

[35] Chen T., Guestrin C. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco: ACM. 785–794.

[36] Hastie T., Tibshirani R., Friedman J. 2009. *The elements of statistical learning*. 2nd ed. New York: Springer. 745 p.

[37] Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*. 27(8): 861–874.

[38] Pedregosa F., Varoquaux G., Gramfort A. [et al.] 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12: 2825–2830.

УДК 004.912:81'322

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ ФРАЗЕОЛОГИЗМОВ В ТЮРКСКИХ ЯЗЫКАХ

[1]***Назирова Э.Ш.***, [2*]***Мадетбаева Н.Б.***, [1]***Мирзаева Ф.Ш.***, [3]***Исломова М.Б.***

*madetbaeva@gmail.com

[1]Ташкентский университет информационных технологий имени Мухаммада аль-Хоразмий,

100084, Узбекистан, г. Ташкент, просп. Амира Темура, 108;

[2]Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта,

100125, Узбекистан, г. Ташкент, м-в Буз-2, 17А;

[3]Университет Инха в Ташкенте,

100170, Узбекистан, г. Ташкент, ул. Зиёлилар, 9

№ 4(74) 2026

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

AiRI
RAQAMLI TEXNOLOGIYALAR VA SUN'IY INTELLEKTNI RIVOJLANTIRISH ILMIY-TADQIQOT INSTITUTI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 4(74) 2026

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(998) 71 263-41-98.
Э-почта: journals@airi.uz.
Веб-сайт: https://journals.airi.uz.

Дизайн и вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 25.08.2026 г.
Формат 60x84 1/8. Заказ № 4. Тираж 100 экз.

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

No. 4(74) 2026

The journal was established in 2015.
6 issues are published per year.

Founder:
Digital Technologies and Artificial Intelligence Development Research Institute.

The journal is registered by Agency of Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan.
Certificate of Registration No. 0856 of 5 August 2015.

ISSN 2181-8460, eISSN 2181-046X

Address:
100125, Tashkent, Buz-2, 17A.
Tel.: +(998) 71 263-41-98.
E-mail: journals@airi.uz.
Web-site: https://journals.airi.uz.

Layout design:
Sharipov Kh.D.

DTAIRI printing office.
Signed for print 25.08.2026
Format 60x84 1/8. Order No. 4. Print run of 100 copies.

Содержание

Ахмедов Д.Д., Туркменова Р.Т.
Моделирование мезомасштабного переноса мелкодисперсной примеси в атмосфере с учётом стохастических пульсаций ветра 5

Чжан Х., Варламова Л.П., У Б.
Динамический анализ стохастической SICR-модели передачи раннего и позднего фитофтороза картофеля 27

Равшанов Н., Шафиев Т.Р., Бобожонова М.А., Кобилова Д.О.
Исследование ключевых факторов, влияющих на распространение пылевых частиц в атмосфере 47

Бердиёров Ш.Ш., Эрмонова М.У.
Сравнение методов конечных разностей и конечных объёмов для численного моделирования фильтрации в цилиндрическом пористом фильтре 65

Равшанов Н., Садуллаев С.
Моделирование процесса нестационарной фильтрации грунтовых вод со свободной поверхностью с учётом инфильтрационных потоков 85

Туракулов Ж.А., Жураев И.Р., Таштемирова Н.Н., Рахматуллаева Д.О.
Математическое моделирование фильтрации суспензии в пористых средах . 102

Хаётов А.Р., Исмоилов С.И.
Квадратурная формула для оптимального приближения определенных интегралов с полиномиальным весом 110

Нуралиев Ф.А., Едилбекова Р.М.
Оптимальная квадратурная формула в пространстве дифференцируемых функций 120

Эшмаматова Д.Б., Очилова Н.К.
Вырожденно-сингулярные эллиптические операторы: точная коэрцитивность, весовая корректность и методы функции Грина для моделей диффузии в опухолевых тканях 133

Назирова Э.Ш., Мадетбаева Н.Б., Мирзаева Ф.Ш., Исломова М.Б.
Математическая модель автоматического выделения фразеологизмов в тюркских языках 152

Ахмеджанова Д.А.
Модель семантического связывания тюркских языков: подход с ограничением справедливости для малоресурсных условий 171

Contents

Akhmedov D.D., Turkmenova R.T.
Modeling mesoscale transport of fine particulate matter in the atmosphere accounting for stochastic wind fluctuations 5

Zhang H., Varlamova L.P., Wu B.
Dynamical analysis of a stochastic SICR transmission model for potato early and late blight 27

Ravshanov N., Shafiev T.R., Bobozhonova M.A., Kobilova D.O.
Study of the key factors affecting the dispersion of dust particles in the atmosphere 47

Berdiyorov Sh.Sh., Ermonova M.U.
Comparison of finite difference and finite volume methods for numerical simulation of filtration in a cylindrical porous filter 65

Ravshanov N., Sadullaev S.
Modeling of the process of unsteady filtration of groundwater with a free surface taking into account infiltration flows 85

Turakulov J.A., Juraev I.R., Tashtemirova N.N., Raxmatullayeva D.O.
Mathematical modeling of suspension filtration in porous media 102

Hayotov A.R., Ismoilov S.I.
Optimal quadrature for definite integrals with polynomial weight 110

Nuraliev F.A., Edilbekova R.M.
Optimal quadrature formula in the space of differentiable functions 120

Eshmamatova D.B., Ochilova N.K.
Degenerate-singular elliptic operators: sharp coercivity, weighted well-posedness, and Green's function methods for tumor-diffusion models 133

Nazirova E.Sh., Madetbayeva N.B., Mirzayeva F.Sh., Islomova M.B.
A mathematical model for the automatic extraction of idioms in Turkic languages 152

Akhmedjanova D.A.
A semantic linking model for Turkic languages: a fairness-constrained cross-lingual approach for low-resource settings 171