




UDC 81'322.2

DOI: 10.18413/2313-8912-2024-10-4-0-6

Lyudmila S. Shurykina¹
Ekaterina A. Latukhina²
Tatiana V. Petrova³

Using CNN and LSTM neural networks
for Arkhangelsk dialect word identification
and classification

¹ Northern (Arctic) Federal University named after M.V. Lomonosov,
17 Severnaya Dvina Emb., Arkhangelsk, 163002, Russia
E-mail: l.shurykina@narfu.ru
ORCID: 0009-0004-8547-1967

² Northern (Arctic) Federal University named after M.V. Lomonosov,
17 Severnaya Dvina Emb., Arkhangelsk, 163002, Russia
E-mail: e.latukhina@narfu.ru
ORCID: 0000-0001-5145-5994

³ Northern (Arctic) Federal University named after M.V. Lomonosov,
17 Severnaya Dvina Emb., Arkhangelsk, 163002, Russia
E-mail: t.petrova@narfu.ru
ORCID: 0009-0004-0341-2470

Received 29 October 2024; accepted 15 December 2024; published 30 December 2024

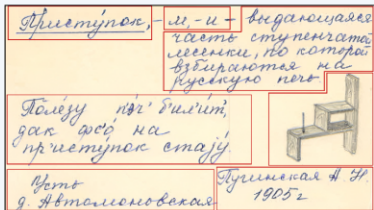
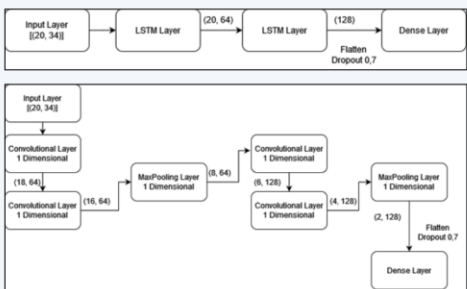
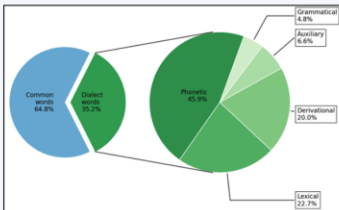

Abstract: The study of dialects provides an opportunity to gain an understanding of the culture and history of a people, which are reflected in language. Dialectal vocabulary differs from standard vocabulary in terms of both meaning and pronunciation, as well as word formation and grammatical structures, primarily in morphology. Similar patterns can also be observed in the Arkhangelsk dialects. The aim of this paper is to develop a dialect words classifier, which can be used to identify dialect words within a given text and categorize them into one of the pre-defined groups. The novelty of this research lies in the lack of an automated system for classifying dialect words based on Arkhangelsk dialect materials. The article describes the development of a neural network for dialect words identification and classification. Dialect words were identified from dialect texts gathered during dialectological research conducted between the 1960s and the present day. LSTM (long short-term memory) and CNN (convolutional neural network) architectures are compared. One of the main challenges in the task of dialect word classification is that the neural network is trained using a limited amount of data. To overcome these limitations, we are investigating the possibility of using a bigram-based approach in addition to the unigram-based words encoding. A trained model that demonstrated the best results was integrated into our application for dialect words processing and analysis. Confusion matrix was constructed for the best model which demonstrates the highest performance for the derivational class and the lowest for the lexical class.

Keywords: Dialect word classification; Natural language processing; Convolutional neural network; Long short-term memory

Acknowledgements: The research was supported by the Russian Science Foundation, project No. 23-28-01380, «Thematic dictionary of Arkhangelsk dialects with electronic support» (<https://rscf.ru/project/23-28-01380/>).




How to cite: Shurykina, L. S., Latukhina, E. A., Petrova, T. V. (2024). Using CNN and LSTM Neural Networks for Arkhangelsk Dialect Word Identification and Classification, *Research Result. Theoretical and Applied Linguistics*, 10 (4), 106–125. DOI: 10.18413/2313-8912-2024-10-4-0-6

USING CNN AND LSTM NEURAL NETWORKS FOR ARKHANGELSK DIALECT WORD IDENTIFICATION AND CLASSIFICATION

Problem	Solution	Results
<p>The dialect materials from the Arkhangelsk region are catalogued in a paper card index that has been compiled since the 1960s</p> <p>The data is heterogeneous and challenging to automatically process, recognize, and analyze</p> 	<p>A dataset of dialectal words from cards has been compiled</p> <p>Several CNN and LSTM models have been developed to classify dialectal words into selected categories</p> 	<p>The best model demonstrated the highest performance for the derivational class and the lowest for the lexical class</p> <p>The best trained model is available on GitHub</p> 
<p>Tools for dialect analysis are required</p>	<p>The research is supported by Russian Science Foundation</p> 	

УДК 81'322.2

DOI: 10.18413/2313-8912-2024-10-4-0-6

Шурыкина Л. С.¹
Латухина Е. А.²
Петрова Т. В.³

Применение нейронных сетей CNN и LSTM для идентификации и классификации диалектизмов на материалах архангельских говоров

¹ Северный (Арктический) федеральный университет имени М.В. Ломоносова
наб. Северной Двины, 17, Архангельск, 163002, Россия

E-mail: l.shurykina@narfu.ru

ORCID: 0009-0004-8547-1967

² Северный (Арктический) федеральный университет имени М.В. Ломоносова
наб. Северной Двины, 17, Архангельск, 163002, Россия

E-mail: e.latukhina@narfu.ru

ORCID: 0000-0001-5145-5994

³ Северный (Арктический) федеральный университет имени М.В. Ломоносова
наб. Северной Двины, 17, Архангельск, 163002, Россия

E-mail: t.petrova@narfu.ru

ORCID: 0009-0004-0341-2470

Статья поступила 29 октября 2024 г.; принята 15 декабря 2024 г.;
опубликована 30 декабря 2024 г.

Аннотация: Изучение диалектов позволяет составить представление о культуре и истории народа, которые находят отражение в лексике языка. Диалектная лексика отличается от нормативной как значением, так и произношением, способами словопроизводства и грамматической структурой, прежде всего морфологией. Подобные закономерности характерны и для архангельских говоров. Цель исследования – разработать классификатор диалектных слов, который поможет выделить диалектные слова в конкретном заданном тексте и отнести их к одной из заранее определенных категорий. Новизна исследования состоит в том, что в настоящее время отсутствует автоматизированная система для классификации диалектизмов, основанная на материалах архангельских говоров. В статье описывается разработка нейронных сетей для идентификации и классификации диалектных слов, извлеченных из диалектных текстов, которые были собраны во время диалектологических практик, проводившихся с 1960-х годов по настоящее время; сравниваются архитектуры LSTM (Long Short-Term Memory, нейронная сеть с долгосрочной кратковременной памятью) и CNN (Convolutional Neural Network, свёрточная нейронная сеть). Нейронная сеть обучается на малом количестве материала, что является одним из основных ограничений в задаче классификации диалектных слов. Чтобы обойти эти ограничения, исследуется возможность использовать биграммный подход кодирования слов в дополнение к униграммному. Обученная модель, показавшая наилучшие результаты, встроена в разрабатываемое приложение для обработки и анализа диалектизмов. Для этой модели была построена матрица ошибок, согласно которой лучше всего распознаются слова из словообразовательной категории, хуже всего – из лексической.

Ключевые слова: Классификация диалектизмов; Обработка естественного языка; Свёрточные нейронные сети; Нейронные сети с долгой краткосрочной памятью

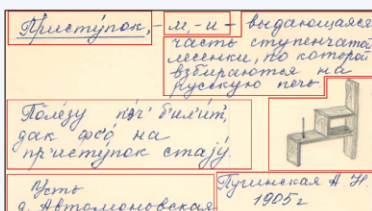
Информация об источниках финансирования или грантах: Исследование выполнено при финансовой поддержке Российского научного фонда № 23-28-01380, «Тематический словарь архангельских говоров с электронной поддержкой» (<https://rscf.ru/project/23-28-01380/>).

Информация для цитирования: Шурыкина Л. С., Латухина Е. А., Петрова Т. В. Применение нейронных сетей CNN и LSTM для идентификации и классификации диалектизмов на материалах архангельских говоров // Научный результат. Вопросы теоретической и прикладной лингвистики. 2024. Т. 10. № 4. С. 106–125. DOI: 10.18413/2313-8912-2024-10-4-0-6

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ CNN И LSTM ДЛЯ ИДЕНТИФИКАЦИИ И КЛАССИФИКАЦИИ ДИАЛЕКТИЗМОВ НА МАТЕРИАЛАХ АРХАНГЕЛЬСКИХ ГОВОРОВ

Проблема

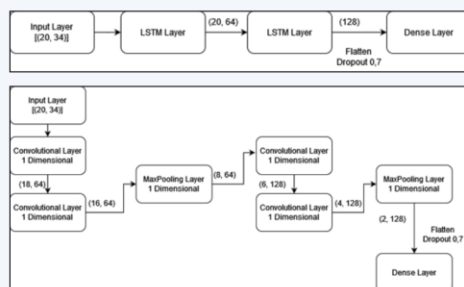
Диалектный материал архангельских говоров хранится в бумажной картотеке, наполнявшейся с 1960-х годов. Карточки разнородные, сложно автоматически распознавать, обрабатывать и анализировать.



Нужны инструменты диалектного анализа

Решение

Собран датасет из диалектизмов с карточек. Обучено несколько моделей CNN и LSTM для классификации диалектных слов по выделенным категориям.

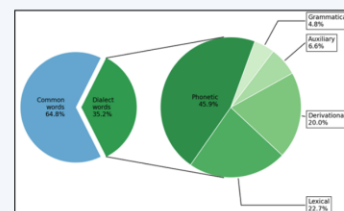


Исследование выполнено при поддержке Российского научного фонда



Результаты

Лучше всего распознаются слова из словообразовательной категории, хуже всего — из лексической. Лучшая обученная модель доступна на GitHub.



1. Introduction

Dialectology is a branch of linguistics that focuses on the study of linguistic variation within a given language community. This field of study highlights the richness and diversity of languages and provides a fascinating avenue for researchers to explore dialect phenomena and create dialect corpora. The study of dialectal features is an important aspect of modern linguistics, as dialects reflect cultural and historical aspects of different regions (Samsitova, 2020).

Dialectisms are linguistic features that are specific to certain dialects associated with a particular geographical area. These features include words, phrases, and expressions that are unique to that region. The classification of dialectisms assists in organizing the differences between regional language variations, as well as the cultural differences that are reflected in these variations (Kornauhova, Goloshtanova, 2022; Kolkova, 2023).

The dialects of a language are numerous and varied. In some countries, such as Italy, knowledge of local dialects can help individuals integrate into society (Shamshin, 2024). In other countries, dialects are actively

utilized in online spaces, including by younger generations (Zhang, Ren, 2022; Han et al., 2024).

Therefore, in (Høyland, Nesse, 2023), both traditional and more recent classification methods are examined based on the data of Norwegian dialects. Depending on the chosen method, the number of primary dialect regions varies from 3 to 12. A comparable study conducted on data from the regional dialects of Kuzbass permits the identification of four groups of lexical units that can be classified as dialectal (Kositsina, 2024). A character N-gram based dialect classifier is used in (Buckley, 2021).

The use of traditional methods for dialect research, such as statistical analysis or dialectometry-based classification approach (Sciarretta, 2024; Arkhangelskiy, 2021), requires a significant amount of manual effort. In order to ensure high accuracy and effectiveness in automatic dialect classification, it is necessary to employ sophisticated approaches such as neural networks.

The study of dialects using neural networks has been conducted by numerous researchers. Novel approaches to the

automatic classification of dialects, based on various technologies and neural network architectures, have been proposed. The authors of (Kethireddy et al., 2022) experimented with representations of features in order to capture the intrinsic articulatory differences between dialects. Researches (Themistocleous, 2017; Themistocleous, 2019; Alali et al., 2019) focused on using the acoustic characteristics of speech to distinguish dialects. Two classification approaches were proposed by the authors, comparing a deep neural network with Support Vector Machines, Random Forests, and Decision Trees. In Russia, oral speech corpora such as the corpus of Udmurt dialects (Vernyaeva, Zhdanova, 2023) are also used.

It is important to recognize dialects not only by their sounds, but also through their written form. In (Alali et al., 2019), an investigation into the application of a convolutional neural network for the classification of Arabic dialects was conducted. A detailed study of the use of various architectures, simple and complex classifiers such as Support Vector Machine and Deep Neural Networks, in relation to Arabic dialects, is presented in (Azim et al., 2022). It was demonstrated that convolutional neural networks on the proposed material outperform other classifiers in terms of accuracy, recall, and F1-score metrics.

Different approaches are used to increase performance of classifiers. The authors of (Ye et al., 2019) proposed an ensemble learning method for dialect classification, in which three models are used to classify dialects and the final decision is determined by voting. The GRU, CNN and DNN models are used to build an ensemble. This method makes dialect classification more precise and amplifies the initially low accuracy.

Large amount of data allows researchers to use powerful tools like transformer architecture. In (Laith, Kang, 2023) the authors demonstrate the transformer usage on the material of Arabic dialects and suggest an approach to dissecting sentiment in conditions

of limited available resources. The proposed model demonstrates better performance than empirical findings. A similar study was conducted in (Adel et al., 2024).

More often than not, dialect data is not enough to use transformers and similar architectures. To train neural networks using other architectures, a significant amount of dialectic data is also required. Therefore, it is necessary to collect additional datasets. New methods are proposed to achieve this goal. For instance, in (Yamani et al., 2024) the authors describe a collaborative approach for gathering high-quality textual data. In (Kuparinen, 2024) the author utilizes messages from forums that distinguish between standard Finnish and dialect variations.

On the other hand, dialect words may be defined as those that are not part of the model's vocabulary. The task of identifying and categorizing such words, although challenging, is essential for the practical application of models in real-world scenarios. For instance, in (Huang et al., 2024) the authors explore the significance of out-of-vocabulary words in LLM-powered recommendation systems.

Dialect studies involve various methods, ranging from traditional statistical analysis to more advanced techniques such as neural networks. The number of primary dialect regions may vary significantly depending on the method used, as demonstrated in the studies on Norwegian and Kuzbass dialects. Traditional methods require significant manual effort, while neural networks offer a more automated and potentially more accurate solution. Researchers have explored various neural network architectures and techniques to classify dialects based on both acoustic and written data. On the one hand, Convolutional Neural Networks (CNNs) have shown promising results in dialect classification, outperforming other classifiers in terms of accuracy, recall, and F1-score. On the other hand, Long Short-Term Memory are specifically designed to handle sequential

data, which is common in natural language processing tasks. Dialects often have unique patterns and sequences in their phonetic, grammatical, and lexical structures that can be effectively captured by LSTMs. Our idea is to compare the architectures of CNN and LSTM networks on the task of classifying Arkhangelsk dialects, given that the dataset is limited in size.

In the present study, we employ a corpus of dialect data from the Arkhangelsk region, which has been collected since the 1960s as part of a larger dialectological fieldwork project in the area and is currently archived in a card catalogue.

The aim of this paper is to develop a dialect words classifier, which can be used to identify and categorize dialect words within a given text. The novelty of this research lies in the lack of an automated system for classifying dialect words based on Arkhangelsk dialect materials. The practical significance of this study lies in the possibility of applying the technique to analyze dialect material collected not only in the Arkhangelsk region, but also in other areas, provided that the necessary data is gathered and the neural network is fine-tuned accordingly.

2. Methods

In the study, we employed such methods as analysis and synthesis. We analyzed various components of convolutional neural networks (CNNs) and long short-term memory (LSTM) neural networks, including LSTM, convolutional, max pooling, flatten, and other layers, as well as the connections between them. We also examined their architectures and functional characteristics. Based on this analysis, we synthesized these components to develop a model that can identify and classify dialect words.

Furthermore, the study incorporates a computer-based experiment, which entails the utilization of software and models for the simulation of actual processes and systems. In the study, we conducted experiments using various neural network architectures in order to determine which combinations yield the

most effective results for the detection and classification of dialect words. Next, the machine learning technique was applied to specific models and subsequently tested to evaluate the quality of the results.

The results of the experiment were subjected to statistical analysis and evaluated using a range of metrics to assess the quality of the model, including accuracy, precision, and F1 score. A comparative method was employed to evaluate the performance metrics of various neural network models, with a view to determining their relative effectiveness in the task of identifying and classifying dialect words.

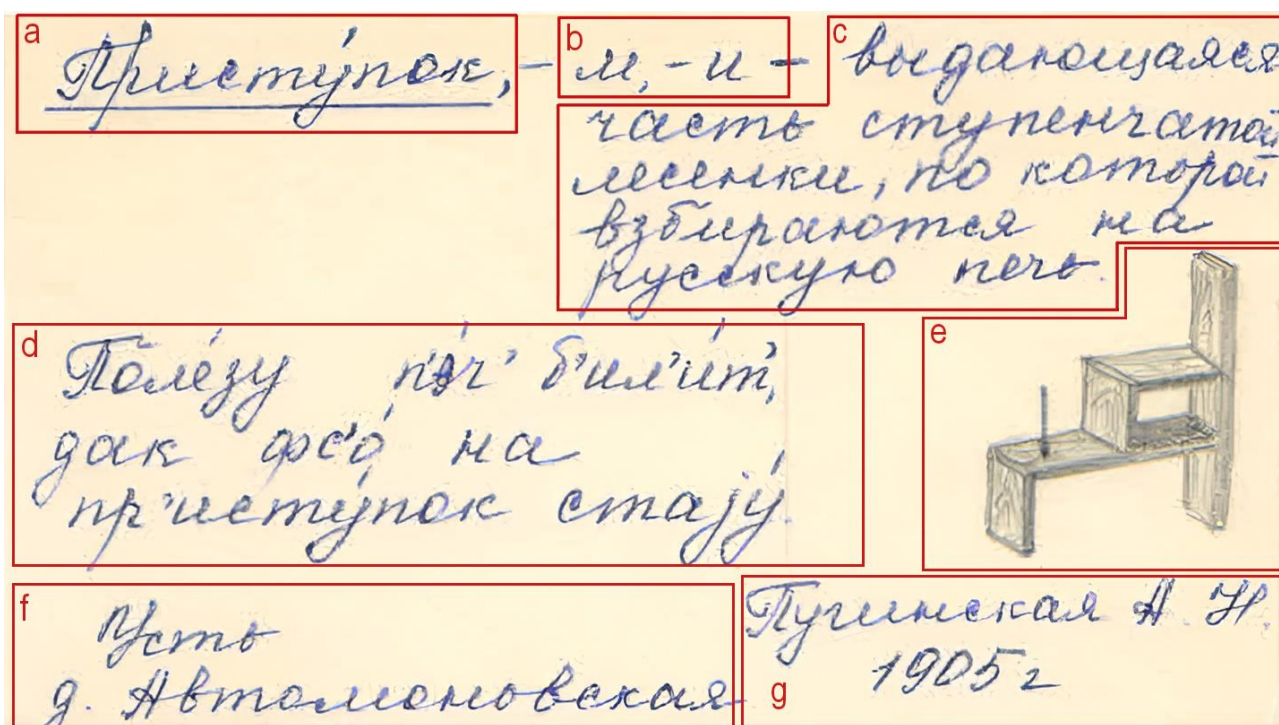
2.1 Dataset

The data collected during dialectological fieldwork is stored in the form of a card index, which consists of specially designed catalogue cards. Each dialect card includes a specific word, along with any necessary emphasis, softening, or brevity markers. Additionally, where possible, there are examples of the word's usage, an explanation of its meaning, the geographical location in the region where it was recorded and some other relevant information.

Examples of dictionary cards are illustrated in Figures 1 and 2.

The card features the word "Pristupok" (Figure 1a, Russian: «Приступок» which means "Step") and provides essential information regarding its grammatical characteristics (Figure 1b), namely, that it is a masculine noun in the nominative case. The definition is: "the prominent part of the staircase, which leads to the Russian stove" (Figure 1c). There is an example of usage "when going to climb the stove to whitewash, then stand on the step" (Figure. 1, d, Russian: «Полезу печь белить, так все на приступок (в)стаю»). The card was recorded in the village of Avtomonovskaya, Ustyansky District (Figure 1e), according to informant Puginskaya A.N., born 1905 (Figure 1f). The text on the card is accompanied by an image of the subject matter (Figure 1g), which was created by a collector of local dialect material.

Figure 1. Dictionary card from the Ustyansky District. Sections: a – word, b – explanation, c – grammatical features, d – example, e – illustration, f – location, g – dialect card source
Рисунок 1. Диалектная карточка из Устьянского района. Разделы: а – слово, б – толкование, с – грамматические характеристики, d – пример употребления, е – иллюстрация, f – место сбора, g – источник диалектной карточки



The same word can be used in different areas and, accordingly, have different meanings. Figure 2 illustrates the same word “Pristupok” with a different meaning, as the material was collected in a different area of the region.

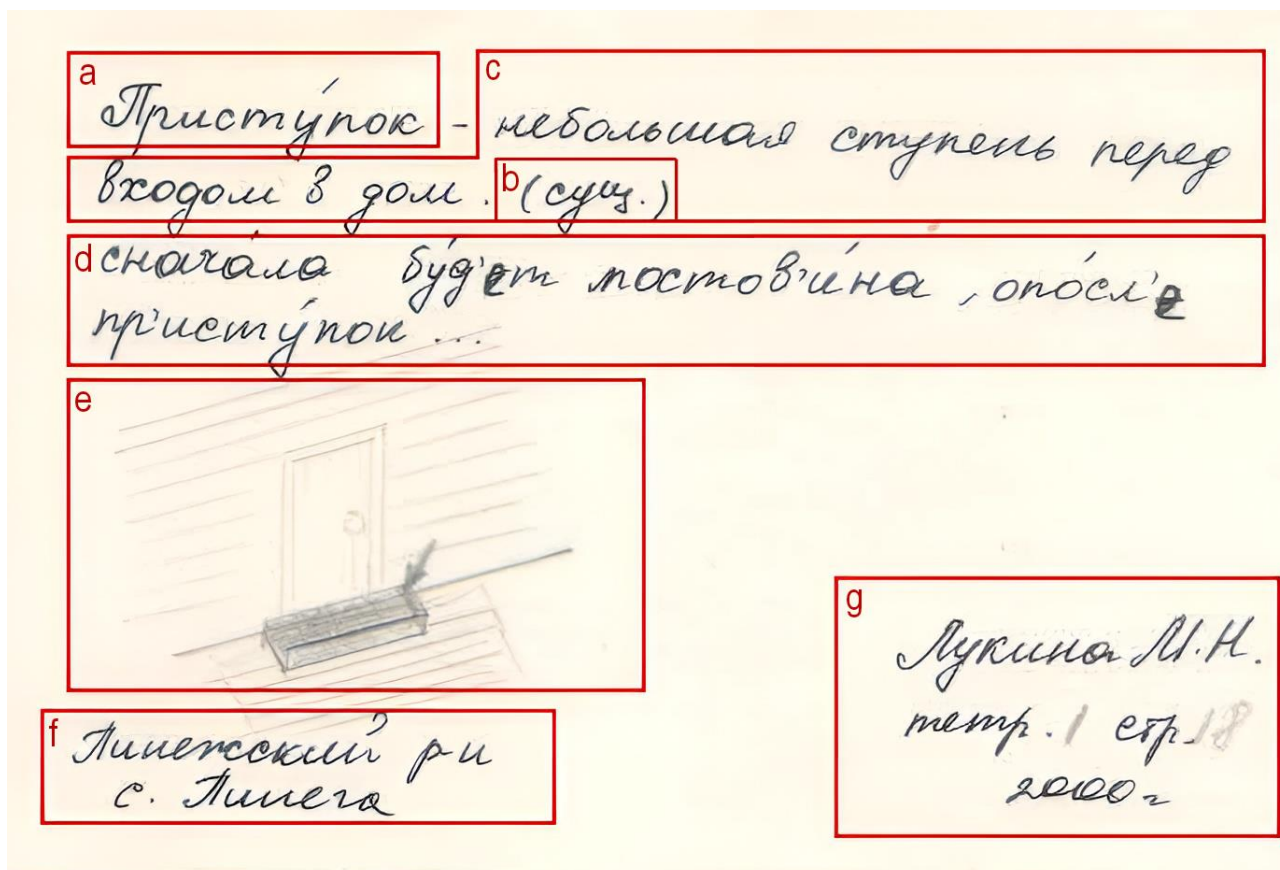
It may be noted that the cards have a similar overall design, although there is no clearly defined arrangement or organization of fields. The specific details may vary slightly from one instance to another. For example, the first card contains the personal details, such as their name and date of birth, while the second card includes information about the collector and the date when the material was collected. Also there are differences among the specified grammatical characteristics. In the first card, the grammatical gender and case of the given word are specified, while the second card only indicates the part of speech of the given word. Some cards contain the complete names

of districts and human settlements, whereas others contain abbreviated versions. Several cards include illustrations of the relevant objects, while the majority of cards do not contain any illustrations.

All of the above factors make it challenging to automate card processing through the use of computer vision technologies.

The current dataset of dialect materials was manually assembled using a custom-developed application called “Kopilka slov” (Russian: «Копилка слов» which means “A Piggy Bank of Words”) (Shurykina, Latukhina, 2023). The application and its operation are described in greater detail in (Shurykina, Latukhina, 2024). The collected words are also incorporated into the materials of the dictionary (Nenasheva, 2023). The dictionary and its digital counterpart are described in further detail in (Nenasheva, Shurykina, 2024).

Figure 2. Dictionary card from the Pinezhsky District. Sections: a – word, b – explanation, c – grammatical features, d – example, e – illustration, f – location, g – dialect card source
Рисунок 2. Диалектная карточка из Пинежского района. Разделы: а – слово, б – толкование, с – грамматические характеристики, d – пример употребления, е – иллюстрация, f – место сбора, g – источник диалектной карточки



A total of 1,387 dialect words have been manually classified by experts. For the purposes of classification, we have identified four categories of dialect words based on their distinctive features: phonetic, grammatical, derivational or word-formational, lexical. Researchers use similar categories when solving practical problems. For instance, in (Karbysheva, Radchenko, 2020), dialect words are translated into a foreign language using the similar categories within that language.

Let us examine the four categories in more detail:

– *Phonetic* – dialect words that reflect the characteristics of the sounds in the dialect’s phonetic system, capture the nuances

in pronunciation that differentiate one dialect from another. There are 293 phonetic units.

– *Grammatical* – dialect words that highlight the peculiarities of the dialect grammatical structure, which include variations in syntax, morphology, and other grammatical features that set the dialect apart from the standard language (118 grammatical dialect units).

– *Derivational or word-formation* – dialect words that differ in its structure from literary language words and involve unique word-formation processes that are specific to the dialect, such as affixation, compounding, and other morphological changes (561 derivational units).

– *Lexical* – local terms for subjects and phenomena specific to the dialect (415 lexical units).

It can be observed that the classes are not balanced. It is not feasible to balance the sample due to the following reasons:

– Reducing the size of the classes would result in a significant loss of information, which is already scarce.

– Increasing the sample size is challenging given the lack of available data.

– Generating new dialect data for the purpose of expanding a dataset can be challenging for several reasons.

Let us examine these reasons in greater detail.

Firstly, dialects frequently exhibit unique lexical, phonetic, and grammatical characteristics that are challenging to replicate using automated methods. These features may include specific vocabulary, phrases, pronunciation patterns, and grammatical structures that are not easily formalized. Secondly, the presence of dialect features in text is often contingent upon contextual and cultural factors, the influence of which can be challenging to account for in the process of text generation. For instance, certain phrases or words may only be used in specific contexts or social settings. Thirdly, a significant amount of high-quality data is required in order to generate high-quality examples of dialect text. Collecting a sufficient number of examples of dialect text to train a model may be a challenging task. Additionally, dialects may vary significantly even among members of the same region or social group. Thus, in (Nenasheva, 2021) several names are provided that are used to refer to the concept of “clasp” (Russian: «застежка») in the Russian vernacular: “zástebka, záshhepka, zastezhnica, zaporina, zaporinka, babka, gápel’ka, gáplik, záboloka, ostebka” (Russian: «зástebka, záщепка, застежница, запорина, запоринка, бабка, гáпелька, гáплик, záболока, остebка») and others. Examples of dialect words that are used to refer to word “button” that are homonyms of commonly used words with

different meanings: “blyashka” (Russian: «бляшка») with meaning of “metal plate” or “plaque on the surface of the skin, tissues, and walls of blood vessels”), “bant” (Russian: «бант») with meaning of “decoration made of ribbon, braid, etc., knotted with loosely released loops or fastened with a buckle”), “bantochka” (Russian: «банточка»), “kostyashka” (Russian: «костяшка») with meaning of “protruding joint” or “small bone product”), “kostyanka” (Russian: «костянка»), “plashchik” (Russian: «плащик») is a diminutive of “raincoat”), “puga, pugitsa, puglitsa, pugolka, puvichka, pubochka” (Russian: «пуга, пугица, пуглица, пуголка, пувичка, пубочка»). However, the words for buttons may not apply to all buttons as a whole. The word “klyapyshek” (Russian: «кляпышек») is used in Vladimir dialects to refer to buttons on a coat or jacket, not on a shirt. “Kuzik” (Russian: «кузик») in Smolensk, Bryansk, Kursk dialects means “a metal button”. “Mastashka” in Voronezh dialects – “a bone button”. Dialect diversity and variation in word usage are also observed in other Russian dialects (Smetanina, Ivanova, 2020). Finally, dialects may differ significantly from the standard form of a language, even to the extent that some argue for considering them as separate languages (Mutalov, 2020; Purtova, 2023).

Assessing the quality of generated dialect text may also be challenging, as it requires appropriate qualifications and expertise in dialects. In other words, it is necessary to involve experts or native speakers of the target language in order to evaluate the accuracy and fluency of the generated content, which can be a laborious and costly process.

2.2 Neural networks architectures

Following architectures are preferable for natural language processing (Ye et al., 2019):

- transformers;
- networks with long short-term memory;
- convolutional neural networks.

Transformer neural networks are designed for processing full-length text fragments and may be challenging to implement. Additionally, transformers require large volumes of text for training (Laith, Kang, 2023; Devlin et al., 2019). Neural network models based on long short-term memory (LSTM) are better suited for processing sequences of text, as they function effectively with sequential data, rather than individual words. Convolutional neural networks (CNNs) can process words at the character level, taking into account the position of each character within a word. Furthermore, CNNs have a simpler architecture and require less computational resources compared to LSTM networks, making them potentially more convenient for use in certain applications (Sainath et al., 2015; Ye et al., 2019).

It should be noted that the lack of contextual information in the problem-solving process is a limitation, but the type of word does not depend heavily on context. In these cases, a convolutional neural network can be

used, which works directly with the structure of the word without relying on its context.

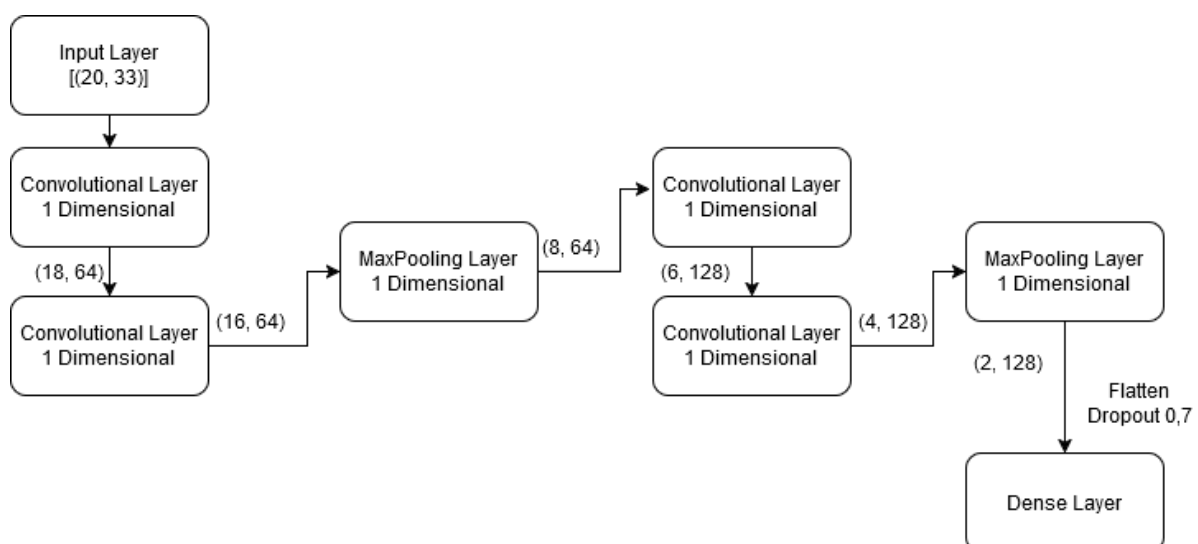
There are two methods for converting words into a vector representation: unigram encoding and bigram encoding. Unigram encoding involves creating a dictionary in which each character is mapped to a vector of zeros and ones. In contrast, bigram encoding uses pairs of consecutive characters. The use of bigram encoding allows for the consideration of combinations of characters, enabling us to obtain more information about a word's structure and features. However, the data encoded using bigrams may contain redundant information and be more challenging to process.

To determine the most suitable neural network for our purpose, we developed, trained, and compared four different neural network models. Specifically, we utilized convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, each operating on both unigrams and bigrams.

The CNN architecture used in this study is illustrated in Figure 3.

Figure 3. Convolutional neural network architecture

Рисунок 3. Архитектура сверточной нейронной сети



The convolutional neural network comprises an input layer, a series of one-dimensional convolutional layers (Conv1D), and a series of max-pooling layers

(MaxPooling1D). There is also a fully connected layer that operates after the data has been flattened into a one-dimensional

array (Flatten) and passed through a dropout (Dropout) layer.

The input to the network is a sequence of words from a dataset, each of which has been transformed into a two-dimensional matrix with dimensions equal to the maximum possible word length and the size of the dictionary. We assume that the maximum word length is 20 characters. The maximum word length in the provided dataset is 14 characters. For the sake of model scalability, six additional characters have been added.

The size of the vocabulary depends on the specific approach used.

For the unigram approach, the dictionary comprises the Russian alphabet without the letter «ё», and a padding symbol. Each letter in the word is represented as a vector of length 33, where each position corresponds to one of the dictionary characters. For each letter, the position corresponding to this letter has a value of 1, and all other positions have a value of 0. For the bigram approach, the dictionary contains 1156 combinations of characters (the same 32 letters, specific start and end tokens, 34 squared because of multiplication principle), and a padding symbol. In this case, each letter

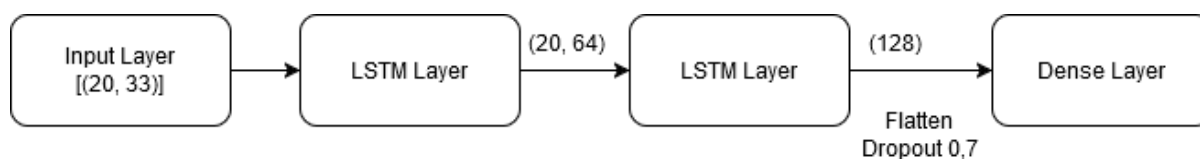
in the word is also presented as a vector, but with a length of 1157.

Convolutional layers use a filter to emphasize significant aspects of the input data. This filter is a one-dimensional array of weights that slides over the input data, performing a convolutional operation. During this process, the filter multiplies each element of the input data by the corresponding weight, and then the resulting products are summed to form an output value for each filter position. During the training of the neural network, the parameters of the filter are tuned to optimize its performance. We have used the ReLU activation function on these layers as the most effective (Ramachandran et al., 2017). The Max Pooling layer reduces the dimensionality of the data while retaining the most significant features. It takes the maximum value from each section of one-dimensional data and passes it to the next layer. The Flatten layer transforms the data obtained in the previous stages into a one-dimensional representation. After that we remove the least significant ones using the Dropout layer and predict the outcome on a fully connected layer.

The network implemented with long short-term memory architecture is depicted in Figure 4.

Figure 4. LSTM neural network architecture

Рисунок 4. Архитектура нейронной сети с долгосрочной кратковременной памятью



The long short-term memory network (LSTM) consists of two layers, each containing a long short-term memory unit. These units are specialized for the efficient processing of sequential data, as opposed to the convolutional layers and max-pooling techniques used in convolutional neural networks. The main components of an LSTM unit include:

1. The cell state maintains information transmitted across the network and is updated

at each step, based on the input data and the current status of the gates.

2. The input gate determines which portion of the new information is added to the memory unit.

3. The forget gate decides which portion of information from the previous state of the memory unit should be erased.

4. The output gate controls which portion of information from the memory unit will be transferred to the subsequent stage.

LSTM layers allow you to save important data and forget unnecessary ones, which makes them a powerful tool for processing sequential data, including text (Sainath et al., 2015).

3. Results and Discussion

The following metrics are employed to assess the quality of models: accuracy, precision, and F1-score.

Accuracy is a measure that indicates the proportion of correct classifications among all examples. It reflects the model's ability to predict classes. The closer the accuracy is to 1, the better the model's predictive ability.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP (True Positive) is the number of correctly classified positive instances, in this case, TP refers to the number of dialect words that were correctly assigned to a specific class, for which accuracy is measured,

TN (True Negative) is the number of correctly classified negative instances, in this case, TN is the number of dialect words that were correctly assigned to any class other than the specific class, for which accuracy is measured,

FP (False Positive) is the number of incorrectly classified positive instances, in this case, FP is the number of words that were incorrectly assigned to a specific class, for which accuracy is measured,

FN (False Negative) is the number of incorrectly classified negative instances, in this case, FN is the number of words that were incorrectly assigned to any class other than the specific class, for which accuracy is measured.

Accuracy allows us to evaluate the overall performance of the model, particularly when the classes in the dataset are balanced. If the classes are not well-balanced, accuracy may be an inaccurate measure of performance, as it does not take into account the uneven distribution of the different classes. For instance, if there are 100 dialect

words in a dataset used to evaluate a model, and 90 of these belong to the word-formation category and 10 belong to the phonetic category, the accuracy metric will be 0.9 even if the model assigns all words to the word-formation category.

To determine how infrequently the model makes errors in incorrectly classifying objects, a precision metric is utilized. Precision is a metric that measures the proportion of accurately classified positive instances from the total number of classified positive instances. In our study, it is the number of dialect words correctly assigned by the model to a particular class, divided by the total number of words assigned to that class by the model.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the proportion of correct positive classifications among all positive instances, i.e., the number of dialect words correctly assigned to a specific class by the model, divided by the total number of dialect words in that class.

$$Recall = \frac{TP}{TP + FN}$$

Recall differs from Precision in that Precision considers the positive instances that are correctly classified by the model, whereas Recall assumes the use of true classes as defined by experts.

The F1-score is a metric that represents the harmonic mean between Precision and Recall. It takes into consideration both the Precision and the Recall of a model, making it a useful tool for evaluating its performance, particularly when the classes in the dataset are not balanced.

$$F1_{Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The results of the model training are presented in Table 1. The table displays the values of accuracy, precision, recall and F1-score metrics on the test dataset, as well as the training time.

Table 1. Model training results

Таблица 1. Результаты обучения моделей

Architecture	Accuracy	Precision	Recall	F1-score	Time of training, s.
CNN	0.39	0.69	0.58	0.63	33.8
LSTM	0.35	0.65	0.58	0.61	146.7
CNN, bigrams	0.34	0.62	0.38	0.47	42.2
LSTM, bigrams	0.43	0.65	0.44	0.52	174.2

The unigram convolutional neural network model demonstrates the highest level of accuracy. It also trained in the shortest period with a sufficiently high level of F1-score, which reflects both precision and recall.

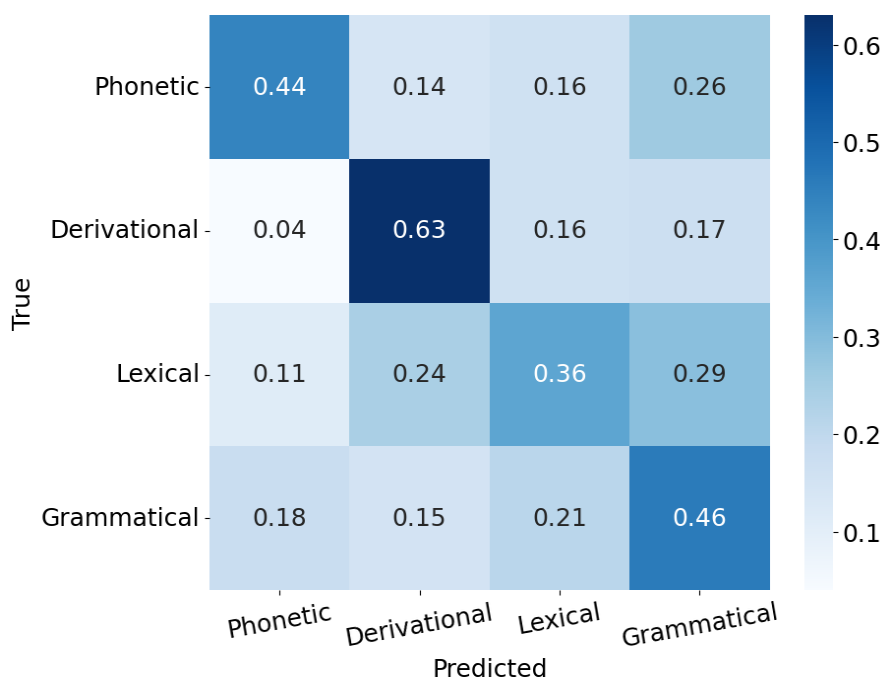
The confusion matrix is a valuable tool for assessing the performance of a classifier. It demonstrates how frequently the classifier misclassifies objects from one class as belonging to another. The confusion matrix is a table where the rows correspond to the true classes and the columns correspond to the predicted classes. The values in the matrix

cells represent a subset of the instances that were assigned by the neural network to the column class, despite the fact that they actually belong to the row class. Therefore, the values along the main diagonal represent the proportion of instances that the neural network has classified correctly. The higher the calculated value of the fraction, the greater the intensity of color in the cell.

Figure 5 illustrates the confusion matrix for the unigram convolutional neural network model.

Figure 5. Confusion matrix

Рисунок 5. Матрица ошибок



Let us consider the example of word-formation dialect words. The lightest cell in the intersection of the Derivational row and Phonetic column, with a value of 0.04, indicates that only 4% of word-formation dialect words were erroneously classified as Phonetic. On the other hand, the darkest cell, which is situated in the same row as the intersection with the Derivational column, represents the proportion of instances that the model classified correctly.

The model demonstrates the best performance in the largest class – derivational class (63%). The accuracy for other classes is lower, but it is still possible to determine their categorical membership with a reasonable degree of certainty (44% for phonetic category and 46% for grammatical category).

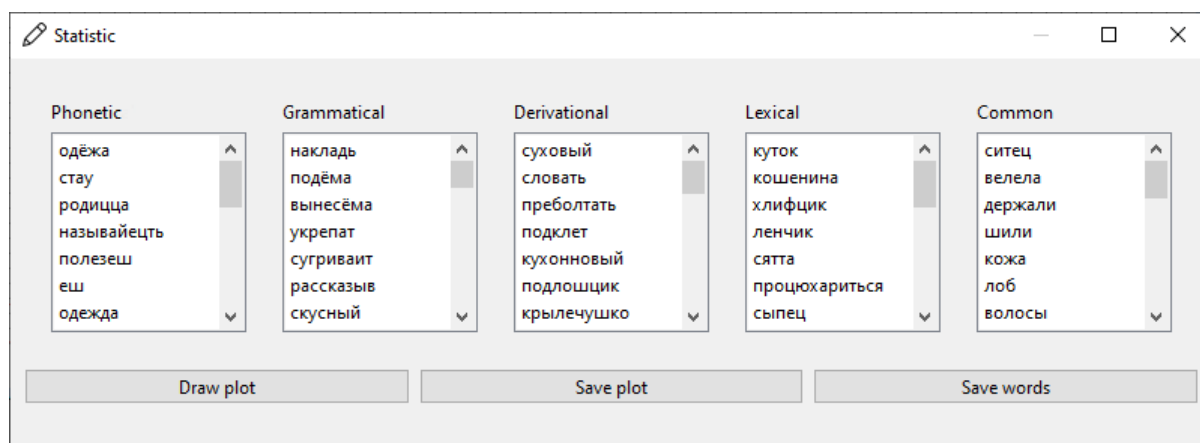
Most often, the classifier misclassifies phonetic dialect words with grammatical (26%), word-formation dialect words with lexical and grammatical features (17% each), lexical dialect words with grammatical dialect words (29%), and grammatical dialect words with phonetic dialect words (18%). The

majority of the errors are due to the classifier’s tendency to assign words to a particular word-formation category. This is caused by the predominance of word-formation class objects in the training dataset.

The trained unigram convolutional neural network model has been integrated into the developed application in order to facilitate the process of working with dialect variations. This integration is described in greater detail in (Shurykina, Latukhina, 2024). A researcher enters a list of dialect words into the application and initiates the classification process. As a result, the dialect variants are separated into multiple categories, each category corresponding to a previously defined class. The last category is reserved for words that the CNN was unable to classify unambiguously. If necessary, the researcher can subsequently rearrange the position of words in the final category by dragging them to their desired location. The resulting categorization can then be saved for future reference (Figure 6).

Figure 6. Graphical interface for the dialectism classifier

Рисунок 6. Графический интерфейс для классификатора диалектизмов

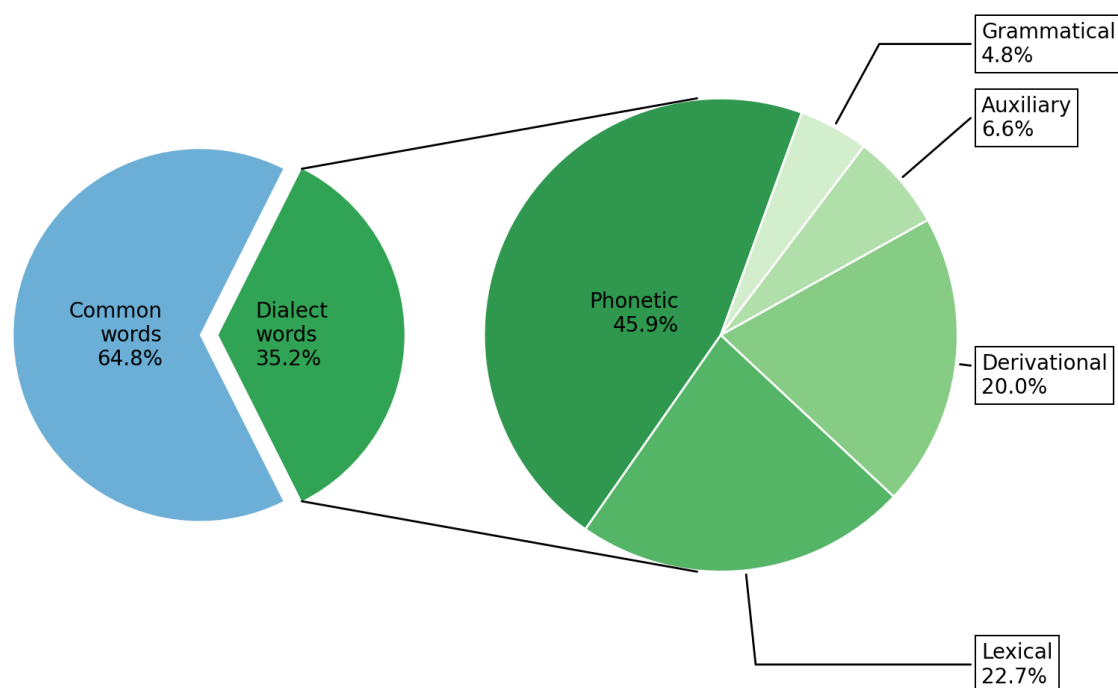


Based on the classification results, we can create a pie chart that demonstrates the proportion of words in each class based on the

final distribution. Figure 7 illustrates an example of such a pie chart.

Figure 7. Pie chart of classes distribution

Рисунок 7. Круговая диаграмма распределения классов



The above pie chart shows that slightly more than a third of the words are dialectal in the collected dataset. Almost half of these are phonetic dialect words, 22% of dialect words belong to the lexical category and 20% are derivational, least of all (4.8%) are grammatical dialect words. Also 6% of dialect words are auxiliary.

The trained model has been published on GitHub¹.

4. Conclusions

Neural networks have been widely used for the identification and classification of dialect words; however, no relevant research has been conducted on the material from the Arkhangelsk dialect before.

The classifier developed as a result of the research makes it possible to automate the processes of identifying and classifying dialect words with an accuracy of up to 63% for the derivational category. The accuracy

can be further improved through additional training on more dialect data. It is also possible to enhance the recognition of dialect words from the grammatical perspective. Grammatical dialect words are words of the literary language that conform to an incorrect pattern (e.g., «мыш», «ребенков», «польта»). Therefore, they can be detected using existing morphological analysers, such as pymorphy2. The word distribution produced as a result of classifier operation can be applied for further dialect investigation or for model training.

It is planned to apply the above approaches in future research. The expansion of the dataset by incorporating dialect material will represent the most significant enhancement. It is also planned to further experiment with more advanced neural network architectures. The solution can be extended by training the model with dialect materials from various regions.

The developments described in the paper, and similar ones, will significantly reduce the amount of manual labour required

¹ Arkhangelsk Dialect Classifier (2024), available at: https://github.com/oat4cat/arkh_dialect_classifier (Accessed: 05.12.2024).

for the pre-processing of dialect research material.

References

- Adel, B., Eddine, M. C., Laouid, A., Chait, K. and Kara, M. (2024). Using Transformers to Classify Arabic Dialects on Social Networks, *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, El Oued, Algeria, 1–7. DOI: 10.1109/PAIS62114.2024.10541289 (In English)
- Alali, M., Sharef, N., Murad, M. and Husin, N. A. (2019). Narrow Convolutional Neural Network for Arabic Dialects Polarity Classification, *IEEE Access*, 7. DOI: 10.1109/ACCESS.2019.2929208 (In English)
- Arkhangelskiy, T. A. (2021). Application of the dialectometric method to the classification of Udmurt dialects, *Uralo-altaiskie issledovaniya*, 2(41), 7–20. DOI 10.37892/2500-2902-2021-41-2-7-20 (In Russian)
- Azim, M. A., Hussein, W., Badr, N. (2022). Automatic Dialect identification of Spoken Arabic Speech using Deep Neural Networks, *International Journal of Intelligent Computing and Information Sciences*, 22, 4, 25–34. DOI: 10.21608/ijicis.2022.152368.1207 (In English)
- Buckley, K. (2021). Uncovering linguistic lineage through using a character N-gram based dialect classifier, *The languages of Scotland and Ulster in a global context, past and present. Selected papers from the 13th triennial Forum for Research on the Languages of Scotland and Ulster*, Munich, Germany, 139–76. (In English)
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint*. DOI: 10.48550/arXiv.1810.04805 (In English)
- Han, M., Zhu, D., Wen, X., Shu, L. and Yao, Z. (2024). Research on Dialect Protection: Interaction Design of Chinese Dialects Based on BLSTM-CRF and FBM Theories, *IEEE Access*, 12, 22059–22071. DOI: 10.1109/ACCESS.2024.3364098 (In English)
- Høyland, B. and Nesse, A. (2023). Norwegian Dialect Classifications, *Dialectologia*, 10, 255–298. DOI: 10.1344/Dialectologia2022.2022.10 (In English)
- Huang, T. J., Yang, J. Q., Shen, C., Liu, K. Q., Zhan, D. C. and Ye, H. J. (2024). Improving LLMs for Recommendation with Out-Of-Vocabulary Tokens?, *arXiv preprint*. DOI: 10.48550/arXiv.2406.08477 (In English)
- Karbysheva, D. Y. and Radchenko G. I. (2020). Types of dialectisms and methods of their translation into a foreign language (based on the novel M.A. Sholokhov's 'Quiet Don'), *Eurasian Scientific Union*, 8-5(66), 294–297. (In Russian)
- Kethireddy, R., Kadiri, S. and Gangashetty, S. (2022). Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations, *The Journal of the Acoustical Society of America*, 151, 1077–1092. DOI: 10.1121/10.0009405 (In English)
- Kolkova, D. E. (2023). Self-identification of personality through the use of dialect (based on the example of the Scottish dialect), *Creative Linguistics: Collection of Scientific Articles*, 6, 106–111. (In Russian)
- Kornaukhova, T. V. and Goloshtanova, A. A. (2022). Reflection of modern realities in the dialects of the English language (based on the example of the Cockney dialect), *Proceedings of the All-Russian scientific and practical conference "Avdeev Readings"*, Penza, Russia, 90–94. (In Russian)
- Kositsina, Y. (2024). Dialectisms in the modern regional dialect of the village of Usmanka, Chebulinsky District, Kemerovo Region, *Philology. Theory Practice*, 17, 1577–1583. DOI: 10.30853/phil20240228 (In Russian)
- Kuparinen, O. (2024). Murre24: Dialect Identification of Finnish Internet Forum Messages, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 12003–12015. (In English)
- Laith, B. and Kang, S. (2023). Transformer Text Classification Model for Arabic Dialects That Utilizes Inductive Transfer, *Mathematics*, 11, 4960. DOI: 10.3390/math11244960 (In English)
- Mutalov, R. O. (2020). On the problem of distinguishing Dargin languages and dialects, *The Newman in Foreign Policy*, 6, 57 (101), 6–8. (In Russian)
- Nenasheva, L. V. (2021). Each garment has its own clasp, *Cuadernos De Rusística Española*, 17, 211–221. DOI: 10.30827/cre.v17.21023 (In Russian)
- Nenasheva, L. V. (2023). *Tematicheskii slovar arkhangel'skih govorov* [Thematic dictionary of Arkhangelsk dialects], Arkhangelsk:

Limited Liability Company “Consulting Information Advertising Agency”, Arkhangelsk, Russia. (*In Russian*)

Nenasheva, L. V. and Shurykina, L. S. (2024). Electronic dictionary of Arkhangelsk dialects, *Arctic and North*, 55, 243–252. DOI: 10.37482/issn2221-2698.2024.55.243. (*In Russian*)

Purtova, G. M. (2023). Meyankieli: dialect or language?, *Proceedings of the International scientific and practical conference “The World Historical and Cultural Heritage of the Arctic”*, Saint-Petersburg, Russia, 27–28. (*In Russian*)

Ramachandran, P., Zoph, B., and Le, Q.V. (2017). Searching for Activation Functions, arXiv:1710.05941. DOI: 10.48550/arXiv.1710.05941 (*In English*)

Sainath, T. N., Vinyals, O., Senior, A. and Sak, H. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks, *IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, Australia, 4580–4584. DOI: 10.1109/ICASSP.2015.7178838 (*In English*)

Samsitova, L. H. (2020). Dialect as a reflection of the linguistic picture of the world (using the example of the Northwestern dialect of the Bashkir language), *Mir nauki, kul'tury, obrazovaniya*, № 6 (85), 474–476. DOI: 10.24412/1991-5500-2020-685-474-476 (*In Russian*)

Sciarretta A. (2024). Dialectometry-based classification of the Central–Southern Italian dialects, *Journal of Linguistic Geography*, 12 (1), 13–23. DOI: 10.1017/jlg.2024.7 (*In English*)

Shamshin, A. L. (2024). The role of knowledge of Italian dialects in intercultural communication: their importance for successful adaptation in Italy, *Proceedings of VIII International scientific and methodological conference “Problems of teaching philological disciplines to foreign students”*, Voronezh, Russia, 221–225.

Shurykina, L. S. and Latukhina, E. A. (2023). Certificate of State Registration of the Computer Program No. 2023668038, 22 Aug 2023. (*In Russian*).

Shurykina, L. S. and Latukhina, E. A. (2024). Automated creation of dialect dictionaries organization, *Current Problems of Applied Mathematics, Informatics and Mechanics*, Voronezh, Russia, 1017–1022. (*In Russian*)

Smetanina, Z. V. and Ivanova, G. A. (2020). The variation of the word in the “Regional dictionary of Vyatka dialects”, *Vestnik Tomskogo gosudarstvennogo universiteta*, 451, 56–68. DOI: 10.17223/15617793/451/8 (*In Russian*)

Themistocleous, C. (2017). Dialect classification using vowel acoustic parameters. *Speech Communication*, 92, 13–22. (*In English*)

Themistocleous, C. (2019). Dialect Classification From a Single Sonorant Sound Using Deep Neural Networks, *Frontiers in Communication*, 4. DOI: 10.3389/fcomm.2019.00064 (*In English*)

Vernyaeva, R. A. and Zhdanova, E. A. (2023). Multimedia Corpus of Russian Dialects of Udmurtia: Electronic Subcorpus of Spoken Speech. *Cuadernos De Rusística Española*, 19, 67–79. DOI: 10.30827/cre.v19.28131 (*In Russian*)

Yamani, A., Alziyady, R., AlYami, R., Albelali, S., Albelali, L., Almulhim, J., Alsulami, A., Alfarraj, M., and Al-Zaidy, R. (2024). The kind dataset: A social collaboration approach for nuanced dialect data collection, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 32–43. (*In English*)

Ye, S., Zhao, R. and Fang, X. (2019). An Ensemble Learning Method for Dialect Classification, *IOP Conference Series: Materials Science and Engineering*, 569 052064. DOI: 10.1088/1757-899X/569/5/052064 (*In English*)

Zhang, Y. and Ren, W. (2022). From *hǎo* to *hǒu* – stylising online communication with Chinese dialects, *International Journal of Multilingualism*, 21 (1), 149–168. DOI: 10.1080/14790718.2022.2061981 (*In English*)

Список литературы

Adel B. Using Transformers to Classify Arabic Dialects on Social Networks / Adel B., Eddine M. C., Laouid A., Chait K., Kara M. // 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), El Oued, Algeria. 2024. Pp. 1–7. DOI: 10.1109/PAIS62114.2024.10541289

Alali M., Sharef N., Murad M. et al. Narrow Convolutional Neural Network for Arabic Dialects Polarity Classification // IEEE Access. 2019. № 7. DOI: 10.1109/ACCESS.2019.2929208

Архангельский Т. А. Применение диалектометрического метода к классификации удмуртских диалектов // Урало-алтайские исследования. 2021. № 2 (41).

C. 7–20. DOI 10.37892/2500-2902-2021-41-2-7-20.

Azim M. A., Hussein W., Badr N. Automatic Dialect identification of Spoken Arabic Speech using Deep Neural Networks // *International Journal of Intelligent Computing and Information Sciences*. 2022. DOI: 10.21608/ijicis.2022.152368.1207

Buckley K. Uncovering linguistic lineage through using a character N-gram based dialect classifier // *The languages of Scotland and Ulster in a global context, past and present. Selected papers from the 13th triennial Forum for Research on the Languages of Scotland and Ulster*, Munich, Germany. 2021. Pp. 139 Pp.76.

Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Devlin J., Chang M.-W., Lee K., Toutanova K. // *arXiv preprint*. 2019. DOI: 10.48550/arXiv.1810.04805

Han M. Research on Dialect Protection: Interaction Design of Chinese Dialects Based on BLSTM-CRF and FBM Theories / Han M., Zhu D., Wen X., Shu L., Yao Z. // *IEEE Access*. 2024. № 12. Pp. 22059–22071. DOI: 10.1109/ACCESS.2024.3364098.

Høyland B., Nesse A. Norwegian Dialect Classifications // *Dialectologia*. 2023. № 10. Pp. 255–298. DOI: 10.1344/Dialectologia2022.2022.10.

Huang, T. J., Yang, J. Q., Shen, C., Liu, K. Q., Zhan, D. C. and Ye, H. J. (2024). Improving LLMs for Recommendation with Out-Of-Vocabulary Tokens. DOI: 10.48550/arXiv.2406.08477

Карбышева Д. Ю., Радченко Г. И. Типы диалектизмов и способы их перевода на иностранный язык (на материале романа М.А. Шолохова «Тихий Дон») // *Евразийское Научное Объединение*. 2020. № 8–5 (66). С. 294–297.

Kethireddy R., Kadiri S. and Gangashetty S. Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations // *The Journal of the Acoustical Society of America*. 2022. № 151. Pp. 1077–1092. DOI: 10.1121/10.0009405.

Колкова Д. Е. Самоидентификация личности посредством использования диалекта (на примере шотландского диалекта) // *Креативная лингвистика: сборник научных статей*. 2023. № 6. С. 106–111.

Корнаухова Т. В., Голоштанова А. А. Отражение современных реалий в диалектах английского языка (на примере диалекта кокни) // *X Авдеевские чтения: Сборник статей по материалам Всероссийской научно-практической конференции*, Пенза. 2022. С. 90–94.

Косицина Ю. В. Диалектизмы в современном региолекте с. Усманка Чебулинского района Кемеровской области // *Филологические науки. Вопросы теории и практики*. № 17. С. 1577–1583. DOI: 10.30853/phil20240228.

Kuparinen O. Murre24: Dialect Identification of Finnish Internet Forum Messages // *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024. Pp. 12003–12015.

Laith B., Kang S. Transformer Text Classification Model for Arabic Dialects That Utilizes Inductive Transfer // *Mathematics*. 2023. № 11. 4960. DOI: 10.3390/math11244960.

Муталов Р. О. К проблеме разграничения даргинских языков и диалектов // *The Newman in Foreign Policy*. 2020. Т. 6. № 57 (101). С. 6–8.

Ненашева Л. В. Для каждой одежды своя застежка. // *Cuadernos De Rusística Española*. № 17. С. 211–221. DOI: 10.30827/cre.v17.21023

Ненашева Л. В. Тематический словарь архангельских говоров. Архангельск: Общество с ограниченной ответственностью «Консультационное информационно-рекламное агентство», 2023. 192 с.

Ненашева Л. В., Шурыкина Л. С. Электронный словарь архангельских говоров // *Арктика и Север*. 2024. № 55. С. 243–252. DOI: 10.37482/issn2221-2698.2024.55.243

Пуртова Г. М. Меянкиели: диалект или язык? // *Мировое историко-культурное наследие Арктики: Тезисы Международной научно-практической конференции*, Санкт-Петербург. 2023. С. 27–28.

Ramachandran P., Zoph B., Le Q. V. Searching for Activation Functions // *arXiv preprint*. 2017. DOI: 10.48550/arXiv.1710.05941

Sainath T. N. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks / Sainath T. N., Vinyals O., Senior A., Sak H. // *IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, Australia. 2015. 4580–4584. DOI: 10.1109/ICASSP.2015.7178838.

Самситова Л. Х. Диалект как отражение языковой картины мира (на примере северо-западного диалекта башкирского языка) // Мир науки, культуры, образования. 2020. № 6 (85). С. 474–476. DOI: 10.24412/1991-5500-2020-685-474-476.

Sciarretta A. Dialectometry-based classification of the Central–Southern Italian dialects // Journal of Linguistic Geography. 2024. № 12(1). Pp. 13–23. DOI:10.1017/jlg.2024.7

Шамшин А. Л. Роль знания диалектов итальянского языка в межкультурной коммуникации: их важность для успешной адаптации в Италии // Проблемы преподавания филологических дисциплин иностранным учащимся: Сборник материалов VIII Международной научно-методической конференции, Воронеж. 2024. С. 221–225.

Свидетельство о государственной регистрации программы для ЭВМ № 2023668038 Российская Федерация. Программа для заполнения базы данных диалектных слов «Копилка слов»: № 2023667071: заявл. 15.08.2023: опубл. 22.08.2023 / Л. С. Шурыкина, Е. А. Латухина, Л. В. Ненашева; заявитель Федеральное государственное автономное образовательное учреждение высшего образования «Северный федеральный университет имени М.В. Ломоносова».

Шурыкина Л. С., Латухина Е. А. Организация автоматизированного создания диалектных словарей // Актуальные проблемы прикладной математики, информатики и механики: сборник трудов Международной научной конференции, Воронеж. 2024. С. 1017–1022.

Сметанина З. В., Иванова Г. А. Вариантность слова в «Областном словаре вятских говоров» // Вестник Томского государственного университета. 2020. № 451. С. 56–68. DOI: 10.17223/15617793/451/8.

Themistocleous C. Dialect classification using vowel acoustic parameters // Speech Communication. № 92. Pp. 13–22. (2017).

Themistocleous C. Dialect Classification From a Single Sonorant Sound Using Deep Neural Networks // Frontiers in Communication. 2019. № 4. DOI: 10.3389/fcomm.2019.00064.

Верняева Р. А., Жданова Е. А. Мультимедийный корпус русских говоров Удмуртии: электронный подкорпус устной

речи // Cuadernos De Rusística Española. № 19. С. 67–79. DOI: 10.30827/cre.v19.28131

Yamani A. The kind dataset: A social collaboration approach for nuanced dialect data collection / Yamani A., Alziyady R., AlYami R., Albelali S., Albelali L., Almulhim J., Al-Zaidy R. // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. 2024. С. 32–43.

Ye S., Zhao R., Fang X. An Ensemble Learning Method for Dialect Classification // IOP Conference Series: Materials Science and Engineering. 2019. 569 052064. DOI: 10.1088/1757-899X/569/5/052064.

Zhang Y., Ren W. From hǎo to hǒu – stylising online communication with Chinese dialects // International Journal of Multilingualism. 2022. № 21 (1). С. 149–168. DOI: 10.1080/14790718.2022.2061981

All authors have read and approved the final manuscript.

Все авторы прочитали и одобрили окончательный вариант рукописи.

Conflicts of interests: the authors have no conflicts of interest to declare.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Shurykina Lyudmila Sergeevna, Assistant, Department of Information Systems and Information Security, Northern (Arctic) Federal University named after M.V. Lomonosov, Arkhangelsk, Russia.

Шурыкина Людмила Сергеевна, ассистент кафедры информационных систем и информационной безопасности, Северный (Арктический) федеральный университет имени М.В. Ломоносова, Архангельск, Россия.

Latukhina Ekaterina Aleksandrovna, Senior Lecturer at the Department of Higher and Applied Mathematics, Northern (Arctic) Federal University named after M.V. Lomonosov, Arkhangelsk, Russia.

Латухина Екатерина Александровна, старший преподаватель кафедры высшей и прикладной математики, Северный (Арктический) федеральный университет имени М.В. Ломоносова, Архангельск, Россия.

Petrova Tatyana Viktorovna, Associate Professor, Candidate of Philological Sciences, Associate Professor of the Department of Russian Language and Speech Culture, Northern (Arctic) Federal University named after M.V. Lomonosov, Arkhangelsk, Russia.

Петрова Татьяна Викторовна, доцент, кандидат филологических наук, доцент кафедры русского языка и речевой культуры, Северный (Арктический) федеральный университет имени М.В. Ломоносова, Архангельск, Россия.