

Система полносвёрточного распознавания ваханского языка

А. Д. Корольков*, А. В. Шешкус**,

* ООО «Смарт Энджинс Сервис», г. Москва, Россия

** Федеральный исследовательский центр «Информатика и управление»
Российской академии наук, г. Москва, Россия

Поступила в редколлегию 10.10.2025 г. Принята 10.12.2025 г.

Аннотация—Задача распознавания ваханского языка имеет особое значение, так как ее решение позволит автоматизировать процесс оцифровки существующих данных, что в свою очередь поможет лингвистам работать над пониманием языка и над сохранением культурного наследия. В данной статье мы рассмотрим специфику ваханского языка и инструмент распознавания символов в виде нейронной сети. Для построения системы распознавания выбрана полносвёрточная нейронная сеть, так как она хорошо зарекомендовала себя при решении сходных задач, а в качестве создания обучающих данных в работе использован подход, основанный на генерации синтетических данных. Дополнительно обсуждаются перспективные направления использования разработанного подхода в смежных областях.

КЛЮЧЕВЫЕ СЛОВА: Ваханский язык, полносвёрточное распознавание, синтетические данные, глубокое обучение.

DOI: 10.53921/18195822_2025_25_4_888

1. ВВЕДЕНИЕ

Исчезновение языков — явление, напрямую связанное с динамикой государств: их возникновением, объединением и распадом. Языки доминирующих культур или политических образований вытесняли другие, в результате чего последние постепенно теряли функциональность и число их носителей сокращалось [1]. Классический пример — латынь, которая, будучи *lingua franca* Римской империи, позже перестала быть чьим-либо родным языком, трансформировавшись в группу романских языков. Тем не менее, латынь сохранилась благодаря огромному корпусу письменных памятников, научных, юридических и литературных текстов, что обеспечило её детальное изучение и преемственность.

Однако далеко не все языки обладают подобным преимуществом обширной письменной традиции. Языки малых народов, часто имеющие исключительно устную форму, исчезают, не оставляя после себя значительных документальных свидетельств. Ярким примером подобной уязвимости является ваханский язык. Он находится под серьёзной угрозой исчезновения, располагая минимальным количеством письменных материалов, что затрудняет его документацию и сохранение [2].

Ваханский язык — это язык региона Вахан, который расположен на стыке таких стран, как Афганистан, Китай, Пакистан, Таджикистан (см. рис. 1).

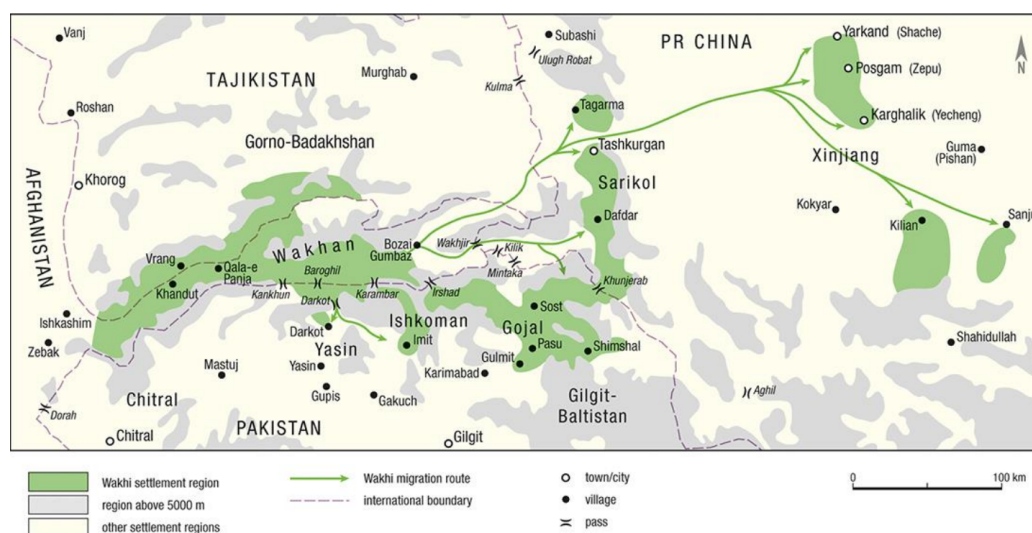


Рис. 1. Расселение носителей ваханского языка.

Расселение ваханского народа имеет географию стыка четырёх стран. У ваханцев нет своей единой письменности, поэтому в каждом государстве ваханцы адаптировались под письменность этого государства. Для исследования проблемы была выбрана Таджикская вариация ваханского алфавита.

Этот алфавит примечателен тем, что он на латинской основе с вкраплением кириллических символов, таким образом можно совместить распознавание ваханских символов с распознаванием кириллических символов.

Структура статьи организована следующим образом. В разделе 2 приведены особенности алфавита ваханской письменности и метод генерации обучающего датасета. В разделе 3 представлена архитектура нейронной сети и схема работы алгоритма. В разделе 4 представлена оценка качества и количественные признаки эксперимента. Раздел 5 завершает статью.

2. ОСОБЕННОСТИ ВАХАНСКОЙ ПИСЬМЕННОСТИ

Ваханский язык таджикской вариации обладает рядом особенностей. В нем достаточно много кириллических и латинских символов, но также есть символы, которых нет в наиболее полном стандарте кодировки символов — Unicode [3] (см. рис. 2).



Рис. 2. Пример символов, которых нет в стандарте кодировки Unicode.

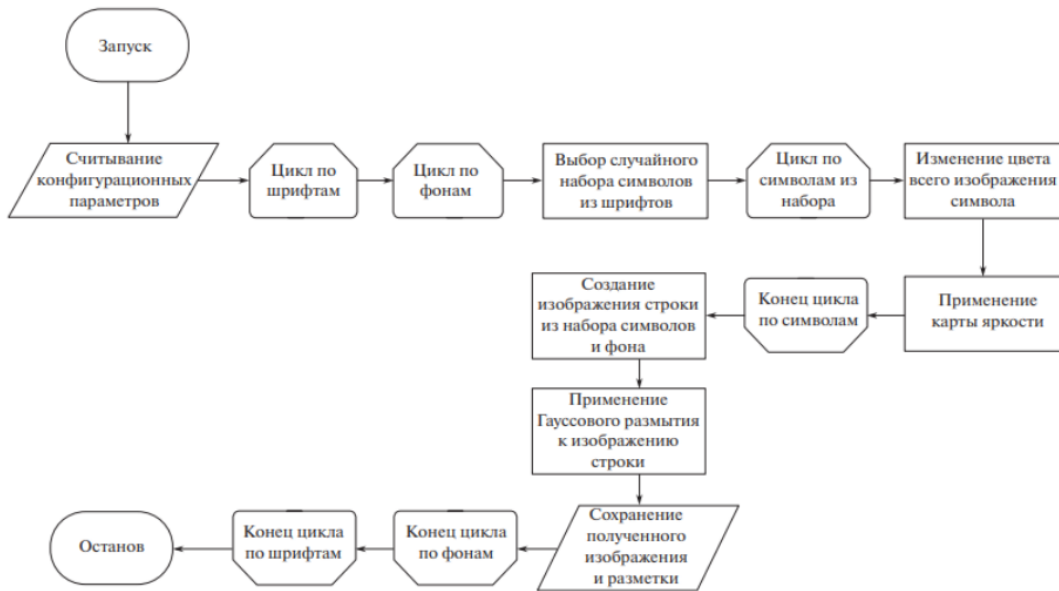


Рис. 5. Алгоритм работы фабрики.

После генерации получаем строки со всеми символами ваханского и русского алфавита, достаточных для обучения сети. Этот датасет учитывает и символы, которых нет в Unicode, они добавлены в строки (см. рис. 6). Символы берутся в случайном порядке, что позволяет сети не выучивать наиболее часто встречающиеся паттерны, а распознавать каждый символ по отдельности. Тем не менее, при генерации данных использовались мягкие ограничения: в большинстве данных слова удовлетворяют некоторой модели: слова разделяются либо пробелом, либо пунктуацией, слова из смеси букв и цифр недопустимы, а в каждом слове заглавная буква либо первая, либо все. Этим требованиям удовлетворяет 90% слов.

10357388961!yŽasō Ēvččc»Úkg:JTZŌQǫHIŠQK
 SĪŌĪ-Xduxv vvžšyrkfčopi?tysapzxsžbĭczž
 δōšžpz Θn»Nç:mδmcnwgzgi jgvnnue(ŠŠÚBCWH
 YČKF"tdékыгы'zpqlqžxδ;Kyvz.csžδ'fkšej

Рис. 6. Пример строки из обучающего датасета.

На рисунке 6 приведен пример из сгенерированных данных. На одной странице имитируется многострочный текст с варьируемыми межстрочными интервалами для большей натуральности данных.

3. АРХИТЕКТУРА НЕЙРОННОЙ СЕТИ

Архитектура сети для решения этой задачи была выбрана полносвёрточная FCN (Fully Convolutional Network) [8] [9]. Такая архитектура хороша тем, что система распознавания,

основанная на ней, совмещает в себе сразу сегментацию и классификацию символов. Так как в батчах лежат не символы по отдельности, а фрагмент строки, то и в разметке находится строка, по ширине совпадающая с изображением. В каждой позиции, которая соответствует центру какого-либо символа находится номер этого символа в алфавите, а во всех остальных позициях записано значение 1 которое означает, что гданиерт с этого места не распространяется. Важно заметить, что в архитектуре сети присутствуют уменьшения размерности и на лосс функцию подается карта признаков, которая имеет другую систему координат. Для того, чтоб исходная разметка соответствовала результирующей карте признаков, она также подвергается этому преобразованию. Архитектура сети в общем показана на рис. 7.

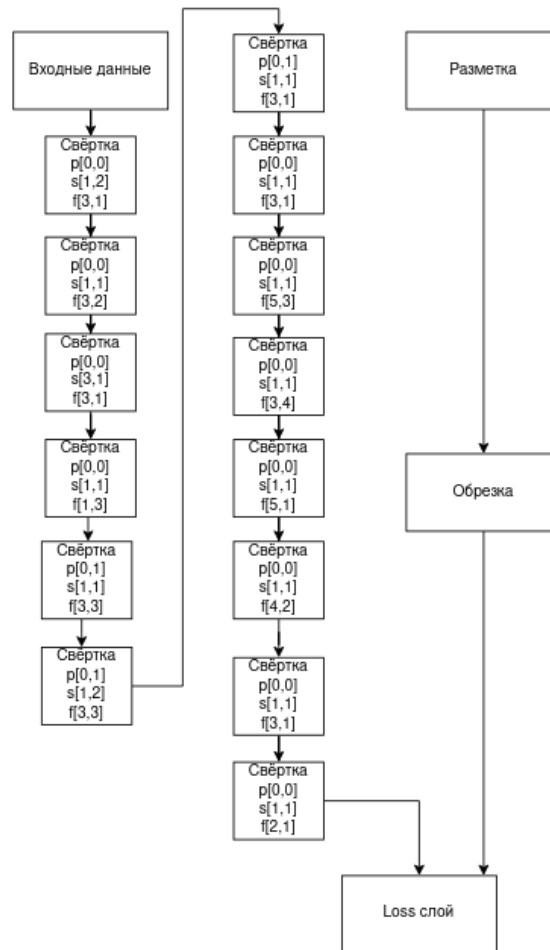


Рис. 7. Архитектура нейронной сети.

Loss функция использовалась SMCE(Softmax cross-entropy) [10]. Функцией активации выступила SymRELU. Для поиска базовых линий использовался метод, приведённый в статье [11]. А для сегментации символов в строке использовался алгоритм, описанный в статье [12] [13].

Рабочий цикл представленного алгоритма состоит из последовательно выполняемых этапов. На первом этапе осуществляется отбор и подготовка шрифтовых наборов, предназначенных для обучения модели. Далее производится синтез обучающей выборки — генерация исходных данных. Следующим шагом применяются методы аугментации данных для увеличения разно-

образия выборки и повышения устойчивости модели. Завершающей фазой процесса является непосредственное обучение нейронной сети на подготовленном таким образом массиве данных с использованием раздутий "на лету".

4. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

Для численной оценки качества представленного решения были сгенерированы данные с помощью описанной выше "фабрики". Генерация включала в себя бинарные текстовые строки в количестве 230400. Для упаковки в обучающие батчи с этих строк скользящим окном наре-зались патчи размера 405X33. Каждый символ алфавита в обучающей базе встречался от 5693 до 102404 раз. На этих данных нейронная сеть обучалась 20 эпох с использованием аугмен-таций на лету [14]: увеличение жирности текста, так как в книге текст жирный; добавление шума на страницу, так как это старый скан книги.

Для тестирования использовалась страница из книги про ваханский язык [15] (см. рис. 8). После обработки этой страницы нейронной сетью, получен следующий результат (см. рис. 9)

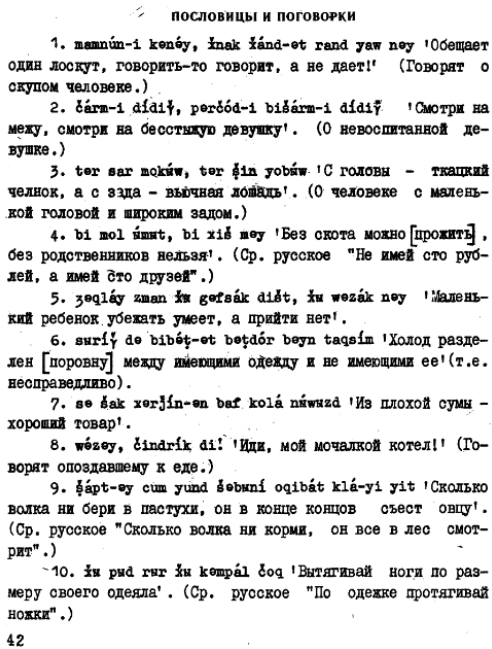


Рис. 8. Страница из книги по ваханскому языку для тестирования.

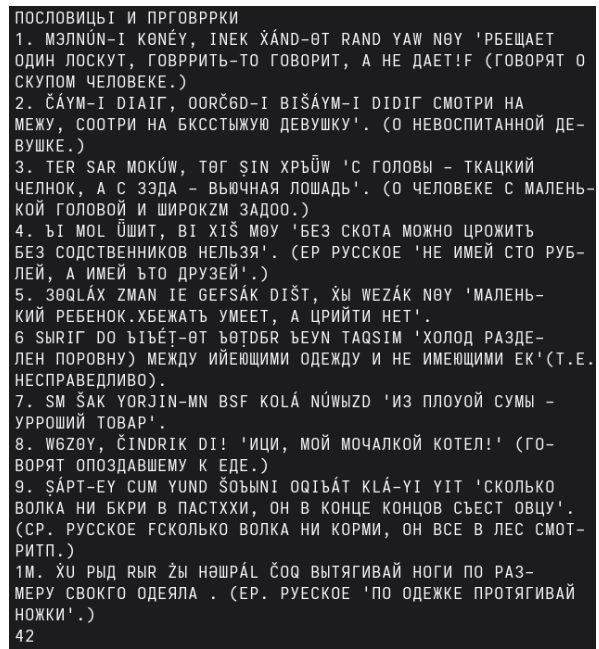


Рис. 9. Результат распознавания.

После обработки получилось посимвольное качество 0.88. Рассчитывалось по формуле

$$\frac{\text{Total} - \text{lev}}{\text{Total}} \tag{1}$$

где Total это количество всех символов на странице, lev — Расстояние Левенштейна.

5. ЗАКЛЮЧЕНИЕ

В настоящей работе была поставлена и решена задача автоматического распознавания текста на ваханском языке — исчезающем языке с ограниченной письменной традицией. Для

этого был разработан и реализован комплексный подход, основанный на применении полносвёрточной нейронной сети (FCN) и специализированного метода генерации синтетических данных.

Ключевые достижения работы можно резюмировать следующим образом:

- Адаптация под низкоресурсную среду: предложенный метод генерации обучающего датасета («фабрика данных») позволил преодолеть главное препятствие — отсутствие крупных размеченных корпусов. Путем синтеза изображений на основе доступных шрифтов, включая символы, отсутствующие в стандарте Unicode, был создан репрезентативный и разнообразный набор данных, достаточный для обучения глубокой модели.

- Валидный архитектурный выбор: Использование полносвёрточной нейронной сети доказало свою эффективность для данной задачи, так как позволило построить решение, способное к выделению инвариантных признаков символов, устойчивое к вариациям шрифтов и артефактам генерации.

- Практический результат: Обученная модель продемонстрировала качество посимвольного распознавания на уровне 0.88 на тестовом синтетическом наборе. И хотя этот результат не является абсолютным, он служит доказательством работоспособности всего пайплайна — от генерации данных до обучения и применения модели, задавая базовый уровень (baseline) для дальнейших исследований.

Перспективы дальнейших исследований лежат в нескольких направлениях:

- Улучшение модели: Исследование более сложных архитектур (например, комбинаций CNN с механизмами внимания или трансформерами), тонкая настройка гиперпараметров и применение аугментаций, имитирующих реальные условия сканирования (деформации, noise, неравномерное освещение).

- Создание большего количества шрифтов для дальнейшего расширения разнообразия данных в обучающей выборке и, тем самым, дальнейшего повышения качества работы системы.

Проведенное исследование подтверждает, что современные методы машинного обучения, в частности, полносвёрточные нейронные сети, в сочетании с универсальным подходом к синтезу данных, представляют собой действенный инструмент для документации и сохранения языкового наследия. Разработанный подход не только вносит конкретный вклад в сохранение ваханского языка, но и задаёт воспроизводимую методологию, применимую для аналогичных задач по сохранению других исчезающих языков с ограниченными письменными ресурсами.

СПИСОК ЛИТЕРАТУРЫ

1. D. Crystal, "Language Death" Cambridge University Press, (2000) - 198 p.
2. UNESCO, "Atlas of the World's Languages in Danger" Paris: UNESCO Publishing, (2010) - 230 p.
3. The Unicode Consortium, "The Unicode Standard, Version 13.0. 2020. Chapter 3" <https://www.unicode.org/versions/Unicode13.0.0/ch03.pdf>. Accessed: 2025-12-26.
4. V. Efimova, V. Shalamov, and A. Filchenkov, "Synthetic dataset generation for text recognition with generative adversarial networks," in Proc. SPIE 11433, Twelfth International Conference on Machine Vision (ICMV 2019) (2020). DOI: 10.1117/12.2558271.
5. T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in Proc. of the 21st International Conference on Pattern Recognition (ICPR) (2012) - pp. 3304–3308.
6. A. V. Gaer, Yu. S. Chernyshova, and A. V. Sheshkus, "Generation of artificial training sample for the task of recognition of symbols of the Russian passport fields," Computer Optics **43**(6), 1023–1031 (2019). DOI: 10.18287/2412-6179-2019-43-6-1023-1031.

7. P. K. Zlobin, Yu. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, "Neural network method for character sequence generation for text images training dataset synthesis," *Proceedings of the Institute for System Analysis of the Russian Academy of Sciences* **73**(2), 40–49 (2023).
8. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(4), 640–651 (2017).
9. J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) - pp. 3150–3158. DOI: 10.1109/CVPR.2016.343.
10. J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," *NATO ASI Series, Series F: Computer and Systems Sciences* **68**, 227–236 (1990).
11. Chernyshova, Yulia S. and Sheshkus, Alexander V. and Arlazarov, Vladimir V., "Two-step CNN framework for text line recognition in camera-captured images," *IEEE Access* **8**, 32587–32600 (2020). DOI: 10.1109/ACCESS.2020.2974051.
12. M. A. Povolotskiy, D. V. Tropin, T. S. Chernov, and B. I. Savelyev, "Dynamic programming approach to textual structured objects segmentation in images," *Information Technologies and Computing Systems* **3**, 66–78 (2019).
13. V. V. Arlazarov, E. I. Andreeva, K. B. Bulatov, D. P. Nikolaev, O. O. Petrova, B. I. Savelev, and O. A. Slavin, "Document image analysis and recognition: A survey," *Computer Optics* **46**(4), 567–589 (2022).
14. Gayer, Aleksandr Vyacheslavovich and Chernyshova, Yuliya Sergeevna and Sheshkus, Aleksandr Vladimirovich, "Effective real-time augmentation of training dataset for the neural networks learning," *ICMV 2018* **11041**, 1104111I–1104111I7 (2019). DOI: 10.1117/12.2522969.
15. A. L. Gryunberg, I. M. Steblin-Kamenskiy, "Yazyki vostochnogo Gindukusha. Vakhanskiy yazyk: teksty, slovar', grammaticheskiy ocherk" *Izdatel'stvo 'Nauka'*, (1976) - 670 p.

Fully Convolutional System for Wakhi Language Recognition

A. D. Korolkov, A. V. Sheshkus

The task of Wakhi language recognition has particular significance, as its solution will allow automating the process of digitizing existing data, which in turn will help linguists work on understanding the language and on preserving cultural heritage. In this article, we consider the specifics of the Wakhi language and a character recognition tool in the form of a neural network. The chosen approach is based on a fully convolutional neural network. An approach based on synthetic data generation was used for creation of the training data. Additionally, promising directions for using the developed approach in related fields are discussed.

KEYWORDS: Wakhi language, fully convolutional recognition, synthetic data generation, deep learning.