DOI: 10.15514/ISPRAS-2025-37(6)-42

CC BY

Clarifying Knowledge about Early Contacts of Native Speakers of the Proto-Finno-Volgaic Language Using Neural Networks

[1,2] *Ju.V. Normanskaja, ORCID: 0000-0002-2769-9187 <julianor@mail.ru>*
[1] *O.V.Goncharova, ORCID: 0000-0002-8665-6240 <oxanavgoncharova@gmail.com>*
[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*
[2] *Institute of Linguistic RAS, 1, B. Kislovskiy lane. Moscow 125009, Russia.*

Abstract. The article explores the potential of artificial intelligence for discovering new etymologies. It consists of two parts: the first describes the structure of the neural network, while the second provides examples of new types of etymologies, including Erzya additions to existing well-known etymologies, separate Finnic-Erzya parallels, and new hypotheses regarding borrowings from Baltic and Germanic languages. The purpose is to demonstrate the kinds of new etymologies that can be proposed within a relatively short time frame for languages with an established etymological tradition through the use of a neural network. The study utilizes a Finnish-Russian dictionary containing 17,212 lexemes and an Erzya-Russian dictionary comprising 8,512 lexemes, both hosted on the LingvoDoc platform. A neural network capable of proposing new etymologies for dictionaries on the lingvodoc.ispras.ru platform has been developed. Using this tool, Finnish and Erzya dictionaries were processed, resulting in the identification of over 100 new etymologies. Among these, 16 etymologies are discussed in the article, pertaining both to native Finno-Ugric vocabulary and borrowings.

Keywords: neural network; Finnish language; Erzya language.

For citation: Normanskaja Ju.V., Goncharova O.V. Clarifying knowledge about early contacts of native speakers of the Proto-Finno-Volgaic language using neural networks. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 6, part 3, 2025, pp. 149-162. DOI: 10.15514/ISPRAS-2025-37(6)-42.

Acknowledgements. This work was supported by a grant RSCF № 25-78-20002. The results were obtained using the services of the Ivannikov Institute for System Programming (ISP RAS) Data Center.

Уточнение информации о ранних контактах носителей прафинно-волжского языка с помощью нейросети

[1,2] *Ю.В. Норманская, ORCID: 0000-0002-2769-9187 <julianor@mail.ru>*
[1] *О.В. Гончарова, ORCID: 0000-0002-8665-6240 <oxanavgoncharova@gmail.com>*
[1] *Институт системного программирования им. В.П. Иванникова РАН,*
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.
[2] *Институт языкознания РАН,*
Россия, 125009, Москва, Б. Кисловский пер., д. 1.

Аннотация. Статья посвящена возможностям технологии искусственного интеллекта для поиска новых этимологий. Она состоит из двух частей, первая посвящена описанию того, как устроена нейросеть, во второй части приведены примеры типов новых этимологий: эрзянские дополнения для существующих общеизвестных этимологий, сепаратные финско-эрзянские параллели, новые гипотезы о заимствовании из балтийских и германских языков. Целью работы - показать, какие новые этимологии возможно предложить в достаточно сжатые сроки для языков с разработанной этимологической традицией благодаря использованию нейросети. Материалы исследования: финско-русский словарь, который содержит 17 212 лексем, и эрзянско-русский словарь, объемом 8 512 лексем, размещенные на платформе ЛингвоДок. В результате работы создана нейросеть, которая может предлагать новые этимологии для словарей, размещенных на платформе lingvodoc.ispras.ru, с ее помощью обработаны словари финского и эрзянского языков, выявлено более 100 новых этимологий, из них в статье разобрано 16 этимологий, которые относятся как к исконной финно-угорской лексике, так и к заимствованиям.

Ключевые слова: нейросеть; финский язык; эрзянский язык.

Для цитирования: Норманская Ю.В., Гончарова О.В. Уточнение информации о ранних контактах носителей прафинно-волжского языка с помощью нейросети. Труды ИСП РАН, том 37, вып. 6, часть 3, 2025 г., стр. 149–162 (на английском языке). DOI: 10.15514/ISPRAS–2025–37(6)–42.

Благодарности. Работа выполнена при поддержке Российского научного фонда, проект № 25-78-20002**.** Результаты получены с использованием услуг Центра коллективного пользования Института системного программирования им. В.П. Иванникова РАН – ЦКП ИСП РАН.

1. Introduction

In the present day, artificial intelligence (AI) technologies are actively addressing significant social challenges and have developed into a substantial industry. Linguistics has amassed a substantial amount of material that is suitable for the application of AI technologies. However, to the best of our knowledge, neural networks have not yet been utilized in comparative-historical linguistics in Russia.

Globally, the first studies on the automatic computational detection of cognates across languages appeared roughly twenty-five years ago. Initially, it was attempted to identify cognate words in the basic vocabulary of different languages using clustering methods (see [1]). These methods rely on comparing words with the same meaning from the basic vocabulary. Phonemes are grouped into clusters, which can vary greatly in size – for instance, clusters of vowels in medial position or consonants in final position may be quite large, whereas clusters of plosives or affricates may be relatively small.

In 2012, German researchers J.M. List and R. Forkel developed the well-known LexStat algorithm [2], which is available online as part of their collection of related tools, some of which were built upon this algorithm. This program, which combines a clustering algorithm with optimally calibrated weights, determines which words from two lists of basic vocabulary are cognates and which are not.

In a 2017 paper, G. Jäger developed a machine vector model and state-of-the-art (SOTA) phonetic alignment algorithms that make it possible to identify cognates in multilingual wordlists [3]. Based on basic vocabulary lists, it was shown that the algorithm performs rather poorly on small datasets, but its accuracy improves as the number of lexicons compared increases. Interestingly, the evaluation was conducted using data from three language families – Indo-European, Altaic, and Austronesian. The highest error rate was observed for the Indo-European languages, which have the longest tradition of study. This demonstrates that even clearly related words, for example in the Indic and Germanic branches, can differ phonetically to a considerable degree. The authors of this highly cited paper on clustering methods concluded that such methodologies cannot replace comparative-historical linguistic analysis.

As far as we know, one of the earliest attempts to use neural networks for cognate detection was undertaken in 2007 by Russian programmers working at the time in the United Kingdom and Bulgaria (see [4]). Their method was tested on English, French, German, and Spanish text corpora. The aim was to train a neural network to distinguish cognate words from those that merely exhibit "false similarity." The accuracy of cognate detection ranged from 81% for the German-English pair to 89.6% for the English-French pair. Overall, considering that the method was applied not only to basic vocabulary but also to broader text corpora in well-studied languages, this was a rather successful experiment. This study showcased the potential of neural networks in the field of comparative-historical linguistics. This paper has been extensively cited over the past fifteen years, particularly in studies focused on developing neural networks for historical-comparative linguistic research. (see [5-7]).

In [5], the authors applied neural network technology to five genetically related languages – Romanian, French, Spanish, Portuguese, and Italian – and to one unrelated language, Turkish. Their goal was to create, based on existing etymological research on Romanian, a neural network capable of distinguishing cognate and non-cognate words across the five languages. The accuracy of their method reached 87% when using a Support Vector Machine (SVM) with alignment-based features.

In [6], the author sought to move beyond language groups with well-established etymologies, although, in our opinion, the data used were not entirely reliable. The primary source was the Austronesian cognate database, which was available online in 2015 at http://language.psy.auckland.ac.nz/austronesian/ (now inactive). It is well known that no generally accepted system of regular phonetic correspondences exists for the Austronesian family. The second source was the Indo-European etymological database at http://ielex.mpi.nl/ (also currently inaccessible). The printed version of the corresponding monograph [8] shows that it includes only basic vocabulary etymologies, not full dictionaries. A third source was an etymological database of Mesoamerican Mayan languages described in [9]. However, as the authors themselves note, these etymologies were not the result of detailed linguistic work but were obtained automatically using the Levenshtein distance metric (which measures the difference between two sequences of characters). From studies of Indo-European, Uralic, and Altaic languages, we know that two words can have a minimal Levenshtein distance yet not be cognate according to regular phonetic correspondences. Thus, it seems that only the Indo-European etymologies could be considered somewhat reliable – and even they are currently unavailable for verification.

For training, the dataset included 167,676 word pairs (cognates and non-cognates) from Austronesian languages, 83,403 from Indo-European, and 63,028 from Mayan – meaning that only 26% of the material came from a language family with a well-established system of regular correspondences. In our view, such unreliable training data makes further use of this neural network impractical. The network itself is available online at GitHub – PhyloStar/SiameseConvNet: Performs cognate identification using Siamese Convolutional Networks, but the author has not updated it in the last eight years, and we have found no references to its use by other researchers.

In terms of dataset scale, a major breakthrough was achieved with CogNet, developed jointly by scholars from Australia, France, and Italy (see [10]). This resource combines material on eight

million cognates across 338 languages (see CogNet – UKC – Universal Knowledge Core for details [11]. According to its creators, the accuracy of automatic cognate detection reaches 94%. However, a closer examination reveals that for the languages of the Russian Federation, the database primarily relies on 100-word lists, many of which are extremely incomplete – for example, at the time of access, only 17 words were recorded for Even and 15 for Dolgan. In many cases, translations are erroneous, or only one of several synonyms is given, often an archaic or obsolete one – such as the Russian word lik instead of the common word litso ('face') (see Universal Knowledge Core | .:DataScientia: [11]. Even the examples displayed on the project's main page (GitHub – kbatsuren/CogNet: CogNet: [12] a large-scale, high-quality cognate database for 338 languages, 1.07M words, and 8.1 million cognates) contain errors in Altaic etymologies, while Uralic data is barely represented.

Nevertheless, the merits of this resource cannot be denied: it successfully integrates dozens of large online explanatory dictionaries for various world languages and, in many cases, constructs quite accurate etymologies while offering excellent visualizations of the results on a map.

In this concise overview, it is impractical to enumerate all neural network projects dedicated to cognate detection. However, it is noteworthy that, in our opinion, this research direction currently stands as one of the leading trends in comparative-historical linguistics globally. Developments in this field are not confined to Europe, America, and Australia. Notably, in 2017, a paper was published on the identification of etymologies in Arabic using neural networks (refer to [13]). Subsequently, in 2020, similar research emerged for the languages of India (refer to [14]). Furthermore, in 2024, research was conducted on Chinese (refer to [15]).

In contrast, the impact of this process on the languages of the peoples of the Russian Federation has been minimal to date. Indeed, CogNet query maps indicate that the territory of Russia remains largely devoid of linguistic data (see Universal Knowledge Core | .:DataScientia: [11]).

From this overview, it becomes clear that existing neural network models achieve very high accuracy in cognate detection (87–95%) when applied to languages included in their training data, provided that the quality of that data is high. We believe it is crucial that, for the languages of the Russian Federation, training data be based on reliable sources – compiled from the most authoritative existing etymological dictionaries and supplemented with modern research conducted by professional linguists – rather than on the kinds of crowdsourced or automatically generated materials used in many lesser-known neural networks. Furthermore, it is essential that neural network outputs undergo expert evaluation by specialists, as is standard practice in medicine, rather than being published in raw form as in CogNet, since such practices devalue genuine scientific achievements and often propagate inaccurate information.

Regrettably, the code for all known neural networks specifically designed for etymological research is inaccessible to users, and there are no publicly available datasets that facilitate the verification of their accuracy. Furthermore, it is noteworthy that foreign projects developing neural networks for etymology are predominantly led by programmers, and their primary objective is not the discovery of novel etymologies but rather the assessment of network performance on existing datasets. Given that these neural networks are not open-access, the research conducted abroad has yielded limited value from a comparative-historical linguistic perspective.

At present, the LingvoDoc platform [16] hosts dictionaries representing more than two thousand dialects of the languages of the Russian Federation. These dictionaries have been partially interconnected through etymological links. This was accomplished manually, utilizing LingvoDoc tools that rely on phonetic and semantic regular correspondences. Consequently, over 1.5 million words were connected through etymological relationships. This material formed the basis for training the neural network developed by the authors.

In the first part of the present article, we describe the technical features of this neural network; in the second, we present an overview of the types of etymological proposals generated by processing the Finnish and Erzya dictionaries using the network.

2. Neural network principles

In the present study, a two-stage approach is proposed for the task of automatic cognate identification in Uralic language corpora. At the first stage, a Siamese neural network was implemented. It takes into account graphical (orthographic) information from the input examples. Evaluation on the validation set demonstrated an average accuracy of approximately 78%.

At the second stage, the model architecture was expanded with an additional processing path for word translations and several heuristic techniques (a Boolean feature **exact_match** with a fixed correction factor, learnable weight coefficients α/β, and a threshold τ = 0.9), which allowed the classification accuracy to be increased to 92%.

The following sections describe the data used, the model architecture, and key training and tuning parameters that ensure network convergence and generalization capacity.

2.1 Dataset

The training data for the model were sourced from the LingvoDoc platform [16], which offers tools for the creation, annotation, and storage of electronic dictionaries, corpora, and concordances for diverse languages, encompassing phonetic and etymological analysis.

The initial corpus contained approximately 98,000 unique lexical entries from dictionaries of Uralic languages, each entry manually verified and supplied with a list of cognates. The lists were converted into all possible "cognate–cognate" combinations, after which duplicate pairs were removed; the final collection of positive examples contained about 1 million records.

A balanced negative class was generated using random negative sampling, excluding any pairs that overlapped with the positive examples.

For both stages, the training set has the general form

$$D = \{(x_i, y_i)\}_{i=1}^{N},\ y_i \in \{0,1\},$$

where the structure differs by stage:

- **Stage 1:** $x_i = \left(w_i^{(1)}, w_i^{(2)}\right)$
- **Stage 2:** $x_i = \left(w_i^{(1)}, t_i, w_i^{(2)}, t_i\right)$,

where $w_i^{(j)}$ is a word from language *j*, and t_i is its translation into Russian.

$$y_i \begin{cases} 1, \text{если } w_i^{(1)} \text{ и } w_i^{(2)} \text{когнаты}, \\ 0, \text{если они не когнаты } \wedge. \end{cases}$$

In the testing phase, the task for each pair is to predict the label y ^∈{0,1}.

Example of a training record:

Ала́ ала́-мъ 1 (for positive example)

kvɛ'ʃkəldʑəgu simy 0 (for negative example)

Initially, the model included 16-dimensional binary vectors of phonetic features [17]. However, subsequent analysis revealed that the lack of unified transcription standards in the source corpora introduced considerable noise. As a result, the phonetic vectors were excluded, and the focus shifted to orthographic features and the semantic representations of translations.

For the fine-tuning stage, an additional corpus of 700,000 translated pairs was prepared, symmetrically distributed between the two categories – cognates and non-cognates.

Example of a fine-tuning record:

ала́ город ала́ − мъ город 1 $(for\ positive\ example)$

$kvɛ'ʃkəldʑəgu$ отдохнуть $simy$ глаз 0 $(for\ negative\ example)$.

2.2. Model Architecture Description

At the first stage, a Siamese neural network was used, consisting of two identical branches, each processing one of the words in the analysed pair $w_i^{(1)}$ and $w_i^{(2)}$.

Each branch takes as input a sequence of characters, represented by two types of embeddings with dimensionality d = 128:

- **Character embedding** $E_{char}(X) \in R^{L\times D}$, which maps each character into a continuous vector space of dimension *D*;
- **Positional embedding** $E_{pos}(X) \in R^{L\times D}$, encoding the position of each character within the sequence (where *L* is the word length).

As a result, an input representation matrix is formed: $X^{(0)}=E_{char}(X) + E_{pos}(X)$.

To prevent overfitting, SpatialDropout1D (probability = 0.2) was applied, implemented as **Dropout2d** on a tensor of shape (*batch*, *d*, *L*), which zeroes out entire embedding channels [18].

Sequence processing is performed by a bidirectional LSTM (BiLSTM) with a hidden state size of 64, capturing contextual dependencies in both forward and backward directions [19]. The resulting features are passed through four transformer blocks, each implementing a multi-head attention mechanism that models complex nonlinear dependencies between characters:

$$Attention(Q,K,V) = softmax\left(\frac{QK^t}{\sqrt{d_K}}\right)V,$$

where *Q*, *K*, and *V* are projections of the input data, with the dimensionality of the keys being (*num_heads = 4*, key dimension per head = 128 / 4 = 32) [6].

After the attention layer, a position-independent feed-forward network (FFN) with 128 neurons and ReLU activation is applied, along with normalization and dropout layers for training stability.

The features from the transformer blocks are averaged across the temporal axis (*torch.mean*):

$$u = \frac{1}{L}\sum_{i=1}^{L} h_{j,}u \in R^{2h},$$

which aggregates the entire sequence into a single fixed-size vector. The similarity between words is then computed via cosine similarity between their hidden representations:

$$s = \frac{\langle u_1, u_2\rangle}{\|u_1\|\|u_2\|} \in [-1,1].$$

All resulting values are concatenated into a single vector:

$$z = [u_1; u_2; s] \in R^{4h+1},$$

which serves as the input to the final classifier. The classification module is a three-layer fully connected perceptron. The first layer applies *LayerNorm* and *ReLU* to an affine transformation of *z*; the second hidden layer is processed similarly, and a linear output followed by a sigmoid σ yields the predicted probability $\hat{y}$:

$$h_1 = ReLu\big(LayerNorm(W_1 z + b_1)\big),$$

$$h_2 = ReLu\big(LayerNorm(W_2 h_1 + b_2)\big),$$

$$\hat{y} = \sigma(W_3 h_2 + b_3).$$

At the second stage, the Siamese network was fine-tuned on a dataset that included word translations, and several heuristic strategies were introduced to further improve accuracy.

The base architecture was expanded to handle dual pathways:

- The word_encoder consists of a BiLSTM layer (*hidden_size = 64, bidirectional*) that extracts contextual features, followed by two transformer blocks with multi-head attention ($h = 4$) and an FFN with 128 neurons to model global word dependencies.
- The translation_encoder uses an analogous BiLSTM layer, followed by four transformer blocks to process translations.

Features from both pathways are combined with trainable coefficients:

$$p_i = \alpha t_i + \beta w_{i,} \alpha + \beta = 1,$$

where α and β are parameters for translations and words, initialized as (0.7, 0.3) and trained jointly with the other network weights.

Additionally, a Boolean feature **exact_match** was introduced, equal to 1 if the first four characters of the translations (considering padding) coincide, and 0 otherwise:

$$m = 1\left\{\tau_{1:4}^{(1)} = \tau_{1:4}^{(2)}\right\}.$$

At the classification stage, vectors and are p_1 and p_2 concatenated together with their element-wise difference and product:

$$v = [p_1; p_2; |p_1, p_2|; p_1 \times p_2]$$

The standard MLP then outputs a base logit l_{base}, which is adjusted by a heuristic contribution:

which is combined with the heuristic term

$$l = l_{base} + \gamma m,$$

where the coefficient $\gamma(match_{coef})$ is fixed and is not updated during fine-tuning. For example, if the first four characters of the translations are identical (e.g., *lesnoy* → *lesnye*), the model adds a fixed correction *match_coef = 0.8* to the classifier output.

To convert logits into binary labels, a threshold τ (*threshold*) [2] was applied. Initially, the validation threshold was determined dynamically: the ROC curve on a held-out set yielded τ values in the range [0.70, 0.78], representing an optimal trade-off between true positives (TP) and false positives (FP), accounting for penalties on false positives [1].

However, when working with real data, a large number of pairs were incorrectly identified as cognates. To reduce false positives, a fixed threshold $\tau = 0.90$ was set – meaning that the model must be highly "confident" (probability ≥ 90%) before classifying a pair as cognate. In practice, this reduced false positives by approximately 2–3 times while maintaining acceptable recall, as most true cognates received high scores $p \geq 0.90$.

Training examples for both stages were randomly divided into training (90%) and validation (10%) sets. Optimization used the AdamW algorithm with initial parameters learning rate = 10^{-4} (and weight decay = (10^{-4}.The binary cross-entropy loss (BCEWithLogits) function [20] was employed.

3. Types of New Etymologies Obtained using the Neural Network

In the present article, we provide examples of types of etymological proposals generated by the neural network, identified during the processing of two dictionaries:

1. the Finnish–Russian dictionary, which contains 17,212 lexemes, available online [21];
2. the Erzya–Russian dictionary, which contains 8,512 lexemes [22].

The Finnish words were taken into new etymologies were checked by [23], Erzya words were checked by [25]. As a result of the neural network's operation, 146,474,120 possible pairings were analyzed, and 16,055 etymological proposals were presented to the user; see Fig. 1 for the output format.

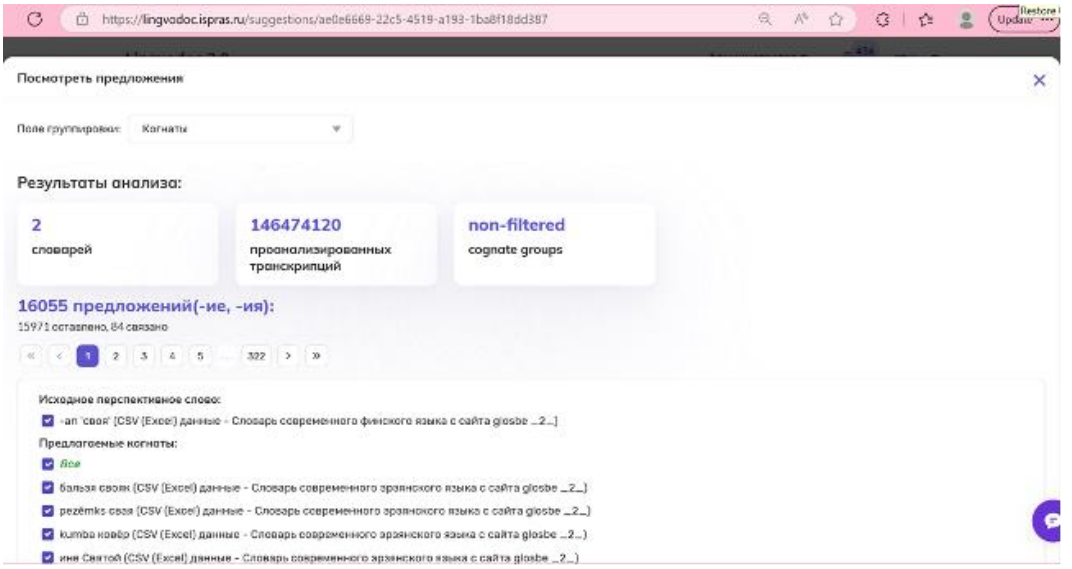

Fig. 1. Interface for displaying etymological proposals after dictionary processing by the neural network.

The user of the neural network must evaluate all of the network's proposals and highlight those deemed convincing, then each time click the "link" checkbox. As a result of our evaluation, 805 of the 16,055 etymological proposals were recognized as correct; these can be viewed in the dictionaries under the "cognates" column. Most of these proposals replicate existing etymologies reflected in the etymological dictionaries [23, 25]. However, more than 100 etymologies proposed by the neural network turned out to be new. They can be classified into three types:

1. Erzya additions to the existing "major" etymologies in [25], which have cognates in Finnish and several other Uralic languages but whose Erzya reflexes were previously unknown;
2. New distinct Finnish–Erzya etymological groups have been identified, comprising words that, according to [23], previously lacked cognates outside the Finnic (Baltic-Finnic) language family.
3. new Erzya parallels for Finnish words that, according to [23], were previously considered to be loanwords from Baltic or Germanic languages. In these cases, it is possible to demonstrate an earlier time of borrowing, already in Proto-Finno-Volgaic (with a split approximately 4,000 years ago), or to abandon the borrowing hypothesis and assume a native Uralic origin of the word.

It is evident that assessing the quality and novelty of the neural network's etymological proposals requires the user to undertake substantial research: verifying all hypotheses against existing etymological dictionaries and evaluating them in terms of the system of regular correspondences that underlies the etymologies accepted in [25; 23]. Below, as an illustration, we present 15 new etymological proposals across these groups.

3.1 Erzya Additions to Existing Etymologies

1. For the etymology FU **aja*- 'to hunt' > Fin. *aja*- 'to drive (game), to hunt, to lead', Est. *aja*- 'to move forward, to watch, to chase', Saami *vuoggje- -j*- 'to ride (a horse, reindeer), to lead' (N), *vuodjē*- 'to drive a car' (L), *vȳjje*- (T), *vujje*- (Kld), *vuajje*- (Not.) 'to drive a car, to direct', Udm. *uj*-, *uĺ*- (S), *ujî*- (G) 'to chase', Komi *voj*- (Skr) 'to be headstrong, to ignore the reins, to descend very fast from a mountain', *vojlį*- (Vm. I) 'to run, gallop, race', *vojed*- 'to run, to start running', *vojledlį*- 'to hunt', Mansi *wujt*- (K), *wojt*- (KM), *vujt*- (KO) 'to hunt' [14:4], we propose to add the comparison with Erz. *ajdäms* 'to drive, roll, rock',

where *-d-* is a suffix forming verbs with the meaning "single occurrence of an action", see [26].

2. For the etymology FU **tärmä* 'strength, strong' > Fin. *tarmo* 'energy, strength', Saami *dar'bmo -bm-* (N) 'energy, strength', *dar'bme -rbm-* (L), *tar'mē* 'strength, energy', Khanty *tärəm* (V), *tə̀rəm* (DN), *tarəm* (Kaz.) 'strong, energetic'?, Mansi *tēriŋ* (P), *tērəŋ* (K) 'agile, quick, wild', *tēr* (K) 'giftedness' ?, [25: 517], we propose to add the comparison with Erz. *tarmo* 'energy'. The Erzya forms provide evidence for the reconstruction of the back-vowel series. It may be necessary to separate the Ob-Ugric forms from this etymology, as the authors of [25: 517] also mark them with a question mark. Alternatively, it may be assumed that the same alternation of front and back vowels as in PU **kakta* ~ **käktä* 'two' applies to these forms.
3. For the etymology PF **kansa* 'people, folk, friends' > Fin. *kansa* 'people, nation, folk', Est. *kaasa* 'friend, companion, spouse', Saami *gaʒ'ʒe -ʒʒ-* (N) 'household, community', (T) *kaince*, (Kld) *kāinc*, 'companion', Udm. *kuz* (S) 'pair', (S G) *kuzo*, (K) *kuzo* 'pair', Komi *goz* (S P) 'pair', *guz* (PO) 'pair', *gozja* (S P) 'paired, married couple', [25: 645], we propose adding Erz. *kan* 'clan (line of generations), tribe'. In [25] there are no other etymologies with the cluster **-ns-* that have a reflex in Erzya, and therefore such development cannot be excluded.
4. For the etymology FU **kira* 'oath, curse, swearing' > Fin., Karel. *kiro* 'curse', Est. *kiruda* 'cursed', Veps *kirota* 'cursed', Votic *tširota*, Saami (N) *gârro* 'curse', Mansi (L) *kor-* 'cursed', Khanty (I) *korəm*, (E) *χurəm* 'to get angry', we propose adding Erz. *krojams* 'to swear', *krojsema* 'obscenity, foul language', where *-j(a)-* is a frequent verbal suffix, see [26: 92]. The loss of **i* in the Mordvinic languages is typical when the stress in Proto-Finno-Volgaic fell on the ending; see [16: 33-45], cf. FV **šišna* 'belt' > Mord. *šna* (E, M), *kšna* (E) [25: 786, 28: 909].
5. In FU **aŋa* 'to open, to release' > Fin. *avaa-* 'to open, to widen', *avanto* 'ice hole', Est. *ava-* 'to open', Mord. *ankšima, avśima* (E), *ańćema* (M) 'ice hole', Khanty *aŋə-* (V) 'to untie (a knot), to open', *eŋ-* (O) 'to take off clothes, shoes', Mansi *ę̄ŋk-* (TJ, LU), *āŋoχo-* (So.) 'to take off a dress', *ōŋkws-* (P) 'to skin a bear', Hung. *old-* (Old Hung. *ód-*) 'to release; to untie (knots)' [25: 11], it seems to us that Fin. *avanto* 'ice hole' and Mord. *ankšima, avśima* (E), *ańćema* (M) 'ice hole' do not fully match the semantics of the other forms. This is also noted by the authors of [23: I 92], who mark the connection between Fin. *avanto* 'ice hole' and *avaa-* 'to open, to widen' with a question mark. We propose comparing Fin. *avaa-* 'to open, to widen' with Erz. *avtems* 'to open up, to open', where *-t-* is a verbal suffix; for its meaning and frequency see [26: 169]. It should be noted that Erz. *-v-* is a frequent reflex of FU *-ŋ-* in the position after back vowels, cf. FU *čaŋV-* 'to beat' > Erz. *Čavo-* [25: 53].
6. In FU **ćärke-* 'to break, to hurt' > Fin. *särke-* 'to break, to hurt', Saami *čērgiidi-* (N) 'to go numb (of limbs)', *tjär'ka* (U) 'strong tingling in limbs fallen asleep after sleep', *íšĕär̄Ḡa-* (Ko. P) 'to hurt (of a wound)', Mari *šäryе-* (W) 'to open, to destroy', Khanty *t̓erəɣ-* (Trj.), *śarĭj-* (Ni.), *śarĭ-* (Kaz.) 'to hurt', Mansi *ćärk-* (TJ), *śə̄rɣ-* (KU), *śarr-* (P), *śārɣ-* (So.) 'bedauern', *śāriɣ-* (So.), *śārɣ-* (N) 'weh tun', Hung. *sér-* 'Schmerzen haben, weh tun'.

3.2 New Finnish–Erzya Comparisons and Their Additional Cognates According to LingvoDoc Dictionaries

7. To the etymological group Fin. *nikka* 'hiccup', Est. *nigu* 'hiccup', Karel. *nikko* 'hiccup', Veps *niki-* 'hiccup', Votic *nikottaa* 'hiccup', Saami (N) *njâkkâstit* 'hiccup' [23: II 220], we propose adding Erz. *niktnetems* 'to hiccup, to sob, to belch'. As noted in [26: 50, 26: 169], *-tńe-* is an unusual but known verbal suffix, whereas *-t-* is a typical verbal suffix. In [29],

Komi *ńiktiniі* 'to choke, to sputter (from coughing, tears, etc.)', *ńiktini* (Der., P.) [31] are also compared with the Finnish and Saami forms, this comparison also supports the Proto-Finno-Permic origin of these forms and allows to reconstruct the PFPerm **niktV-* 'hiccup'.

8. To the etymological group Fin. *sirkka* 'grasshopper, cricket', Karel. *sirk* 'grasshopper', Karel. *tširkka* 'grasshopper, cricket', Veps *tširk* 'grasshopper, cricket', Votic *tširk, širk* 'grasshopper, cricket' [23: III 186], we propose adding Erz. *ciŕkun* 'grasshopper, cricket', where *-un* is an unusual nominal suffix; see [30: 210] for details. On the LingvoDoc platform, other cognates of this proto-form can also be found: Moksha (Koldais) *ćiŕkun*, (Mordovskie Yurtkuli) *ćerku·n*, (Uryum) *ćiŕkun* 'grasshopper'. In [29], Mari *č8rkiem* 'to chirp, to twitter' and Komi *ćirk* 'grasshopper' are also compared to this etymological group, allowing the reconstruction FPerm *ćirkV 'grasshopper, cricket'*.
9. To the etymological group Fin. *kierittä* 'to roll', *kiero* 'bent, twisted', Est. *keer* 'bent, twisted', Karel. *kiero* 'bent, twisted', Veps *keŕ* 'roll', Votic *tšērtā* 'wheel', Saami (N) *gierre* 'thread, cord' [23: I 354], we propose adding Erz. *keverdems* 'to roll' and Hung. *kever* 'to spin, to stir', which according to [32: 746] also lacks an etymology. In Finnish, the change **w > 0* is frequent in intervocalic position, cf. *lewe* > *lyö-* 'to strike' [25: 247], *luwe* > *luu* 'bone' [25: 254], *puwe* > *puu* 'tree' [25: 410]. Thus, comparison of Finnic, Erzya, and Hungarian forms allows the reconstruction FU **kiwer- 'to roll, to turn'*.
10. The forms Fin., Karel. *vain* 'only', Est. *vaid*, Veps *vaiše*, Votic *vaitas* 'only' [23: III 392] can be compared with Erz. *vańks* 'only, merely, pure', where *-ks* is an Erzya nominal suffix; for its meaning, see [33]. One can reconstruct FV **wajn* '*only'*. In [25] there are no words with intervocalic **-jn-* that have Erzya reflexes, but the cluster **-jm-* > Erz. *-m-*, cf. FU **śajma* 'wooden vessel, boat' > Erz. *śuma, śima* 'log' [25: 456], while the cluster *-jn-* > Komi, Udm. *ń*. Therefore, a similar change in Erzya seems possible.

3.3 Erzya Parallels for Loanwords in (Baltic-)Finnic (and Saami) Languages

3.3.1 Examples of Germanic Loanwords

11. Fin. *tuoni* 'death', *tuonela* 'hell', Karel. *tuoni* 'death', Est. dial. *toonekurg, toonkurg*, Saami (N) *duodnâ* 'poor thing, death, hell'. To this etymological group we propose adding Erz. *tonači* 'hell', where *-či* is a typical affix for forming abstract nouns, cf. *мазы-чи* 'beauty', *сюпав-чи* 'prosperity, wealth'; see [34] for details. In [23: III 330] it is considered that the Finnic words are borrowed from PGmc. **dawīni > Old Norse dán* 'death', Norw. *dån* 'weariness' [23: III 330]. If this hypothesis is accepted, then a Proto-Germanic borrowing into Proto-Finno-Volgaic should be posited.
12. Fin. *kampa* 'comb' < Old Norse *kambr*, Norw. *kam* [23: I 294]. To this etymological group we propose adding Erz. *kaba* 'comb'. The Erzya reflex *-b-* < **mp* is not typical, but there are examples of **mp* > *-p-* in native vocabulary, cf. *kumpa* 'wave' > Erz. *kopildi-* 'to move in a wave-like motion'. It may be assumed that in borrowing, voicing could have occurred, which sporadically appears in Volgaic and Permic languages.
13. Fin. *avioliittovälittäjä* 'matchmaker', *avioliitto* 'marriage', *avio* 'spouse, marital partner, marriage', Est. *abielu* 'marriage' [23: I 92], we propose to compare with Erz. *avakuda* 'matchmaker', *ava* 'woman', Moksha *ava* 'woman', Mari *ava* 'mother'. In [23: I 92] a possible loan of the Finnic forms from Proto-Germanic **aiwō, *aiwa* 'law, norm' is noted as questionable; the closest semantic reflex to the Finnish is Old English *ǣw* 'law, marriage, spouse'. Note that the Mordvinic and Mari words may also be reflexes of PU **apV 'older female relative, aunt, elder sister'* [25: 15], although these parallels are absent in [25: 15], possibly due to semantic differences, since the standard phonetic reflex of **-p-* is Mord., Mari -v-. Phonetically similar words for 'woman' are also found in Turkic languages, cf.

PTurk. **apa* ‘mother, elder sister, aunt’, which in Tuvan has the reflex *ava*, see [35: 158-159]. Thus, in this case the hypothesis of Finnic forms borrowed from Germanic seems unlikely.

14. Fin. *viltti* ‘blanket’, Karel. *vilti* we propose to compare with Erz. *veĺtävks* ‘blanket’, *veĺtäms* ‘to cover with a blanket’, Moksha (Koldais) *véĺʹtə̂rda* ‘luxuriously embroidered blanket’. Comparing these forms allows for the reconstruction of PFV **wiltV ‘blanket’*. In [23: III 450] the Finnic forms are considered to be borrowed from Germanic, cf. Norw. *filt* ‘blanket’. This borrowing hypothesis appears quite plausible and, if accepted, would imply Germanic borrowings into Proto-Finno-Volgaic.

3.3.2 Examples of Baltic Loanwords

15. Fin. *vuode* ‘mattress’, Est. *voodi* ‘mattress’, which according to [23: III 472] is derived from *vuota* ‘skin flayed from a large animal’, we propose to compare with the Baltic-Finnic forms: Erz. *vatola* ‘mattress’, where *-la* is a suffix forming nouns, cf. *чова-ля* ‘bead’. The correspondence Erz. *a* – Fin. *uo* is also typical, cf. Erz. *kaŕ* – Fin. *kuori* < FU **kore* (**kōre)* ‘bark’ [25: 184]. According to [23: III 476], the Finnic words are considered borrowings from Baltic, cf. Lith. *óda* ‘skin’, Latv. *āda* ‘skin’. Without pausing here to evaluate the likelihood of borrowing these words from Baltic, we note that if the comparison with Erzya is accepted, then borrowing into Proto-Finno-Volgaic must be assumed.

16. Fin., Karel. *härkä* ‘ox’, Est., Veps *härg* ‘ox’, Votic *ärtšä* ‘ox’, we propose to compare with Erz. *šeŕge* ‘ox’, which go back to FV **čärkV*. In [23: I 210] the Finnic words are considered Baltic borrowings, cf. Lith. *žìrgas*, Latv. *zirgs* ‘horse’. Here, as in the previous etymology, the phonetics and semantics do not precisely correspond to the Finno-Volgaic, but if the borrowing hypothesis is accepted, they again indicate the greater intensity of early contacts between speakers of Baltic and Finno-Volgaic languages.

4. Conclusion

It may be concluded that the neural network we developed is sufficiently effective for identifying new etymologies even in languages as well studied as Finnish and Erzya. The analysis of sizable dictionaries (several thousand lexemes) using neural networks makes it possible to supplement existing etymologies, propose new ones, and refine the chronology of borrowings. From a linguistic perspective, the most intriguing result is that some words previously regarded as separate Baltic or Germanic borrowings into the Finnic (Baltic-Finnic) and Saami languages have parallels in the Mordvinic languages, whose phonetic correspondences in several cases (for example, Fin. *härkä* ‘ox’ – Erz. *šeŕge* ‘ox’) indicate a common and ancient source of borrowing. In such cases, we must assume either that these are not borrowings at all but inherited words in the Finno-Volgaic–and possibly Finno-Ugric–languages, and that the similarity to Germanic or Baltic is coincidental or points to a shared Nostratic origin; or else we must posit direct contacts between speakers of Proto-Finno-Volgaic and Baltic and Germanic groups.

While the hypothesis of contacts with the Balts and/or Balto-Slavs was proposed already in [36: 191] – even though that study identified no more than ten lexemes that could be regarded as borrowings into Proto-Finno-Volgaic–and is widely accepted and further substantiated in [37], to our knowledge there has previously been no evidence adduced for contacts between Germanic speakers and speakers of Proto-Finno-Volgaic. It has generally been assumed that borrowings from Germanic could have entered the Volgaic languages only via Russian; see [38]. However, among the four Germanic borrowings into Finno-Volgaic considered here, there is only one case in which a similar word appears in Tambov Russian dialects–*ava* ‘woman’ [39: I 196] – and that item is regarded as a borrowing from the Mordvinic languages, given that it is attested solely within the contact area.

List of abbreviatons

E. – Erzya
Est. – Estonian
PF – Finno-Permic
Fin. – Finnish
FU. – Proto-Finno-Ugric
PFV. – Proto-Finno-Volgaic
G. – Glazov Udmurt dialect
Hung. – Hungarian
I. – Irtysh Khanty dialect / Izhma Komi dialect (context-dependent)
K. – Kondinsk Mansi dialect
Kaz. – Kazym Khanty dialect
Karel. – Karelian
Khanty – Khanty
Kld. – Kildin Saami
Ko. – Skolt Saami
Komi – Komi
KM. – Middle Kondinsk Mansi dialect
KU. – Lower Kondinsk Mansi dialect
Latv. – Latvian
Lith. – Lithuanian
L. – Lozva Mansi dialect / Lule Saami (context-dependent)
LU. – Lower Lozva Mansi dialect
M. – Moksha
Mansi – Mansi
Mari – Mari
Moksha – Moksha
Mord. – Mordvinic
Ni. – Nizyam Khanty dialect
Norw. – Norwegian
Not. – Notozero Saami
O. – Obdorsk Khanty dialect
Old Hung. – Old Hungarian
Old Norw. – Old Norse
P. – Pechora Komi dialect / Pelym Mansi dialect / Pite Saami (context-dependent)
PFV – Proto-Finno-Volgaic
PF – Proto-Finno-Permic
PFPerm – Proto-Finno-Permic (if retained)
PGmc. – Proto-Germanic
PO. – East Permyak dialect
PU. – Proto-Uralic
S. – Middle Sysola Komi dialect / Sarapul Udmurt dialect (context-dependent)
Saami – Saami
Skr. – Sysola (Syktyvkar) Komi dialect
So. – Sosva Mansi dialect
TJ. – Tavda Mansi dialect
T. – Ter Saami
Trj. – Tremjugan Khanty dialect
U. – Ume Saami
Udm. – Udmurt
V. – Vakh Khanty dialect

Veps. – Veps
Vm. – Vym Komi dialect
Vot. – Votic
W. – Western (Forest) Mari dialect

References

[1]. Bergsma Sh., Kondrak G. Alignment-based discriminative string similarity. In Proc. ACL. 2007
[2]. Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists // Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 117–125, Avignon, France. Association for Computational Linguistics.
[3]. Jäger G., List J.-M., Sofroniev P. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists // Conference: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, 2017, Long Papers.
[4]. Mitkov R., Pekar V., Blagoev, D. et al. Methods for extracting and classifying pairs of cognates and false friends. Machine Translation 21, 29–53 (2007).
[5]. Dinu L.P., Ciobanu A.M. Building a Dataset of Multilingual Cognates for the Romanian Lexicon // Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC 2014, p. 3313-3318.
[6]. Rama T. Siamese Convolutional Networks for Cognate Identification // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, p. 1018–1027.
[7]. Fourrier C., Sagot B. Probing Multilingual Cognate Prediction Models // Findings of the Association for Computational Linguistics: ACL 2022. p. 3786-3801.
[8]. Dyen I., Kruskal J. B., Black P. An Indo-European classification: A lexicostatistical experiment. Transactions of the American Philosophical Society 1992, 82(5), p. 1–132.
[9]. Wichmann S., Holman E.W. Languages with longer words have more lexical change // Approaches to Measuring Linguistic Differences, 2013, p. 249–281.
[10]. Batsuren Kh., Bella G., Giunchiglia F. A large and evolving cognate database // Language Resources and Evalution, vol. 56, 2022, p. 1-25.
[11]. https://ukc.datascientia.eu/concept, дата обращения 26.11.2025.
[12]. https://datascientiafoundation.github.io/LiveLanguage/datasets/cognet/, дата обращения 26.11.2025.
[13]. Alreshidi H., Aldhlan K. Auto-Extracting Method of Cognates Words in Arabic and English Languages // International journal of advanced studies in Computer Science and Engineering (IJASCSE), vol. 6, issue 01, 2017.
[14]. Kanojia D., Bhattacharyya P., Kulkarni M., Haffari G. Challenge Dataset of Cognates and False Friend Pairs from Indian Languages // LREC 2020, p. 1-12.
[15]. Pulini M., List J.-M. Finding language-internal cognates in Old Chinese // Bulletin of Chinese Linguistics 2024, 17(1), p. 53–72.
[16]. https://lingvodoc.ispras.ru/, дата обращения 26.11.2025.
[17]. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.
[18]. Tompson, J., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient Object Localization Using Convolutional Networks. Proceedings of CVPR. https://arxiv.org/pdf/1411.4280**Ошибка! Недопустимый объект гиперссылки.**
[19]. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681.
[20]. Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. ICLR 2019.
[21]. https://lingvodoc.ispras.ru/dictionary/11457/163277/perspective/11457/163278/view, дата обращения 26.11.2025.
[22]. https://lingvodoc.ispras.ru/dictionary/11459/62466/perspective/11459/62467/view, дата обращения 26.11.2025.
[23]. Suomen sanojen alkuperä, ed. by Forsberg U.-M., Itkonen E. Helsinki, 1992-2000.
[24]. Erzya-Russian dictionary: 27000 lexems, ed. By Serebrennikova B. A. , Buzakovoj R. N., Mosina M. V. M., 1993.
[25]. Uralisches etymologisches Wörterbuch, ed. by K.Rédei. Budapest, 1986 – 1988.
[26]. Mesarosh E. Verb-forming suffixes in the Erzya language. Studia Uralo-Altaica 42. Seged, 1999.

[27]. Normanskaya Ju.V. Reconstruction of the Proto-Uralic paradigmatic stress and its influence on the development of the vocalism system. M., 2018.
[28]. Paasonen H. Mordwinisches Wörterbuch. II Band. Helsinki, 1992.
[29]. Lytkin V.I., Gulyaev V.G. A short etymological dictionary of the Komi language. Moscow, 1970.
[30]. Ariskina T.P. Suffixed nouns in the Erzya language: semantics and functioning // Vestnik ugrovedeniya № 2(8), 2018.
[31]. Comparative dictionary of Komi-Zyr'an dialects, complied by Zhilina T. I., Saxarova M. A., Sorvacheva V. A. Syktyvkar, 1961.
[32]. Etymologisches Wörterbuch des Ungarischen, ed. by Loránd Benkő. Budapest, 1993.
[33]. Ryabov I.N. Word-formation relations between parts of speech in the Erzya language. Saransk, 2000.
[34]. Vodyasova L.P. Ways of expressing grammatical meanings in the morphology of the Erzya language // Lingvistika №30, 2016 (https://sci-article.ru/stat.php?i=1455731730).
[35]. Etymological dictionary of Turkic language, ed. by Sevortyan E.V., t. 1, M., 1974.
[36]. Kalima J. Itämerensuomalaisten kielten balttilaiset lainasanat. Helsinki.
[37]. Napol'skix V.V. The Balto-Slavic language component in the Lower Kama region in the middle of the 1st millennium AD // Slavyanovedenie, 2006, 2, 3-19.
[38]. Butylov N.V. Foreign language vocabulary in Mordovian languages (Indo-European borrowings). Saransk, 2006.
[39]. Dictionary of Russian dialects, ed. by F.P.Filin, vol. I. Leningrad, 1965.

Информация об авторах / Information about the author

Юлия Викторовна НОРМАНСКАЯ – доктор филологических наук, главный научный сотрудник, заведующая лабораторией «Лингвистические платформы» Института системного программирования им. В.П. Иванникова РАН, ведущий научный сотрудник отдела Урало-алтайских языков Института языкознания РАН.

Yulia Viktorovna NORMANSKAYA – Dr. Sci. (Philology), Chief Researcher, Head of the Laboratory "Linguistic Platforms" at Ivannikov Institute for System Programming of the Russian Academy of Sciences; Leading Researcher, of the Department of the Ural-Altaic Languages at the Institute of Linguistics of the Russian Academy of Sciences.

Оксана Владимировна ГОНЧАРОВА – кандидат филологических наук, старший научный сотрудник лаборатории «Лингвистические платформы» Института системного программирования им. В.П. Иванникова РАН.

Oxana Vladimirovna GONCHAROVA – Cand. Sci. (Philology), Senior Researcher of the Laboratory "Linguistic Platforms" at Ivannikov Institute for System Programming of the Russian Academy of Sciences.