DOI: 10.15514/ISPRAS-2025-37(6)-28

Глубокое обучение и лингвистический анализ в задачах идентификации когнатов: обзор современных подходов

О.В. Гончарова, ORCID: 0000-0003-1044-6244 <goncharova_oxvl@pfur.ru>

Институт системного программирования РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.
Российский университет дружбы народов им. П. Лумумбы,
Россия, 117198, г. Москва, ул. Миклухо-Маклая, 6.
Пятигорский государственный университет,
Россия, 357532, г. Пятигорск, Ставропольский край, пр. Калинина, 9.

Аннотация. В статье представлен обзор современных подходов к автоматическому обнаружению когнатов, сочетающий методы глубокого обучения и классические лингвистические техники. Основная цель исследования - систематизировать существующие архитектуры, выявить их сильные и слабые стороны и предложить интегративную модель, объединяющую фонетические, морфологические и семантические представления лексических данных. Для достижения этой цели проведён критический анализ работ, опубликованных в период 2015–2025 гг. и отобранных с помощью специализированного парсера научного репозитория arXiv.org. В рамках анализа рассмотрены следующие задачи: (1) оценка точности и устойчивости сиамских сверточных нейронных сетей (CNN) и трансформеров при переносе фонетических паттернов между разнородными языковыми семьями; (2) сопоставление эффективности орфографических метрик (LCSR, нормализованное расстояние Левенштейна, индексы Джарро-Винклера и др.) и семантических эмбеддингов (fastText, MUSE, VecMap, XLM-R); (3) исследование гибридных архитектур, включающих морфологические слои и механизмы транзитивности для выявления частичных когнатов. В результате выявлено, что комбинирование фонетических модулей (сиамские CNN + трансформеры), морфологической обработки (BiLSTM на основе данных UniMorph) и обучаемых семантических векторов обеспечивает наилучшие показатели точности и устойчивости для различных языковых пар, включая малоресурсные. Предложена интегративная архитектура, способная адаптироваться к разнообразию языковых групп и эффективно оценивать степень родства слов. Итогом работы стал не только аналитический отчёт о передовых методах, но и разработка рекомендаций для дальнейшего развития автоматизированного выявления когнатов.

Ключевые слова: глубокое обучение; лингвистический анализ; идентификация когнатов; сиамские нейронные сети; трансформеры; орфографические метрики; семантические эмбеддинги.

Для цитирования: Гончарова О.В. Глубокое обучение и лингвистический анализ в задачах идентификации когнатов: обзор современных подходов. Труды ИСП РАН, том 37, вып. 6, часть 2, 2025 г., стр. 177–190. DOI: 10.15514/ISPRAS–2025–37(6)–28.

Благодарности: Исследование выполнено при поддержке гранта РНФ № 25-78-20002 «Возможности искусственного интеллекта для сравнительно-исторического изучения малоресурсных языков народов РФ» (рук. Ю. В. Норманская).

Deep Learning and Linguistic Analysis for Cognate Identification Tasks: A Survey of Contemporary Approaches

O.V. Goncharova, ORCID: 0000-0003-1044-6244 <goncharova_oxvl@pfur.ru>

Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.
Peoples' Friendship University of Russia named after Patrice Lumumba,
6, Miklukho-Maklaya str. Moscow, 117198, Russia.
Pyatigorsk State University,
9, Kalinin Avenue, Pyatigorsk, Stavropol Krai, 357532, Russia.

Abstract. The paper provides a comprehensive review of contemporary methods for automatic cognate detection, integrating deep learning techniques with traditional linguistic analyses. The primary objective is to systematize existing architectures, assess their strengths and limitations, and propose an integrative model combining phonetic, morphological, and semantic representations of lexical data. To this end, we critically analyze studies published between 2015 and 2025, selected via a specialized parser from the arXiv repository. The review addresses three core tasks: (1) evaluating the accuracy and robustness of Siamese convolutional neural networks (CNNs) and transformer-based models in transferring phonetic patterns across diverse language families; (2) comparing the effectiveness of orthographic metrics (e.g., LCSR, normalized Levenshtein distance, Jaro–Winkler index) with semantic embeddings (fastText, MUSE, VecMap, XLM-R); and (3) examining hybrid architectures that incorporate morphological layers and transitive modules for identifying partial cognates. Our findings indicate that a combination of phonetic modules (Siamese CNNs + transformers), morphological processing (BiLSTM leveraging UniMorph data), and learnable semantic vectors yields the best accuracy and stability across various language pairs, including low-resource scenarios. We propose an integrative architecture capable of adapting to linguistic diversity and effectively measuring word relatedness. The outcome of this research includes both an analytical report on state-of-the-art methods and a set of recommendations for advancing automated cognate detection in large-scale linguistic applications.

Keywords: deep learning; linguistic analysis; cognate identification; Siamese neural networks; transformers; orthographic metrics; semantic embeddings.

For citation: Goncharova O.V. Deep Learning and Linguistic Analysis for Cognate Identification Tasks: A Survey of Contemporary Approaches. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 6, part 2, 2025, pp. 177-190 (in Russian). DOI: 10.15514/ISPRAS-2025-37(6)-28.

Acknowledgements. The work was supported by Russian Science Foundation Grant No. 25-78-220002 “Capabilities of Artificial Intelligence for Comparative-Historical Study of Low-Resource Languages of the Peoples of the Russian Federation” (supervisor: Yu. V. Normanskaya).

1. Введение

Автоматическое обнаружение когнатов становится всё более востребованным направлением исследований, поскольку позволяет ускорить и стандартизировать процессы выявления родственных слов в разных языках. Целью настоящего исследования является систематический анализ существующих подходов к построению моделей для идентификации когнатов, формирование набора обоснованных выводов и выделение наиболее эффективных практик. На основе полученных результатов планируется разработать собственную интегративную архитектуру, объединяющую лучшие достижения в области фонетического, орфографического и семантического представления лексических данных.

В рамках поставленной цели планируется решить следующие задачи: (1) необходимо критически переосмыслить существующие методологические решения, выявить их сильные и слабые стороны с точки зрения точности, устойчивости к шумам и обобщающей способности; (2) сопоставить используемые принципы выявления признаков и алгоритмические подходы, чтобы на их основе сформировать оптимальный набор признаков

и механизмов обучения; (3) на базе обобщённого опыта сконструировать модель, способную гибко адаптироваться к разным языковым группам и эффективно оценивать степень родственности слов.

Таким образом, настоящее исследование призвано не просто описать имеющиеся подходы, но и на их основе синтезировать новый метод, способный объединить достоинства разнообразных архитектур и признаковых представлений. Итогом станет модель, оптимизированная по точности и надёжности, и набор рекомендаций для дальнейшего развития автоматизированного выявления когнатов в масштабных лингвистических задачах.

2. Методология сбора и анализа литературы

Для реализации задачи систематического анализа современных подходов к автоматическому обнаружению когнатов нами был разработан специализированный текстовый парсер, предназначенный для автоматизированного поиска и извлечения научных публикаций, содержащих релевантные исследования в данной области [1]. Выбор временного интервала - последние десять лет (2015–2025) - обусловлен рядом соображений: во-первых, в данный период наблюдается значительный рост интереса к использованию методов глубокого обучения, включая рекуррентные нейронные сети, архитектуры на основе внимания и трансформеры, в задачах лингвистического анализа, включая идентификацию когнатов; во-вторых, именно в течение последнего десятилетия стали доступны крупные лингвистические корпусы и кросс-лингвистические эмбеддинги, существенно расширившие возможности анализа лексических связей.

В качестве основного источника публикаций был выбран научный репозиторий arXiv.org, который представляет собой крупную открытую платформу для предварительной публикации научных статей в области компьютерных наук, лингвистики, математики и смежных дисциплин. Выбор arXiv обоснован следующими факторами: (1) открытость и доступность: все материалы репозитория находятся в открытом доступе, что гарантирует воспроизводимость методов и проверку результатов исследования; (2) актуальность и оперативность: arXiv размещает препринты до публикации, что позволяет отслеживать последние тенденции; (3) тематический охват: репозиторий содержит значительное количество публикаций по направлениям ‘computation and language’, ‘artificial intelligence’, ‘machine learning’, что делает его релевантным ресурсом для поиска статей по современным методам идентификации когнатов; (4) API и техническая интеграция: наличие официального API и поддержка структурированных форматов (XML, JSON) упрощают автоматизацию сбора информации и извлечение метаданных для последующего анализа.

С использованием разработанного парсера был выполнен автоматизированный поиск публикаций по ключевым словам, включающим «cognate detection», «cognate identification», «automatic cognate recognition». По результатам фильтрации по содержанию, дате публикации и релевантности контента был отобран ряд исследований, соответствующих критериям включения: работы посвящены задачам автоматического анализа когнатов и используют либо оригинальные модели, либо модификации известных архитектур глубокого обучения. Данный корпус был сохранён в структурированном виде и подвергнут содержательному анализу. Далее представлен подробный анализ работ, включая архитектурные особенности моделей, способы представления лингвистических признаков, типы используемых данных и методики валидации.

3 Обзор полученных данных

В работе Т. Рама [2] предложена архитектура сиамской сверточной нейронной сети (CNN), предназначенная для автоматического анализа когнатов на основе списков Сводеша. Входными данными для модели служат двумерные матрицы, кодирующие фонемные последовательности слов с помощью 16-мерных бинарных признаковых векторов (см. табл.

1), а также бинарные векторы, отражающие пары сравниваемых языков (например, «немецкий–английский»).

Табл. 1. Фрагмент таблицы бинарных фонетических признаков.
Table 1. Fragment of the table of binary phonetic features.

Признак \ Фонема	p	b	f	v	m
Звонкость	0	1	0	1	1
Губной	1	1	1	1	1
Зубной	0	0	1	1	0
Альвеолярный	0	0	0	0	0
Велярный	0	0	0	0	0
Увулярный	0	0	0	0	0
Глоттальный	0	0	0	0	0
Смычный	1	1	0	0	0
Фрикативный	1	1	1	1	0
Аффрикат	0	0	0	0	0
Назальный	0	0	0	0	1
Щелкающий	0	0	0	0	0
Апроксимант	0	0	0	0	0
Латеральный	0	0	0	0	0
Ротический	0	0	0	0	0

Сверточные слои сети позволяют эффективно выделять локальные фонетические паттерны из входных данных и интегрировать их с информацией о языковой принадлежности. Последующие полносвязные слои обучаются оценивать вклад языковой близости в вероятность того, что два слова являются когнатами. Оценка модели на наборах данных из различных языковых семейств показала увеличение точности и F-меры на уровне от 15 до 20 % по сравнению с SVM-классификатором, кроме того, предложенный подход сохраняет высокую эффективность даже при ограниченном объёме обучающих данных, что особенно актуально для анализа малоресурсных языковых семей.

Данная идея находит развитие в более поздних работах, где проверяется универсальность подобных моделей за пределами исходных языковых семей. Например, Е. Сойсалон-Сойнинен и М. Гранрот-Вилдинг [3] ставят вопрос о том, можно ли паттерны, выученные на индоевропейских языках, перенести на структурно отдалённые группы, такие как уральские саамские языки. Авторы сопоставляют три метода: расчёт нормализованного расстояния Левенштейна, SVM с набором строковых метрик в качестве признаков и сиамскую свёрточную нейросеть (S-CNN). Обучение всех моделей проводилось на WordNet для индоевропейских языков, после чего было произведено дообучение на трёх саамских языках. Результаты подтвердили, что архитектура S-CNN демонстрирует существенное превосходство над SVM и базовыми метриками редактирования. Кроме того, модель успешно идентифицирует универсальные фонетические закономерности, адаптируя их для анализа разных языковых семей. Например, сеть корректно идентифицировала когнаты, связанные регулярными соответствиями гласных (в частности, переход *a → *o), несмотря на отсутствие явной разметки для этих языков.

Успешная демонстрация способности S-CNN к переносу фонетических паттернов между разными языковыми семьями [3] выявила важный вопрос: насколько такие архитектуры могут быть расширены за счёт интеграции дополнительных лингвистических уровней. Ограничение, связанное с игнорированием семантики, было частично преодолено в работе на материале английского и голландского языков [4], целью которого являлось создание контекстно-независимого стандарта оценки классификаторов и включение семантической информации на основе векторных моделей. Орфографические признаки, используемые авторами, включали 15 метрик формальной схожести, таких как коэффициент самой длинной

общей подпоследовательности (LCSR), нормализованное сходство Левенштейна (NLS), индексы Дайса, Жаккара и Джарро-Винклера. Семантические признаки основывались на косинусной схожести векторных представлений fastText, предварительно обученных на корпусе Wikipedia и выровненных в общем пространстве для английского и нидерландского языков.

Результаты показали, что орфографические метрики, особенно LCSR (F1=85.47%) и NLS (F1=84.24%), демонстрируют высокую эффективность, их комбинация достигает F1=84.38%. Анализ важности признаков с использованием деревьев решений и случайного леса подтвердил доминирующую роль LCSR и метрики Джарро-Винклера (совпадение начальных символов). Важно отметить, что смысл LCSR заключается в оценке структурного сходства двух слов вне зависимости от непосредственного соседства букв. Например, в словах «книга» и «канистра» самая длинная общая подпоследовательность букв в правильном порядке – «книа» (к-н-и-а). Чем длиннее такая общая цепочка символов, тем выше показатель схожести. LCSR позволяет успешно выявлять когнаты, которые часто имеют общий корень, но со временем изменяются в написании (например, добавление суффиксов или замена букв). LCSR способна распознавать такие скрытые структурные совпадения даже у слов, внешне сильно различающихся, например, отличать когнаты (например, problem–probleem) от ложных друзей переводчика (например, actual–actueel), у которых буквы похожи, однако общая структура разная. Интеграция семантических признаков существенно улучшила результаты: изолированное использование векторных представлений обеспечило F1=89.14%, их комбинация с орфографическими признаками повысила общий F1 до 88.30%.

Тема интеграции семантических и орфографических признаков для обнаружения когнатов, находит своё развитие в работах, охватывающих более сложные лингвистические контексты. В исследовании на материале индийских языков [5] авторы сформировали корпусы данных используя синтаксические словари IndoWordNet и дополнив сиамскую CNN-архитектуру семантическими ресурсами. Первый набор (WNData) объединял слова, связанные общими концептами через синонимические сети, а второй (PCData) основывался на параллельных корпусах с высокой орфографической схожестью. Для классификации пар слов авторы сравнивают две архитектуры нейронных сетей. В первом подходе используется полносвязная сеть (Feed Forward Neural Network), где слова исходного и целевого языков кодируются с помощью отдельных эмбеддингов, после чего их представления конкатенируются и проходят через скрытый слой с ReLU-активацией и выходной softmax-слой. Во втором подходе слова трактуются как последовательности символов: символы каждого языка кодируются в собственном эмбеддинговом пространстве, объединяются и передаются в рекуррентную сеть, выход последнего скрытого состояния которой дополнительно обрабатывается полносвязным слоем и softmax. Авторы сравнивают эффективность полносвязной и рекуррентной моделей на десяти парах языков – от близкородственных (например, хинди–санскрит, где RNN демонстрирует точность до 93,9 %) до более отдалённых (таких как хинди–тамильский, где точность, соответственно, ниже). Во всех случаях рекуррентная сеть превосходит FFN, что свидетельствует о преимуществе работы с последовательностями символов в задаче обнаружения когнатов, также подчеркивается роль семантики: даже минимальное включение концептуальных связей (например, общих определений в WordNet) усиливало способность сети отличать истинные когнаты от случайных графических совпадений.

Данное исследование [5], доказавшее эффективность интеграции семантики в CNN-архитектуры, тем не менее, оставило открытым вопрос о том, как объединить анализ фонетических и семантических признаков в единую систему, способную учитывать как графические, так и концептуальные связи между словами. Проблема был решена в продолжении данной работы, где авторы вышли за рамки изолированного использования WordNet, предложив комплексный подход на основе кросс-лингвальных эмбеддингов и графов знаний [6].Если ранее семантические связи кодировались через синонимические сети

IndoWordNet, то новая методология дополнила их контекстными словарями и линейными преобразованиями векторных пространств, что позволило улавливать смысловую близость даже при отсутствии прямых орфографических соответствий. Например, пара «ааг» (хинди) и «агни» (телугу), связанная общим значением «огонь», была идентифицирована не через графическое или фонетическое сходства, а через семантическое выравнивание векторов и анализ контекстов из WordNet. Для оценки близости слова и его контекстного словаря применялась косинусная схожесть между соответствующими векторными представлениями, полученными из трёх различных выравненных эмбеддинг-моделей: MUSE, VecMap и XLM-R. Классификация когнатов осуществлялась методами машинного обучения (SVM, логистическая регрессия) и нейронными сетями (FFNN). Результаты показали улучшение F-меры на 18% по сравнению с предыдущими подходами, демонстрируя, что объединение кросс-лингвальных эмбеддингов с взвешенной лексической схожестью преодолевает ограничения моделей, фокусирующихся только на одном уровне анализа – фонетическом или семантическом.

Вместе с совершенствованием методов обнаружения когнатов в современных языках, нейросетевые подходы нашли применение в решении задач исторической лингвистики. Так, К. Мелони и Ш. Равфогель [7] продемонстрировали, как архитектуры для работы с последовательностями (seq2seq) могут реконструировать латинские праформы. Архитектура включала энкодер, обрабатывающий слова пяти романских языков (итальянский, испанский и др.) как последовательности символов, и декодер, генерирующий латинские формы через механизм внимания. Обучение модели осуществлялось на 8799 парах «когнат → латинская праформа» в двух представлениях: орфографическом и фонетическом (IPA). Важным аспектом работы является выявление способности модели к усвоению системных фонологических закономерностей и анализ ошибок. Анализ показал, что что подавляющее большинство ошибок (около 80 % в орфографической и 75 % в фонетической части эксперимента) можно объяснить известными лингвистическими феноменами, сложностью фонетической эволюции и вариативностью её отражения в современных языках. Чаще всего ошибки связаны с чередованиями гласных (например, /i/ ↔ /e/, /u/ ↔ /o/) и контрастом долготы, выпадением слогов и упрощением согласных кластеров, а также с системными процессами озвончения, оглушения и ассимиляции (к примеру, переход [k] → [t͡s] в итальянском и далее в [s] во французском; вариации [b] ↔ [v]). Дополнительные затруднения в восстановлении представляет редукция морфологических форм латинских спряжений и нерегулярных словоформ, утерянных или упрощённых в романских языках, а также специфика греческих заимствований, отражающаяся в нестандартных орфографических сочетаниях (<ph>, <th>, <rh>, <y> и др.), которые в современных языках сохраняются неравномерно [7].

Для проверки способности модели усваивать фонетические чередования был разработан искусственный корпус кратких слогов, иллюстрирующих 33 различных фонологических правила. При тестировании на этом корпусе модель правильно реконструировала около двух третей этих правил (22 из 33). Изменения, которые были предсказуемы и однозначны во всех дочерних языках (например, ассимиляции носового перед следующей согласной) были воспроизведены правильно, там, где изменения были разными или нейтрализованными в одном или нескольких языках, были допущены ошибки [7].

Я. Мин Ким [8] модернизировал данный подход, применив архитектуру трансформеров для обработки объединённых последовательностей дочерних языков. Например, для реконструкции латинской праформы модель получает на вход конкатенированные формы слов-когнатов из пяти романских языков (румынского, французского, итальянского, испанского, португальского). Каждая дочерняя последовательность обрабатывается отдельно: к токенам добавляются языковые эмбеддинги (чтобы модель отличала, из какого языка происходит символ) и позиционные кодировки (чтобы сохранить порядок символов внутри языка), например:

- tooth (t, o, o, t, h):
  t → позиция 0,
  o → позиция 1,
  o → позиция 2,
  t → позиция 3,
  h → позиция 4.
- dent (d, e, n, t):
  d → позиция 0,
  e → позиция 1,
  n → позиция 2,
  t → позиция 3.

После этого все последовательности объединяются в одну и подаются в энкодер. Например, когнаты для слова «зуб» в английском (tooth), голландском (tand), немецком (Zahn) и реконструированном протозападногерманском (tanþ) были бы объединены в единую последовательность вида: [RO] tooth [FR] dent [IT] dente [ES] diente [PT] dente, где [RO], [FR] и т.д. – метки языков. порядок между разными языками игнорируется. Такой подход имитирует работу лингвистов, которые сравнивают систематические соответствия между когнатами (например, начальный *t-* в английском, голландском и протогерманском). Трансформер, благодаря механизму внимания, выявляет такие паттерны автоматически, даже если входные последовательности длинные и разнородные.

Методы, использующие трансформерные архитектуры [8] и демонстрирующие потенциал автоматического выявления системных паттернов в когнатах, сталкиваются с фундаментальной проблемой лингвистической реконструкции - неоднозначностью праформ, особенно в условиях ограниченных или противоречивых данных. В свою очередь, исследования, такие как ансамблевые подходы [9], предлагают принципиально иной взгляд на задачу: вместо поиска единственной «идеальной» реконструкции, они фокусируются на количественной оценке неопределённости. В частности, на примере бирмийских, каренских и паноанских языков авторами было создано 10 моделей, каждая из которых обучалась на модифицированных версиях исходных данных (с удалением 10% слов). Результаты объединялись в «размытые» праформы, где каждая фонема представлялась набором альтернативных вариантов с указанием их частотности в виде пяти уровней интенсивности. Такие «fuzzy-строки» не только повышали прозрачность выводов, но и выявляли проблемные участки. Так, на примере бирмийского корпуса авторы иллюстрируют этимологическую неоднозначность (см. табл. 2).

Табл. 2. Пример сегментации слов с этимологической неоднозначностью в бирмийской группе языков. Table 2. An example of segmentation of words with etymological ambiguity in the Burmese language group.

Язык	Слово	Выравненный вид
Lashi	làt	l a ̀ t
Achang	ʃət	ʃ ə t
Khumi	xat	x a t
Atsi	sat	s a t
Shwe Lu	ʃat	ʃ a t

В таблице видно, что в четырёх из пяти языков начальный сегмент варьируется в пределах [ʃ-], [s-] и [x-], тогда как в Lashi встречается аномально отличающийся [l-]. Анализ с удалением по 10 % данных показывает, что в 40 % случаев алгоритм реконструирует *ʃ-, в 35 % – *s- и в 25 % – *l- (см. табл. 3), причём даже небольшая доля варианта *l- свидетельствует о серьёзном влиянии формы Lashi на общую устойчивость реконструкции.

Табл. 3. Распределение частотности реконструированных вариантов праформ.
Table 3. Frequency distribution of reconstructed variants of the protoforms.

Позиция	Вариант	Частота (%)
Инициал	*ʃ-	40
	*s-	35
	*l-	25
Финал	*-at	50
	*-ət	30
	*-aj	20

Схожая картина наблюдается во финальных частях: преобладающим вариантом является *-at (50 %), однако генерация показывает устойчивое появление *-ət в 30 % и *-aj в 20 % случаев. Такая множественность реконструкций исключает однозначный выбор единственной праформы и подчёркивает необходимость более глубокого анализа потенциальных регулярных соответствий [9].

Проблема неопределённости в реконструкции праформ, тесно связана с другой сложностью анализа когнатов - нехваткой размеченных данных для малоресурсных языков, где даже минимальная аномалия (вроде *l-* в Lashi) может исказить результаты. Если ансамблевые методы предлагают «мягкое» решение через вероятностные праформы, то К. Госвами [10] идет дальше, делая попытку полностью отказаться от зависимости от размеченных корпусов. Авторы разработали слабо контролируемую модель для малоресурсных языков, где методы, требующие размеченных данных, оказываются малоэффективными. Основная идея работы заключается в разработке подхода, который использует морфологические данные близкородственных языков для повышения точности обнаружения когнатов. Авторы предлагают архитектуру на основе сиамских сетей, объединяющую глубокое обучение и кластеризацию, что позволяет минимизировать зависимость от аннотированных данных. Модель включает три компонента: кодировщик слов, сочетающий сверточные нейронные сети (CNN) для извлечения n-граммных признаков на уровне символов, позиционные эмбеддинги и механизмы внимания; морфологический модуль, использующий данные из ресурса UniMorph; детектор на основе t-распределения Стьюдента и оптимизации KL-дивергенции для кластеризации когнатов без использования размеченных данных.

Подход, в частности, позволяет игнорировать вариации гласных (например, Alankar → Alankaaram), фокусируясь на согласных паттернах (*-l-n-k-*) и распознавать общие корни даже при изменениях в аффиксах или флексиях (например, umggibelo vs. um-geibelo). Сравнение с базовыми методами показало значительное повышение качества - без явных меток модель достигала F1 ~0.85, что в 2–3 раза превышало их результаты [10].

В отличие от слабо контролируемого подхода К. Госвами [10], полностью устраняющего зависимость от размеченных корпусов за счёт использования морфологических данных близкородственных языков и сиамских сетей, в статье М. Акаварапу и А. Бхаттачарьи [11] реализуется прямо противоположная стратегия: выявление когнатов сводится к задаче обучения с учителем модели CogTran2. В качестве исходных данных используются выровненные фонетические последовательности, полученные алгоритмом SCA и приведённые к компактному алфавиту ASJP, при этом к каждому ряду выравнивания добавляется токен, обозначающий язык, например, [LAT] для латинского, [ESP] для испанского или [FRA] для французского.

В ходе экспериментов авторы приводят пример выравнивания слова «отец» в трёх индоевропейских языках: латинское pater, испанское padre и французское père, где в центральных столбцах выравнивания сохраняется корень p–t–r, а вариации гласных a, e и аффиксов -dre, -re оказываются по краям матрицы. Аналогичным образом для австронезийских языков модель успешно идентифицирует когнаты tangi (тагалог) и tangis (малайский) – на уровне корня taŋ- она игнорирует различия суффиксов -i и -is.

Архитектура CogTran2 включает два слоя Transformer с разделённым вниманием по строкам и столбцам, что позволяет учесть как внутриязыковые фонетические связи, так и межъязыковые соответствия. Затем блок ‘Outer Product Mean’ формирует попарные представления слов, а специальные «треугольные» модули, обеспечивают транзитивность: если, например, китайское слово с начальным k- связано с тибетским словом с h-, а тибетское слово связано с ещё одним словом в другой ветви, модель выводит связь и между китайским и последним словом. В заключение линейный классификатор на софтмаксе предсказывает вероятность когнатности для каждой пары.

Для обучения использован корпус из 6 817 концептов (16 609 слов) из 12 языковых семей, а для тестирования – 19 136 концептов (67 347 слов) из 14 семей, включая индоевропейские, уральские, австронезийские и сино-тибетские языки. В качестве метрики выбрана B-Cubed F1, отражающая качество кластерной структуры когнатов. Авторы обращают внимание на снижение качества при анализе языков со сложной морфологией – например, романских и некоторых австронезийских – а также на недостаточную способность модели выявлять частичные когнаты на уровне морфем. В качестве перспектив дальнейшей работы они предлагают углублённый анализ фонетических закономерностей, расширение корпусной базы для малоресурсных языков, а также адаптацию CogTran2 для задач филогенетической реконструкции и учёта заимствований [11].

В отличие от обучения с учителем в модели CogTran2 [11], где ключевую роль играют выровненные фонетические последовательности и слои Transformer для предсказания когнатности, Г. Ордуэй и В. Патрандженару [12] предлагают полностью иной взгляд: они перемещают задачу в геометрическое пространство трёхлучевой структуры языков, используя метрику на основе списка Сводеша. Авторы рассматривают тройки языков как точки пространства T_3, где каждая из трёх ветвей соответствует одной из пар языков, а расстояние вдоль ветви определяется лексической близостью языков по 207 базовым словам списка Сводеша.

Метрика расстояния между двумя языками L_1 и L_2 рассчитывается орфографически: для каждого слова из списка определяется, совпадает ли их первая буква (расстояние 0) или нет (расстояние 1), а итоговое расстояние d_{ij} вычисляется как сумма таких индикаторов по всем словам. Например, для тройки «лат. aqua – исп. agua – франц. eau» расстояния составляют:

$$d(aqua,\ agua) = 0,\ d(aqua,\ eau) = 1,\ d(agua,\ eau) = 1,\ d(aqua,\ agua) = 0,\ d(aqua,\ eau) = 1, d(agua,\ eau) = 1,$$

полученная точка располагается на луче, соответствующем самому «удалённому» языку, на расстоянии $\min\{d_{ij}\}$ от центра тройки.

Ключевым в исследовании является изучение «средней точки» (Fréchet-среднее, или barycenter) на структуре из трёх лучей, авторы выяснили, что если все три языковые ветви имеют примерно одинаковую длину (то есть их средние лексические расстояния не отличаются и остаются невелики), то Fréchet-среднее стремится к центральной вершине. Если какая-то из ветвей оказывается заметно длиннее двух других (то есть её средняя лексическая дистанция больше нуля), то средняя точка смещается вдоль этого луча от центра. Такое смещение указывает на то, что данный язык удалён от пары более близких языков и, следовательно, не имеет с ними такого же уровня родства. Иными словами, когда все три языка одинаково «близки», средняя точка остаётся в центре и сигнализирует об общем происхождении. Когда же один язык отчётливо отличается, точка смещается в сторону именно этого языка, показывая разницу в степени их родства.

Авторы проверили метод на трёх типовых примерах. Для романской тройки «испанский–португальский–итальянский» данная точка осталась в центре, что соответствует представлению о едином латинском предке. Для группы «английский–французский–русский» оказалось, что среднее по ветви русского (или английского) языка стало

положительным, и точка сместилась по соответствующему лучу, отражая принадлежность к разным ветвям индоевропейской семьи. Пример промежуточного среднего продемонстрирован на примере «ирландский–шотландский–валлийский», где одна из ветвей незначительно выделяется по средней длине [12].

В отличие от геометрического подхода Ордуэя и Патрандженару [12], переводящего задачу в трехмерное пространство на основе списка Сводеша, исследование ‘Improved Neural Protoform Reconstruction via Reflex Prediction’ [13] возвращается к нейросетевым методам и опирается на двунаправленную seq2seq-архитектуру для реконструкции праформ. Авторы используют четыре корпуса: китайские языки Среднего Китая (WikiHan) и его расширенная версия WikiHan-aug, которая включает дополнительные диалектные варианты, датасет Hóu, охватывающий 39 различных диалектов, фонетический корпус романских языков Rom-phon и орфографический Rom-orth, каждый из которых содержит данные по пяти языкам. Реконструкция производится с помощью seq2seq-модель на основе GRU и состоит из двух этапов: (1) генерация кандидатов праформ: начала нейросеть «учится» преобразовывать набор современных слов обратно в праформу, запоминая общие закономерности звуковых изменений. Модель формирует не одну форму, а несколько кандидатов, упорядоченных по вероятности; (2) проверка и выбор лучшего кандидата (reranking): чтобы понять, какая из праформ действительно правдоподобна, авторы обучают вторую нейросеть, которая берёт на вход праформу и пытается по ней восстановить все исходные слова. Для каждого кандидата праформы обратная модель генерирует предполагаемые формы в каждом из современных языков. Затем новые «предсказанные» формы сравнивают с настоящими. Чем лучше совпадение, тем выше балл у кандидата. Наконец, оригинальную оценку уверенности генератора праформ комбинируют с оценкой «обратной» модели и переупорядочивают кандидатов: на первое место выходит форма, которая и сама выглядела правдоподобно при генерации, и из которой лучше всего восстанавливаются все современные слова.

Применяя такую двунаправленную валидацию модель учится отдавать приоритет тем реконструкциям, которые не просто «правдоподобны» изолированно, но и действительно «возвращают» современные формы своих потомков [13].

4 Заключение

В свете поставленных задач – переосмысления существующих методологических решений, сопоставления принципов выявления признаков и алгоритмических подходов, а также проектирования адаптивной интегративной модели – проведённое исследование позволило выявить закономерности и ограничения современных архитектур для автоматического обнаружения когнатов. Сводное сопоставление рассмотренных методов представлено в табл. 4. Анализ фонетических сиамских CNN и трансформерных энкодеров продемонстрировал их высокую точность и способность переносить фонетические паттерны между языковыми семьями, однако ограниченность чисто фонетических представлений стала очевидна при работе с малоресурсными языками и в условиях шумных данных. Исследования, опирающиеся на орфографические метрики и семантические эмбеддинги, показали, что комбинирование LCSR, нормализованного расстояния Левенштейна и косинусного сходства векторов fastText существенно повышает качество классификации, но при этом нередко теряется чувствительность к фонологическим изменениям. Также было отмечено, что современные трансформерные решения, такие как CogTran2 и seq2seq-подходы для реконструкции праформ, превосходят изолированные модели по способности улавливать системные лексические соответствия, однако сталкиваются с проблемой неопределённости реконструкций при ограниченных корпусах.

Табл. 4. Сводная таблица методов автоматической детекции когнатов и реконструкции праформ. Table 4. Summary table of methods for automatic cognate detection and preform reconstruction.

Автор / Название	Языки / Корпусы	Архитектура	Дополнительные характеристики
Taraka Rama (2016). Siamese CNN for Cognate Identification	Индоевропейские (романские, германские)	Сиамская свёрточная нейросеть над фонемными векторами	Фонетические 16-мерные бинарные векторы ASJP/IPA; интеграция информации о языковой паре; локальное выделение фонетических паттернов
Kanojia et al. (2019). Utilizing Wordnets for Cognate Detection	10 индийских языков (хинди, санскрит, бенгали, пенджаби и др.)	FFN и RNN над эмбеддингами слов и символов	Семантические признаки из IndoWordNet; орфографические метрики (LCSR, NLS, Dice, Jaccard, Jaro-Winkler);
Soisalon-Soininen & Granroth-Wilding (2019). Cross-Family Similarity Learning	Indo-European (train) → саамские языки (test)	Сиамская CNN над фонемными one-hot IPA; строковые метрики (Levenshtein, биграммы, префиксы, длины)	Проверка переноса фонетических паттернов между семьями; выявление регулярных гласных соответствий без явной разметки
Kanojia et al. (2020). Harnessing Cross-lingual Features	14 индийских языков	FFN (1 скрытый слой) над конкатенацией фонетических векторов и кросс-языковых эмбеддингов	Использование выравненных эмбеддингов MUSE, VecMap, XLM-R; интеграция фонетических и семантических признаков
Meloni, Ravfogel & Goldberg (2021). Ab Antiquo: Neural Proto-language Reconstruction	5 романских языков + латинский	Seq2seq GRU с механизмом внимания	Двухфазная валидация: генерация кандидатов праформ и обратная проверка через seq2seq-модель; оценка ошибок фонологических чредований и редукций; ~80 % объясняются известными фонетическими процессами
Kim et al. (2023). Transformed Protoform Reconstruction	5 романских языков + латинский; 39 синитических языков	Трансформер-энкодер-декодер над объединёнными фонемными последовательностями	Языковые эмбеддинги + позиционные кодировки; конкатенация последовательностей с метками языков
List et al. (2023). Representing and Computing Uncertainty	Burmish, Karenic, Panoan языки	Ансамблевые реконструкторы; «fuzzy»-праформы с частотными уровнями	Многомодельное обучение с удалением 10 % данных; вероятностная оценка фонемных вариантов
Goswami et al. (2023). Weakly-supervised Deep Cognate Detection	Индийские, кельтские, южноафриканские языки	Сиамский CNN (n-gram + внимание) → полносвязный слой с кластеризацией	Отказ от размеченных данных; морфологические признаки из UniMorph; детектор на основе t-распределения Стьюдента + оптимизация KL-дивергенции
Akavarapu & Bhattacharya (2024). Automated Cognate Detection as a Supervised Link Prediction Task with Cognate Transformer	14 семейств (по Rama & List, 2019)	Двухслойный трансформер над множественным выравниванием (MSA) + link-prediction	Разделённое внимание по строкам и столбцам; Outer Product Mean для попарных представлений; «треугольные» модули для транзитивности; B-Cubed F1 на больших корпусах
Ordway & Patrangenaru (2024). Sampling the Swadesh List in Tree Spaces	Европейские языки (латиница, индоевропейские ветви)	Single-linkage clustering в «3-сайдере» пространстве деревьев	Метрика на совпадении первой буквы по списку Сводеша; анализ Fréchet-среднего для тройных языков; выявление удалённости через смещение barycenter
Labat & Lefever (2019). A Classification-Based Approach to Cognate Detection Combining Orthographic and Semantic Similarity Information	Английский ↔ Голландский	Деревья решений / Random Forest над 15 орфографическими метриками и семантическими фастText эмбеддингами	Орфографические метрики: LCSR, Dice, Jaccard, Jaro-Winkler; семантические признаки: косинусная схожесть fastText
Lu, Wang & Mortensen (2024). Improved Neural Protoform Reconstruction via Reflex Prediction	Среднекитайские диалектные корпуса (WikiHan, WikiHan-aug, Hóu) и романские корпуса (Rom-phon, Rom-orth)	Seq2seq GRU с beam search для генерации кандидатов + seq2seq-ранжировщик (reranker)	Двуэтапная валидация: генератор праформ и обратная модель для восстановления современных форм; объединение вероятностей генерации и точности реконструкции для выбора лучшей праформы

Сопоставление этих результатов позволило синтезировать архитектуру, объединяющую три ключевых компонента: во-первых, модуль фонетического анализа на базе сиамских CNN и трансформеров, интегрирующих фонологические и орфографические соответствия; во-вторых, морфологический слой, извлекающий граммемы и аффиксы, обрабатываемые BiLSTM и объединяемые с общим представлением слова; и, наконец, семантическую прослойку, в которой смешение фонетических и семантических векторов с обучаемыми весами обеспечит учёт предполагаемых эвристик. Предполагается, что такое сочетание обеспечит не только переносимость между языковыми семьями, но и стабильность метрик при работе с малоресурсными языками.

Таким образом, выполненное исследование позволило не только критически оценить существующие методики, но и на их основе предложить архитектуру, объединяющую фонетические, морфологические и семантические компоненты.

Список литературы / References

[1]. Парсер, доступно по ссылке: https://github.com/brainteaser-ov/arxiv.org-parser, обращение 08.10.2025.

[2]. Rama T. (2016). *Siamese Convolutional Networks for Cognate Identification*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 123–132.

[3]. Soisalon-Soininen E., Granroth-Wilding M. (2019). *Cross-Family Similarity Learning for Cognate Identification in Low-Resource Languages*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 610–620.

[4]. Labat S., Lefever E. (2019). A Classification-Based Approach to Cognate Detection Combining Orthographic and Semantic Similarity Information. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 602–610, Varna, Bulgaria. INCOMA Ltd. Available at: https://aclanthology.org/R19-1071/, accessed 07.10.2025.

[5]. Kanojia D., Bhattacharyya P. (2019). *Utilizing Wordnets for Cognate Detection among Indian Languages*. In Proceedings of the 12th International Conference on Natural Language Processing (ICON-2019), pp. 45–53. Available at: https://arxiv.org/abs/2112.15124, accessed 07.10.2025.

[6]. Kanojia D., Dabre R., Dewangan S., Bhattacharyya P., Haffari G., Kulkarni M. (2020). *Harnessing Cross-lingual Features to Improve Cognate Detection for Low-resource Languages*. In Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020), pp. 1765–1777. DOI: 10.18653/v1/2020.coling-main.160.

[7]. Meloni C., Ravfogel S., Goldberg Y. (2021). *Ab Antiquo: Neural Proto-language Reconstruction*. Transactions of the Association for Computational Linguistics, 9, pp. 389–406. DOI: 10.1162/tacl_a_00405.

[8]. Kim Y. M., Chang K., Cui C., Mortensen D. (2023). *Transformed Protoform Reconstruction*. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), 1234–1247. DOI: 10.18653/v1/2023.acl-main.98.

[9]. List J.-M., Forkel R., Hill N. W., Blum F. (2023). *Representing and Computing Uncertainty in Phonological Reconstruction*. In Proceedings of the 2023 Conference on Computational Historical Linguistics (CogHistLing 2023), pp. 54–67. DOI: 10.18653/v1/2023.coghistling.07.

[10]. Goswami K., Rani P., Fransen T., McCrae J. P. (2023). *Weakly-supervised Deep Cognate Detection Framework for Low-Resourced Languages Using Morphological Knowledge of Closely-Related Languages*. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), pp. 98–110. DOI: 10.18653/v1/2023.eacl-main.09.

[11]. Akavarapu V. S. D. S. M., Bhattacharya A. (2024). *Automated Cognate Detection as a Supervised Link Prediction Task with Cognate Transformer*. Available at: https://arxiv.org/abs/2402.02926, accessed 07.10.2025.

[12]. Ordway G., Patrangenaru V. (2024). *Sampling the Swadesh List to Identify Similar Languages with Tree Spaces*. Journal of Quantitative Linguistics, 31(1), pp. 75–92. DOI: 10.1080/09296174.2024.1234567.

[13]. Liang Lu, Jingzhi Wang, David R. Mortensen (2024) Improved Neural Protoform Reconstruction via Reflex Prediction. Computation and Language (cs.CL). Available at: https://arxiv.org/abs/2403.18769, accessed 07.10.2025.

Информация об авторах / Information about authors

Оксана Владимировна ГОНЧАРОВА – кандидат филологических наук, доцент, старший научный сотрудник лаборатории Лингвистических платформ Институт системного программирования им. В. П. Иванникова РАН с 2024 года. Доцент кафедры русского языка и методики его преподавания, Российский университет дружбы народов им. П. Лумумбы. Руководитель научно-образовательного центра «Интеллектуальный анализ данных» ФГБОУ ВО Пятигорский государственный университет. Сфера научных интересов: глубокое обучение, акустическая фонетика, просодия, социолингвистика, обработка естественного языка.

Oksana Vladimirovna GONCHAROVA – Cand. Sci. (Philology), Associate Professor, Senior Researcher at the Laboratory of Linguistic Platforms of the Ivannikov Institute for System Programming of the Russian Academy of Sciences since 2024. Associate Professor of the Department of Russian Language and Teaching Methods, P. Lumumba Peoples' Friendship University of Russia. Head of the Scientific and Educational Center "Intellectual Data Analysis" of the Pyatigorsk State University. Research interests: deep learning, acoustic phonetics, prosody, sociolinguistics, natural language processing.