УНИВЕРСИТЕТ ИТМО

International Culture & Technology Studies
online multimedia journal

ITMO UNIVERSITY

Культура и технологии
электронный мультимедийный журнал

Journal Homepage: https://cat.itmo.ru
ISSN 2587-800X

Адрес статьи / To link this article: https://cat.itmo.ru/ru/2026/v11-i1/628

Опыт и перспективы применения искусственного интеллекта для исследования иероглифических рукописей на вымерших языках

Е. В. Ягунова, Н. Д. Костыгов

Санкт-Петербургский государственный университет, Россия

iagounova_elena@mail.ru, st123478@student.spbu.ru

Аннотация. В данной статье рассматриваются подходы к обработке и анализу малоресурсных и фрагментарных корпусов, имеющих отношение к иероглифическим письменностям вымерших языков. Основное внимание уделяется применению методов машинного обучения, включая генеративные, сравнительные и контекстно-ориентированные подходы, в условиях ограниченности и неоднородности исходных данных. Подчеркивается, что специфика подобных корпусов требует комплексного учета ряда факторов, связанных с особенностями представления и интерпретации письменных источников упомянутых языков. Отмечается значимость разработки и применения методов, направленных на повышение устойчивости анализа и расширение возможностей обработки данных в условиях их дефицита. В этом контексте рассматриваются общие направления развития современных технологий, ориентированных на работу с подобными материалами, а также их роль в формировании новых исследовательских практик. Особое внимание уделяется потенциалу интеграции различных подходов, обеспечивающей более гибкое и адаптивное решение задач анализа. Делается вывод о том, что использование методов машинного обучения открывает перспективы для дальнейшего развития цифровой палеографии и создания интеллектуальных систем поддержки экспертной деятельности, способствуя более системному и масштабируемому изучению письменного наследия вымерших языков.

Ключевые слова: компьютерная лингвистика, искусственный интеллект, вымершие языки, иероглифические письменности, малоресурсные корпуса, машинное обучение

1. Введение

Последнее десятилетие, вследствие стремительного сближения методов машинного обучения и компьютерной лингвистики, ознаменовалось масштабированием применения технологий искусственного интеллекта для решения комплексных прикладных задач в области языкознания [1]. Следует отметить, что компьютерная лингвистика является одной из ключевых технологических реализаций прикладной лингвистики. Если последняя ставит своей целью решение практических коммуникативных и социальных задач в различных сферах человеческой деятельности [2], то компьютерная лингвистика обеспечивает инструментарий для их автоматизации и масштабного применения.

Особое место среди таких инструментов занимают методы машинного обучения. По сути, это область искусственного интеллекта, ориентированная на разработку алгоритмов и моделей, позволяющих вычислительным системам автоматически выявлять закономерности и улучшать свои результаты на основе данных без явного программирования. Эти алгоритмы обучаются на примерах, формируют обобщаемые модели и применяются для выполнения задач и предсказаний в новых условиях [3].

Тем не менее, сама природа иероглифических источников ставит перед ними ряд нетривиальных вызовов, требующих осторожного и взвешенного подхода. В этом отношении особую научную ценность и методологическую сложность представляет собой область палеографического анализа иероглифических рукописей, относящихся к вымершим языковым системам.

Расшифровка таких письменностей сопряжена с крайне ограниченным количеством источников, отсутствием потомков языка и неопределённостью в интерпретации символов, что делает анализ сложным и экспертозависимым [4]. Интеграция алгоритмов глубокого обучения, компьютерного зрения и статистического моделирования языка открывает принципиально новые возможности для более объективного и масштабируемого исследования подобных корпусов.

Для понимания задач исследования текстового наследия вымерших языков необходимо обратить внимание на особенности различных иероглифических систем, поскольку именно эти особенности определяют методологические подходы к их анализу и автоматизации обработки.

Иероглифическое письмо представляет собой логографо-фонетическую систему, в которой каждый знак сохраняет устойчивую визуально насыщенную форму и одновременно кодирует морфемный и фонетический уровень, создавая плотную сеть семантических и звуковых связей. Такая система формирует комплексное представление языка через интеграцию смысла и звучания в каждом графическом элементе, вместо последовательного разложения на отдельные фонемы [5].

В отличие от алфавитных систем, основанных на фонемном принципе и явной сегментации слов, восточноазиатские письменности характеризуются иным уровнем графической репрезентации и организации текста, что существенно влияет на интерпретацию и обработку письменной речи.

Наиболее известными иероглифическими системами являются и поныне используемые китайская и японская письменности, которые в разной степени опираются на иероглифический принцип, при котором знак соотносится с морфемой или морфемо-слогом и не предполагает автоматической реконструкции звучания, тогда как корейская письменность является алфавитной, но сохраняет слогоблочную графическую организацию и морфологическую прозрачность, унаследованную от более ранней логографической традиции [6, 7, 8]. Эти типологические различия обуславливают специфические требования к лингвистическому анализу и корпусному моделированию, прежде всего в области сегментации и морфологической аннотации [9].

Письменности майя и древнеегипетская иероглифика представляют собой типологически несвязанные и ныне мёртвые системы письма, для которых ключевыми являются практические проблемы формализации и машинной обработки текста [10, 11, 12]. В обеих системах знак функционирует как сложная графическая единица, что затрудняет автоматическое выделение линейных единиц и требует многоуровневого представления данных в корпусах. Отсутствие стандартизованной орфографии, высокая графическая вариативность и зависимость чтения от контекста вынуждают при создании корпусов явно разделять графемный, транслитерационный и лингвистический уровни разметки.

Для задач искусственного интеллекта это означает необходимость специализированных моделей, сочетающих визуальный анализ, контекстную интерпретацию и формализованные экспертные сведения о графических единицах, поскольку прямое применение методов, разработанных для современных алфавитных письменностей, оказывается недостаточным для корректного анализа и интерпретации этих письменных систем.

Исследованию существующего опыта и перспектив дальнейшего применения современного инструментария в исследовании текстового наследия вымерших иероглифических языков и

посвящена настоящая статья. Она является подготовительной частью для более крупного проекта по построению корпуса тангутского языка.

2. Методология

В данной статье будут рассматриваться современные системы и модели, актуальные для периода с 2022 по 2025 год. Выбор дат обусловлен важными изменениями в области искусственного интеллекта, в частности, выходом ChatGPT и стремительным развитием генеративных нейронных сетей, что оказало значительное влияние на сферу компьютерной лингвистики [13].

Объектом анализа в статье являются проекты, посвящённые изучению иероглифических письменностей вымерших языков, включая, но не ограничиваясь, древнеегипетскую, древнекитайскую письменность майя и ряд типологически сопоставимых систем.

Живые языки с иероглифической письменностью, такие как японский и корейский, сознательно исключены из рассмотрения, поскольку их анализ опирается на иные методологические предпосылки, включая наличие обширных современных корпусов, стандартизированной орфографии и активной языковой практики.

Аналогичным образом в статье не рассматриваются алфавитные и слоговые системы письма, поскольку методы, разработанные для фонемно-ориентированных письменностей, не в полной мере отражают специфику визуально сложных иероглифических систем и требуют отдельного аналитического подхода.

3. Обзор основных направлений

Одной из определяющих характеристик иероглифических письменностей вымерших языков является крайняя ограниченность доступных данных. Корпуса, как правило, включают небольшое число памятников, содержащих высокую долю редких или уникальных знаков, часто представленных в повреждённом виде. Это существенно ограничивает применимость классических подходов и требует разработки специализированных методов обучения, ориентированных на малоресурсные сценарии.

В связи с этим современные исследования, посвящённые применению искусственного интеллекта к анализу иероглифических рукописей на вымерших языках, демонстрируют значительное методологическое разнообразие. Вместе с тем, анализ литературы позволяет сгруппировать существующие подходы в ограниченное число устойчивых направлений, каждое из имеет свои архитектурные особенности.

3.1. Синтетическое расширение данных

Для компенсации ограниченности корпусов и высокой стоимости ручной разметки уже долгое время применяются методы генерации синтетических и частично размеченных данных. Пусть исторически данные подходы и развивались для решения задач, связанных с живыми языками, опыт последних лет показывает, что они также успешно могут служить для изучения древних языков [14].

В обработке естественных языков генеративный искусственный интеллект позволяет автоматически создавать сгенерированные наборы данных, которые могут воспроизводить структуру, стиль и смысл реальных данных, тем самым значительно облегчая обучения, снижая затраты на разметку и расширяя возможности моделей в условиях ограниченного объёма исходных данных, что существенно улучшает точность моделей [14, 15, 16, 17]. Однако, при использовании синтетических данных следует помнить, что неправильный выбор алгоритмов может повлечь за собой увеличение разнообразия данных, изменение распределения выборки, потерю информации, снижение качества в малых выборках, трудности с длинными текстами и низкоресурсными языками, а также другие нежелательные эффекты [18,19].

На практике используются как генеративные языковые модели для текстовых корпусов, так и генеративные нейросети для графических данных, что позволяет воспроизводить вариативность графики и расширять контекстное и лексическое разнообразие обучающих выборок [20, 21, 22, 23].

Методы синтетического расширения данных имеют большое количество вариаций, и, например, включают использование 3D-моделей, генерацию данных на основе цифровых шрифтов, а также комбинированные пиксельных и векторных подходы. Эти подходы позволяют уменьшить проблему низкоресурсности, повысить точность OCR и сегментации, а также создавать обучающие наборы, сопоставимые по качеству с реальными памятниками [24, 25].

Помимо генерации новых образцов для редких классов, синтетические методы включают аугментацию уже существующих данных: вращение, размытие по кривизне, тональную коррекцию, и другие операции, связанные с обработкой формы, что обеспечивает балансировку классов и улучшение качества обучающих данных [20, 22, 23, 26]. Экспериментальные исследования показывают, что включение синтетических данных стабильно повышает качество моделей сегментации и распознавания и улучшает переносимость классификаторов между различными памятниками [23, 27].

Таким образом, современные методы синтетического расширения и аугментации данных доказали свою эффективность для малоресурсных корпусов древних иероглифических письменностей, создавая возможности для точного анализа и автоматической обработки языков с ограниченной представленностью в источниках.

3.2. Сравнительный межъязыковой анализ

Существенную роль в языковых исследованиях последних лет играет сопоставительный анализ. Он уже успешно применялся с некоторыми малоресурсными алфавитными языками [28,29], но несмотря на культурные и хронологические различия, многие иероглифические системы также демонстрируют типологические сходства в организации знаков, композиции надписей и принципах графемного варьирования. Это открывает возможность использования методов машинного обучения для выявления устойчивых визуальных и структурных паттернов и переноса знаний между родственными или типологически сходными письменностями [30].

Современные исследования применяют комбинированные подходы, позволяющие сравнивать формы знаков, структуру надписей и распределение графем. Такие методы позволяют выявлять как универсальные когнитивные паттерны, включая семантические классификаторы, так и особенности отдельных письменностей, уточняя культурно-исторические связи между корпусами. Структурные закономерности из хорошо изученных систем применяются для распознавания ранее невиданных или частично утраченных знаков [30, 31, 33].

Методы сравнительного анализа также применяются для изучения эволюции письменных форм. Модели, сравнивающие визуально-структурное сходство графем в разных исторических периодах, позволяют количественно оценивать изменения начертаний и выявлять периоды преемственности или стилистических изменений [34, 35].

В таких подходах особое внимание уделяется моделированию внутренней структуры знаков через последовательности штрихов, радикалов и структурных элементов, что позволяет переносить знания на новые формы и восстанавливать утерянные элементы [32, 35, 36]. Подобные стратегии создают возможность выявления семантических и структурных соответствий между древними и современными письменностями. Такие подходу также активно применяются в исследовательской деятельности [34].

Исходя из этого, мы можем сказать, что современные методы сравнительного межъязыкового и межписьменного анализа объединяют визуально-структурное моделирование, перенос знаний между системами и изучение эволюции графем. Они позволяют количественно и качественно исследовать сходство и развитие письменностей, выявлять универсальные когнитивные и культурно-исторические закономерности и повышать точность распознавания и анализа знаков.

3.3. Контекстно-зависимый анализ

Помимо проблем с малочисленностью источников, иероглифические системы вымерших языков не всегда допускают однозначное распознавание знаков исключительно на основе их

визуальной формы. В условиях повреждённости носителей, высокой графической вариативности и отсутствия стандартизированных орфографических канонов визуальный анализ часто оказывается недостаточным для однозначной интерпретации знаков, что делает контекст ключевым источником информации и оправдывает развитие методов контекстно-зависимого моделирования и восстановления текста.

Эти методы рассматривают надпись как последовательность взаимозависимых элементов и формулируют задачи заполнения лакун и идентификации неоднозначных знаков как задачи вероятностного предсказания с использованием языковых моделей и нейросетевых архитектур. Подходы такого класса позволяют учитывать синтаксические, частотные и структурные закономерности, недоступные чисто визуальным методам, и применяются как для практической реконструкции, так и для количественного анализа свойств письменности [37, 38, 39].

Одно направление исследований представляет собой чисто лингвистическое моделирование последовательностей знаков. В них модели на основе рекуррентных сетей применяются для предсказания вероятных знаков в контексте последовательности и в ряде случаев демонстрируют устойчивость при дефиците контекста, что делает их полезным дополнением к компьютерно-зрительным модулям при реконструкции повреждённых фрагментов [33].

Параллельно развиваются методы восстановления изображений, явно учитывающие визуальный контекст. Общей идеей этих подходов является поэтапное или структурированное моделирование: сначала восстанавливаются ключевые структурные элементы, такие как контуры и границы, а затем с учётом этих подсказок восполняется цвет и текстура. Эта концепция реализуется как через многоэтапные состязательные архитектуры, где комбинируются дальнодействующие зависимости и локальные детали. Такой подход позволяет улучшить согласованность реконструируемых элементов и качество текстуры, особенно при недостатке информации о повреждённых областях, показывая превосходство над одношаговыми или блочно-сопоставительными методами [40, 41].

Наиболее эффективными оказываются гибридные мультимодальные методы, объединяющие лингвистический и визуальный контексты. В таких системах языковая модель генерирует кандидатные восстановления утраченных знаков, а компьютерное зрение анализирует сохранившиеся штрихи и структуру для ранжирования кандидатов и снижения неопределённости. Аналогично, историко-культурный контекст может быть интегрирован в генеративные мультимодальные системы для комплексного анализа текста и поиска параллелей, обогащая представления эмбеддингами с культурно-историческими признаками. Во всех случаях такая мультимодальная стратегия позволяет преодолевать ограничения одномодальных подходов: лингвистические подсказки компенсируют слабую визуальную информацию, а визуальные и культурно-исторические признаки снижают неоднозначность и повышают уверенность в реконструкции [37, 42, 43].

Современные методы контекстно-зависимого моделирования позволяют успешно преодолевать фундаментальные ограничения, связанные с малоресурсностью и неоднозначностью данных, смещая фокус с автоматического распознавания на создание интеллектуальных ассистивных систем, поддерживающих экспертов в процессах реконструкции и анализа [44, 45].

3.4. Самообучение и обучение без учителя

Самообучение и обучение без учителя представляют собой перспективные методологические подходы для анализа древних иероглифических систем письма через анализ неразмеченных или слабо размеченных данных. Это особенно актуально в эпиграфике и палеографии, где экспертные аннотации являются дорогостоящими и труднодоступными, а сами корпуса текстов характеризуются крайней ограниченностью, фрагментарностью и наличием значительного количества повреждённых или уникальных графем.

В контексте работы с иероглифическими системами данные методы решают ряд специфических задач. Алгоритмы самообучения, такие как контрастивное обучение или маскированное моделирование изображений, позволяют строить устойчивые представления графем, обучаясь на задачах восстановления, предсказания контекста или сравнения различных модификаций одного знака. Обучение без учителя, в свою очередь, применяется для выявления

скрытых закономерностей в распределении символов, автоматической кластеризации визуально схожих элементов и анализа пространственных конфигураций внутри надписей. Совместное использование этих подходов закладывает основу для создания автоматизированных инструментов, способных поддерживать экспертов в процессах дешифровки, каталогизации и структурного анализа вымерших письменностей [46, 47, 48].

Применение методов самообучения на основе анализа неразмеченных массивов данных позволяет формировать устойчивые представления графем в условиях острого дефицита аннотированных образцов. Такие методы способствует извлечению глубоких структурных признаков без привлечения экспертных знаний. Важной особенностью таких систем является их способность извлекать семантически значимые характеристики напрямую из необработанных изображений, что минимизирует зависимость от современных каллиграфических имитаций и повышает точность идентификации редких знаков. Это позволяет эффективно классифицировать визуально схожие элементы, преодолевая трудности, связанные с наличием дефектов [49, 50, 51, 52].

Современные методы автоматизации работы с палеографическими материалами объединяют использование алгоритмов обучения без учителя, вероятностных моделей и автокодировщиков для выявления значимых областей, идентификации и группировки символов, а также определения отношений между их вариантами. Основные этапы таких подходов включают удаление визуального шума, разделение перекрывающихся знаков и формирование упорядоченных коллекций без заранее заданных категорий, что позволяет создавать репрезентативные наборы данных непосредственно на основе оригинальных памятников и снижать долю ручного труда при каталогизации [53, 54, 55].

Разрабатываемые алгоритмы самообучения и обучения без учителя успешно комбинируются с другими подходами и обеспечивают высокую точность выделения символов на разнообразных поверхностях и поддерживает автоматизацию обработки от обнаружения графемы до её системной классификации и транслитерации, что расширяет возможности комплексного анализа эпиграфического наследия [36, 56, 57].

4. Заключение

Современные исследования демонстрируют, что интеграция методов искусственного интеллекта существенно расширяет возможности анализа вымерших иероглифических письменностей. Синтетическое расширение данных, межъязыковое сравнение, контекстно-зависимое моделирование и автоматическое извлечение признаков позволяют преодолевать фундаментальные ограничения малоресурсных и фрагментарных корпусов, улучшать точность распознавания графем и создавать репрезентативные наборы данных без полной зависимости от ручной разметки. В совокупности эти подходы обеспечивают более объективное, масштабируемое и системное исследование текстового наследия, повышая эффективность работы экспертов и расширяя возможности цифровой палеографии

Особое значение приобретает гибридный анализ, который сочетает визуальные, лингвистические и культурно-исторические признаки, позволяя реконструировать повреждённые или редкие графемы и выявлять закономерности в развитии письменностей. Применение этих методов создаёт основу для дальнейшего построения интеллектуальных систем поддержки экспертов, способных не только автоматизировать рутинные процессы, но и обеспечивать глубокое понимание структуры, эволюции и семантики древних иероглифических систем, открывая перспективы комплексного исследования текстового наследия вымерших языков.

Литература

[1] Shormani M.Q. What fifty-one years of linguistics and artificial intelligence research tell us about their correlation: A scientometric analysis // Artificial Intelligence Review. 2025. Vol. 58. Iss. 12. Art. 379. DOI: 10.1007/s10462-025-11332-5

[2] Кушнерук С.П. Прикладная лингвистика: вызовы XXI века // Вестник Волгоградского государственного университета. Серия 2, Языкознание. 2017. Т. 16, № 2. С. 6–17. DOI: 10.15688/jvolsu2.2017.2.1

[3] França R.P., Borges Monteiro A.C., Arthur R., Iano Y. An overview of deep learning in big data, image, and signal processing in the modern digital age // Hybrid Computational Intelligence for Pattern Analysis, Trends in Deep Learning Methodologies / Piuri V., Raj S., Genovese A., Srivastava R. (eds.). Academic Press, 2021. P. 63-87, DOI: 10.1016/B978-0-12-822226-3.00003-9
[4] Tamburini F. On automatic decipherment of lost ancient scripts relying on combinatorial optimisation and coupled simulated annealing // Front. Artif. Intell. 2025. Vol. 8. Art. 1581129. DOI: 10.3389/frai.2025.1581129
[5] Houston S., Stauder A. What is a hieroglyph? // Homme. OpenEdition, 2020. No. 233. P. 9–44. DOI: 10.4000/lhomme.36526
[6] Pae H.K. Chinese, Japanese, and Korean writing systems: All East-Asian but different scripts // Script Effects as the Hidden Drive of the Mind, Cognition, and Culture. Literacy Studies. Vol 21. Springer: Cham, 2020. P. 71-105. DOI: 10.1007/978-3-030-55152-0_5
[7] Taylor I., Taylor M.M. Writing and literacy in Chinese, Korean and Japanese. Amsterdam, Netherlands: Benjamins (John) North America, 1995. 412 p.
[8] Zhang H., Bian Z., Ma J., Xue F. Study on the Evolution and Development of the Chinese Language and Writing System // Transactions on Social Science, Education and Humanities Research. 2024. Vol. 11. P. 804-808. DOI: 10.62051/gha8d115
[9] Лу И. Принципы создания корпусов китайского языка // Известия Российского государственного педагогического университета им. А. И. Герцена. 2016. № 181. С. 22–29.
[10] Jauhiainen H. Gly2Mdc v.2.0: Lessons Learned from Building a Toolfor Hieroglyphic Texts // Digital Humanities in the Nordic and Baltic Countries. 2024. Vol. 6. No. 1. DOI: 10.5617/dhnbpub.11486
[11] Prager C., Grube N., Brodhun M., Diederichs K., Diehr F., Gronemeyer S., Wagner, E. 5 The Digital Exploration of Maya Hieroglyphic Writing and Language // Crossing Experiences in Digital Epigraphy: From Practice to Discipline / A. De Santis, I. Rossi (eds.). Warsaw, Poland: De Gruyter Open Poland, 2018. P. 65-83.. DOI: 10.1515/9783110607208-006
[12] Sánchez R.M. When Hieroglyphs Meet Technology: A Linguistic Journey through Ancient Egypt Using Natural Language Processing // Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA). 2024. P. 156–169. DOI: 10.63317/2mhka44cjoqv
[13] Alsagoafi A.A., Jassim Aljamal A., Alahmad M.A., Waleed Buhaimed J., Abdulallah Alhamdan T., Saad Alfadhli M. A Bibliometric Analysis of AI-Powered Technologies in Language Learning: Trends from 2022 to 2025 // Forum for Linguistic Studies. 2025. Vol. 7. Iss. 12. P. 1362–1379. DOI: 10.30564/fls.v7i12.12311
[14] Ayyoob M.P., Muhamed Ilyas P. Stroke-based data augmentation for enhancing optical character recognition of ancient handwritten scripts // IEEE Access. 2024. Vol. 12. P. 186794–186802. DOI: 10.1109/ACCESS.2024.3505238
[15] Nadăş M., Dioşan L., Tomescu A. Synthetic data generation using large language models: Advances in text and code // IEEE Access. 2025. Vol. 13. P. 134615–134633. DOI: 10.1109/ACCESS.2025.3589503
[16] Guan S., Greene D. Advancing post-OCR correction: A comparative study of synthetic data // arXiv. 2024. DOI: 10.48550/arXiv.2408.02253
[17] Sommerschield T., Assael Y., Pavlopoulos J., Stefanak V., Senior A., Dyer C., Bodel J., Prag J., Androutsopoulos I., de Freitas N. Machine Learning for Ancient Languages: A Survey // Computational Linguistics. 2023. Vol. 49. Iss. 3. P. 703–747. DOI: https://doi.org/10.1162/coli_a_00481
[18] Wang Z., Wang P., Liu K., Wang P., Fu Y., Lu C.-T., Aggarwal C.C., Pei J., Zhou Y. A Comprehensive Survey on Data Augmentation // arXiv. 2025. DOI: 10.48550/arXiv.2405.09591
[19] Li B., Hou Y., Che W. Data augmentation approaches in natural language processing: A survey // arXiv. 2021. DOI: 10.1016/j.aiopen.2022.03.001
[20] Li J., Wang Q.-F-, Huang K., Yang X., Zhang R., Goulermas J.Y. Towards better long-tailed oracle character recognition with adversarial data augmentation // Pattern Recognition. 2023. Vol. 140. Art. 109534. DOI: 10.1016/j.patcog.2023.109534
[21] Wang W., Zhang T., Jin X., Mouchère H., Yu X. Improving Oracle Bone Characters Recognition via A CycleGAN-based Data Augmentation Method // International Conference on Neural Information Processing (ICONIP), Nov 2022, New Delhi, India. 2023. P. 88–100. URL: https://hal.science/hal-03919404v1 (дата обращения: 10.02.2026).
[22] Yue X., Li H., Fujikawa Y., Meng L. Dynamic Dataset Augmentation for Deep Learning-based Oracle Bone Inscriptions Recognition // J. Comput. Cult. Herit. 2022. Vol. 15. Iss. 4. Art. 76. DOI: 10.1145/ 3532868
[23] Rest C., Fisseler D., Weichert F., Somel T., Müller G.G.W. Illumination-based Augmentation for Cuneiform Deep Neural Sign Classification // J. Comput. Cult. Herit. 2022. Vol. 15. Iss. 3. Art. 50. DOI: 10.1145/3495263
[24] Creed L.M. Neural Style Transfer for synthesising a dataset of ancient Egyptian hieroglyphs // arXiv. 2025. DOI: 10.48550/arXiv.2504.02163
[25] Gao S., Hui B., Li W. Image Generation of Egyptian Hieroglyphs // Proceedings of the 2024 16th International Conference on Machine Learning and Computing (ICMLC '24). New York: Association for Computing Machinery, 2024. P. 389–397. DOI: 10.1145/3651671.3651771

[26] Su B., Liu X., Gao W., Yang Y., Chen S. A restoration method using dual generate adversarial networks for Chinese ancient characters // Visual Informatics. 2022. Vol. 6. Iss. 1. P. 26-34. DOI: 10.1016/j.visinf.2022.02.001
[27] Shen Y., Li J., Huang S., Zhou Y., Xie X., Zhao Q. Data Augmentation for Low-resource Word Segmentation and POS Tagging of Ancient Chinese Texts // Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages. Marseille, France: European Language Resources Association, 2022. P. 169–173. URL: https://aclanthology.org/2022.lt4hala-1.26/ (дата обращения: 10.02.2026).
[28] Snæbjarnarson V., Simonsen A., Glavaš G., Vulić I. Transfer to a Low-Resource Language via Close Relatives: The Case Study on Faroese // Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). Tórshavn, Faroe Islands. University of Tartu Library, 2023. P. 728–737. URL: https://aclanthology.org/2023.nodalida-1.74/ (дата обращения: 10.02.2026).
[29] Jain P.A. Transfer learning in low-resource language processing applications // Scientific Journal of Artificial Intelligence and Blockchain Technologies. 2025. Vol. 2. No. 3. P. 81-89. DOI: 10.63345/sjaibt.v2.i3.210
[30] Goldwasser O., Handel Z. Introduction: Graphemic classifiers in complex script systems // Journal of Chinese Writing Systems. 2024. Vol. 8. Iss. 1. P. 2–13. DOI: 10.1177/25138502241234025
[31] Zhou W., Liu J., Li J., Li J., Lin L., Fukumoto F., Dai D. Style-independent radical sequence learning for zero-shot recognition of Small Seal script // Journal of the Franklin Institute. 2023. Vol. 360. Iss. 16. P. 11295–11313. DOI: 10.1016/j.jfranklin.2023.09.005
[32] Chen Z., Yang W., Li X. Stroke-based autoencoders: Self-supervised learners for efficient zero-shot Chinese character recognition // Applied Sciences. 2023. Vol. 13. Iss. 3. Art. 1750. DOI: 10.3390/app13031750.
[33] Cai X., Zhang E. HieroLM: Egyptian hieroglyph recovery with next word prediction language model // Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025). Albuquerque, New Mexico. Stroudsburg, PA, USA: Association for Computational Linguistics, 2025. P. 25–31. URL: https://aclanthology.org/2025.latechclfl-1.4/ (дата обращения: 10.02.2026).
[34] Wang M., Cai Y., Gao L., Feng R., Jiao Q., Ma X., Jia Y. Study on the evolution of Chinese characters based on few-shot learning: From oracle bone inscriptions to regular script // PLoS One. 2022. Vol. 17. Iss. 8. Art. e0272974. DOI: 10.1371/journal.pone.0272974
[35] Jiang R. Liu Y., Zhang B., Chen X., Li D., Han Y. OraclePoints: A hybrid neural representation for oracle character // Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA: ACM, 2023. P. 7901–7911. DOI: 10.1145/3581783.3612534
[36] Fuentes-Ferrer R., Duque-Domingo J., Herrera P.J. Recognition of Egyptian hieroglyphic texts through focused generic segmentation and cross-validation voting // Applied Soft Computing. 2025. Vol. 171. Art. 112793. DOI: 10.1016/j.asoc.2025.112793
[37] Assael Y., Sommerschield T., Cooley A., Shillingford B., Pavlopoulos J., Suresh P., Herms B., Grayston J., Maynard B., Dietrich N., Wulgaert R., Prag J., Mullen A., Mohamed S. Contextualizing ancient texts with generative neural networks // Nature. 2025. Vol. 645. P. 141–147. DOI: 10.1038/s41586-025-09292-5
[38] Hu W., Zhan H., Ma X., Liu C., Yin B., Lu Y., Suen C.Y. VGTS: Visually Guided Text Spotting for novel categories in historical manuscripts // Expert Systems with Applications. 2025. Vol. 261. Art. 125557. DOI: 10.1016/j.eswa.2024.125557
[39] Sobhy A., Helmy M., Khalil M., Elmasry S., Boules Y., Negied N. An AI based automatic translator for ancient hieroglyphic language—from scanned images to English text // IEEE Access. 2023. Vol. 11. P. 38796–38804. DOI: 10.1109/ACCESS.2023.3267981
[40] Yan Y., Zhang R., He H., Lei N., Zhang X., Jiang C. Image restoration technology of Tang dynasty tomb murals using adversarial edge learning // Journal on Computing and Cultural Heritage. 2024. Vol. 17. Iss. 3. Art. 52. DOI: 10.1145/3674984
[41] Zhao F., Ren H., Sun K., Zhu X. GAN-based heterogeneous network for ancient mural restoration // Heritage Science. 2024. Vol. 12. Iss. 1. Art. 418. DOI: 10.1186/s40494-024-01517-6
[42] Wang Z., Li Y., Li H. Chinese inscription restoration based on artificial intelligent models // Heritage Science. 2025. Vol. 13. Iss. 1. Art. 326. DOI: 10.1038/s40494-025-01900-x
[43] Chen Y., Hu C., Feng C., Song C., Yu S., Han X., Liu Z., Sun V. Multi-Modal Multi-Granularity Tokenizer for Chu Bamboo Slips // Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE. Association for Computational Linguistics, 2025. P. 6201–6211. URL: https://aclanthology.org/2025.coling-main.414/ (дата обращения: 10.02.2026).
[44] Bhatt R., Rai A., Chanda S., Krishnan N.C. Pho(SC)-CTC—A hybrid approach towards zero-shot word image recognition // International Journal on Document Analysis and Recognition. 2023. Vol. 26. P. 51–63. DOI: 10.1007/s10032-022-00407-6
[45] Li Q., Zhang C. UCR: A unified character-radical dual-supervision framework for accurate Chinese character recognition // Pattern Recognition. 2025. Vol. 162. Iss. C. Art. 111373. DOI: 10.1016/j.patcog.2025.111373

[46] Zhao Z., Alzubaidi L., Zhang J., Duan Y., Gu Y. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations // Expert Systems with Applications. 2024. Vol. 242. Art. 122807. DOI: 10.1016/j.eswa.2023.122807
[47] Gui J., Chen T., Zhang J., Cao Q., Sun Z., Luo H., Tao D. A survey on self-supervised learning: Algorithms, applications, and future trends // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024. Vol. 46. Iss. 12. P. 9052–9071. DOI: 10.1109/TPAMI.2024.3415112
[48] Ganeriwala P., Mitra D. Few-shot learning for grapheme recognition in ancient scripts // Journal on Computing and Cultural Heritage. 2025. Art. 3773290. DOI: 10.1145/3773290
[49] Wang M., Deng W., Su S. Oracle character recognition using unsupervised discriminative consistency network // Pattern Recognition. 2024. Vol. 148. Art. 110180. DOI: 10.1016/j.patcog.2023.110180
[50] Zhang C. Wang B., Chen K., Zong R., Mo B., Men Y., Almpanidis G., Chen S., Zhang X. Data-driven oracle bone rejoining: A dataset and practical self-supervised learning scheme // Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2022. P. 4482–4492. DOI: 10.1145/3534678.3539050
[51] Zhou S., Wang, X., Qiu, J., Bu W., Wang H. OracleNet: enhancing Oracle Bone Script recognition with Adaptive Deformation and Texture-Structure Decoupling // Heritage science. 2025. Vol. 13. Iss. 1. Art. 273. DOI: 10.1038/s40494-025-01839-z
[52] Wang M., Deng W., Liu C.-L. Unsupervised structure-texture separation network for oracle character recognition // IEEE Transactions on Image Processing. 2022. Vol. 31. P. 3137–3150. DOI: 10.1109/TIP.2022.3165989
[53] Born L., Monroe M.W., Kelley K., Sarkar A. Learning the character inventories of undeciphered scripts using unsupervised deep clustering // Proceedings of the Workshop on Computation and Written Language (CAWL 2023). Toronto, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023. P. 92–104. URL: https://aclanthology.org/2023.cawl-1.11/ (дата обращения: 10.02.2026).
[54] Bi X., Sun Z., Chen Z. A novel unsupervised contrastive learning framework for ancient Yi script character dataset construction // Heritage science. 2025. Vol. 13. Iss. 1. Art. 39. DOI: 10.1038/s40494-025-01600-6
[55] Yue X., Wang Z., Ishibashi R., Kaneko H., Meng L. An unsupervised automatic organization method for Professor Shirakawa's hand-notated documents of oracle bone inscriptions // International Association for Pattern Recognition. 2024. Vol. 27. Iss. 4. P. 583–601. DOI: 10.1007/s10032-024-00463-0
[56] Lion P., Trunz E., Klein R. Unsupervised detection and localization of Egyptian hieroglyphs // Eurographics Workshop on Graphics and Cultural Heritage. The Eurographics Association, 2024. DOI: 10.2312/gch.20241259
[57] Nasser A., Mohamed M., Sherif A., Mahmoud B., Yehia S., Saad A., El-Rahmany M.S., Mohamed E.H. HieroGlyphTranslator: Automatic recognition and translation of Egyptian hieroglyphs to English // arXiv. 2025. DOI: 10.48550/arXiv.2512.03817

Experience and Prospects of Applying Artificial Intelligence to the Study of Hieroglyphic Manuscripts in Extinct Languages

E. V. Yagunova, N. D. Kostygov

Saint Petersburg State University, Saint Petersburg, Russia

Abstract. The paper examines approaches to the processing and analysis of low-resource and fragmentary corpora associated with hieroglyphic writing systems of extinct languages. The main focus is placed on the application of machine learning methods, including generative, comparative, and context-oriented approaches, under conditions of limited and heterogeneous data. It is emphasized that the specific nature of such corpora requires a comprehensive consideration of various factors related to the representation and interpretation of written sources. The importance of developing and applying methods aimed at improving the robustness of analysis and expanding data processing capabilities in conditions of scarcity is highlighted. In this context, general directions in the development of modern technologies designed to work with such materials are discussed, as well as their role in shaping new research practices. Particular attention is given to the potential of integrating different approaches, enabling more flexible and adaptive solutions to analytical tasks. It is concluded that the use of machine learning methods opens prospects for the further development of digital palaeography and the creation of intelligent systems for expert support, contributing to a more systematic and scalable study of the written heritage of extinct languages.

Keywords: computational linguistics, artificial intelligence, extinct languages, hieroglyphic writing systems, low-resource corpora, machine learning

References

[1] Shormani, M.Q. (2025). What fifty-one years of linguistics and artificial intelligence research tell us about their correlation: A scientometric analysis. *Artificial Intelligence Review*. Vol. 58. Iss. 12. Art. 379. DOI: 10.1007/s10462-025-11332-5

[2] Kushneruk, S.P. (2017). Applied Linguistics: Chellenges of the 21th Century. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2, Yazykoznanie [Science Journal of Volgograd State University. Linguistics]*. Vol. 16. No. 2. 6-17. DOI: 10.15688/jvolsu2.2017.2.1 [In Russian]

[3] França, R.P., Borges Monteiro, A.C., Arthur, R., Iano, Y. (2021). An overview of deep learning in big data, image, and signal processing in the modern digital age. In: Hybrid Computational Intelligence for Pattern Analysis, Trends in Deep Learning Methodologies. Piuri V., Raj S., Genovese A., Srivastava R. (eds.). Academic Press. 63-87, DOI: 10.1016/B978-0-12-822226-3.00003-9

[4] Tamburini, F. (2025). On automatic decipherment of lost ancient scripts relying on combinatorial optimisation and coupled simulated annealing. *Front. Artif. Intell*. Vol. 8. Art. 1581129. DOI: 10.3389/frai.2025.1581129

[5] Houston, S., Stauder, A. (2020). What is a hieroglyph? *Homme. OpenEdition*, No. 233. 9–44. DOI: 10.4000/lhomme.36526

[6] Pae, H.K. (2020). Chinese, Japanese, and Korean writing systems: All East-Asian but different scripts. In: Script Effects as the Hidden Drive of the Mind, Cognition, and Culture. *Literacy Studies*. Vol 21. Springer. Cham. 71-105. DOI: 10.1007/978-3-030-55152-0_5

[7] Taylor, I., Taylor, M.M. (1995). Writing and literacy in Chinese, Korean and Japanese. Amsterdam, Netherlands. Benjamins (John) North America. 412 p.

[8] Zhang, H., Bian, Z., Ma, J., Xue, F. (2024). Study on the Evolution and Development of the Chinese Language and Writing System. *Transactions on Social Science, Education and Humanities Research*. Vol. 11. 804-808. DOI: 10.62051/gha8d115

[9] Lu, I. (2016). The principles for building chinese corpora. *Izvestija Rossijskogo gosudarstvennogo pedagogicheskogo universiteta im. A. I. Gercena*. No. 181. 22–29.

[10] Jauhiainen, H. (2024). Gly2Mdc v.2.0: Lessons Learned from Building a Toolfor Hieroglyphic Texts. *Digital Humanities in the Nordic and Baltic Countries*. Vol. 6. No. 1. DOI: 10.5617/dhnbpub.11486

[11] Prager, C., Grube, N., Brodhun, M., Diederichs, K., Diehr, F., Gronemeyer, S., Wagner, E. (2018). 5 The Digital Exploration of Maya Hieroglyphic Writing and Language. In: Crossing Experiences in Digital Epigraphy: From Practice to Discipline. A. De Santis, I. Rossi (eds.). Warsaw, Poland. De Gruyter Open Poland. 65-83. DOI: 10.1515/9783110607208-006

[12] Sánchez, R.M. (2024). When Hieroglyphs Meet Technology: A Linguistic Journey through Ancient Egypt Using Natural Language Processing. In: Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA). 156–169. DOI: 10.63317/2mhka44cjoqv

[13] Alsagoafi, A.A., Jassim Aljamal. A., Alahmad, M.A., Waleed Buhaimed, J., Abdulallah Alhamdan, T., Saad Alfadhli, M. (2025). A Bibliometric Analysis of AI-Powered Technologies in Language Learning: Trends from 2022 to 2025. *Forum for Linguistic Studies*. Vol. 7. Iss. 12. 1362–1379. DOI: 10.30564/fls.v7i12.12311

[14] Ayyoob, M.P., Muhamed Ilyas, P. (2024). Stroke-based data augmentation for enhancing optical character recognition of ancient handwritten scripts. *IEEE Access*. Vol. 12. 186794–186802. DOI: 10.1109/ACCESS.2024.3505238

[15] Nadăş, M., Dioşan, L., Tomescu, A. (2025). Synthetic data generation using large language models: Advances in text and code. *IEEE Access*. Vol. 13. 134615–134633. DOI: 10.1109/ACCESS.2025.3589503

[16] Guan, S., Greene, D. (2024). Advancing post-OCR correction: A comparative study of synthetic data. *arXiv*. DOI: 10.48550/arXiv.2408.02253

[17] Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., de Freitas, N. (2023). Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*. Vol. 49. Iss. 3. 703–747. DOI: https://doi.org/10.1162/coli_a_00481

[18] Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C.-T., Aggarwal, C.C., Pei, J., Zhou, Y. (2025). A Comprehensive Survey on Data Augmentation. *arXiv*. DOI: 10.48550/arXiv.2405.09591

[19] Li, B., Hou, Y., Che, W. (2021). Data augmentation approaches in natural language processing: A survey. *arXiv*. DOI: 10.1016/j.aiopen.2022.03.001

[20] Li, J., Wang, Q.-F-, Huang, K., Yang, X., Zhang, R., Goulermas, J.Y. (2023). Towards better long-tailed oracle character recognition with adversarial data augmentation. *Pattern Recognition*. Vol. 140. Art. 109534. DOI: 10.1016/j.patcog.2023.109534

[21] Wang, W., Zhang, T., Jin, X., Mouchère, H., Yu, X. (2023). Improving Oracle Bone Characters Recognition via A CycleGAN-based Data Augmentation Method. In: International Conference on Neural Information Processing (ICONIP), Nov 2022, New Delhi, India. 88–100. Available at: https://hal.science/hal-03919404v1. (accessed date: 10/02/2026).

[22] Yue, X., Li, H., Fujikawa, Y., Meng, L. (2022). Dynamic Dataset Augmentation for Deep Learning-based Oracle Bone Inscriptions Recognition. *J. Comput. Cult*. Herit. Vol. 15. Iss. 4. Art. 76. DOI: 10.1145/ 3532868

[23] Rest, C., Fisseler, D., Weichert, F., Somel, T., Müller, G.G.W. (2022). Illumination-based Augmentation for Cuneiform Deep Neural Sign Classification. *J. Comput. Cult.* Herit. Vol. 15. Iss. 3. Art. 50. DOI: 10.1145/3495263
[24] Creed, L.M. (2025). Neural Style Transfer for synthesising a dataset of ancient Egyptian hieroglyphs. *arXiv*. DOI: 10.48550/arXiv.2504.02163
[25] Gao, S., Hui, B., Li, W. (2024). Image Generation of Egyptian Hieroglyphs. In: Proceedings of the 2024 16th International Conference on Machine Learning and Computing (ICMLC '24). New York. Association for Computing Machinery. 389–397. DOI: 10.1145/3651671.3651771
[26] Su, B., Liu, X., Gao, W., Yang, Y., Chen, S. (2022). A restoration method using dual generate adversarial networks for Chinese ancient characters. *Visual Informatics*. Vol. 6. Iss. 1. 26-34. DOI: 10.1016/j.visinf.2022.02.001
[27] Shen, Y., Li, J., Huang, S., Zhou, Y., Xie, X., Zhao, Q. (2022). Data Augmentation for Low-resource Word Segmentation and POS Tagging of Ancient Chinese Texts. In: Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages. Marseille, France. European Language Resources Association. 169–173. Available at: https://aclanthology.org/2022.lt4hala-1.26/ (accessed date: 10/02/2026).
[28] Snæbjarnarson, V., Simonsen, A., Glavaš, G., Vulić, I. (2023). Transfer to a Low-Resource Language via Close Relatives: The Case Study on Faroese. In: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). Tórshavn, Faroe Islands. University of Tartu Library. 728–737. Available at: https://aclanthology.org/2023.nodalida-1.74/ (accessed date: 10/02/2026).
[29] Jain, P.A. (2025). Transfer learning in low-resource language processing applications. *Scientific Journal of Artificial Intelligence and Blockchain Technologies*. Vol. 2. No. 3. 81-89. DOI: 10.63345/sjaibt.v2.i3.210
[30] Goldwasser, O., Handel, Z. (2024). Introduction: Graphemic classifiers in complex script systems. *Journal of Chinese Writing Systems*. Vol. 8. Iss. 1. 2–13. DOI: 10.1177/25138502241234025
[31] Zhou, W., Liu, J., Li, J., Li, J., Lin, L., Fukumoto, F., Dai, D. (2023). Style-independent radical sequence learning for zero-shot recognition of Small Seal script. *Journal of the Franklin Institute*. Vol. 360. Iss. 16. 11295–11313. DOI: 10.1016/j.jfranklin.2023.09.005
[32] Chen, Z., Yang, W., Li, X. (2023). Stroke-based autoencoders: Self-supervised learners for efficient zero-shot Chinese character recognition. *Applied Sciences*. Vol. 13. Iss. 3. Art. 1750. DOI: 10.3390/app13031750.
[33] Cai, X., Zhang, E. (2025). HieroLM: Egyptian hieroglyph recovery with next word prediction language model. In: Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025). Albuquerque, New Mexico. Stroudsburg, PA, USA. Association for Computational Linguistics. 25–31. Available at: https://aclanthology.org/2025.latechclfl-1.4/ (accessed date: 10/02/2026).
[34] Wang, M., Cai, Y., Gao, L., Feng, R., Jiao, Q., Ma, X., Jia, Y. (2022). Study on the evolution of Chinese characters based on few-shot learning: From oracle bone inscriptions to regular script. *PLoS One*. Vol. 17. Iss. 8. Art. e0272974. DOI: 10.1371/journal.pone.0272974
[35] Jiang, R. Liu, Y., Zhang, B., Chen, X., Li, D., Han, Y. (2023). OraclePoints: A hybrid neural representation for oracle character. In: Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA. ACM. 7901–7911. DOI: 10.1145/3581783.3612534
[36] Fuentes-Ferrer, R., Duque-Domingo, J., Herrera, P.J. (2025). Recognition of Egyptian hieroglyphic texts through focused generic segmentation and cross-validation voting. *Applied Soft Computing*. Vol. 171. Art. 112793. DOI: 10.1016/j.asoc.2025.112793
[37] Assael, Y., Sommerschield, T., Cooley, A., Shillingford, B., Pavlopoulos, J., Suresh, P., Herms, B., Grayston, J., Maynard, B., Dietrich, N., Wulgaert, R., Prag, J., Mullen, A., Mohamed, S. (2025). Contextualizing ancient texts with generative neural networks. *Nature*. Vol. 645. 141–147. DOI: 10.1038/s41586-025-09292-5
[38] Hu, W., Zhan, H., Ma, X., Liu, C., Yin, B., Lu, Y., Suen, C.Y. (2025). VGTS: Visually Guided Text Spotting for novel categories in historical manuscripts. *Expert Systems with Applications*. Vol. 261. Art. 125557. DOI: 10.1016/j.eswa.2024.125557
[39] Sobhy, A., Helmy, M., Khalil, M., Elmasry, S., Boules, Y., Negied, N. (2023). An AI based automatic translator for ancient hieroglyphic language—from scanned images to English text. *IEEE Access*. Vol. 11. P. 38796–38804. DOI: 10.1109/ACCESS.2023.3267981
[40] Yan, Y., Zhang, R., He, H., Lei,, N., Zhang, X., Jiang, C. (2024). Image restoration technology of Tang dynasty tomb murals using adversarial edge learning. *Journal on Computing and Cultural Heritage*. Vol. 17. Iss. 3. Art. 52. DOI: 10.1145/3674984
[41] Zhao, F., Ren, H., Sun, K., Zhu X. (2024). GAN-based heterogeneous network for ancient mural restoration. *Heritage Science*. Vol. 12. Iss. 1. Art. 418. DOI: 10.1186/s40494-024-01517-6
[42] Wang, Z., Li, Y., Li, H. (2025). Chinese inscription restoration based on artificial intelligent models. *Heritage Science*. Vol. 13. Iss. 1. Art. 326. DOI: 10.1038/s40494-025-01900-x
[43] Chen, Y., Hu, C., Feng, C., Song, C., Yu, S., Han, X., Liu, Z., Sun, V. (2025). Multi-Modal Multi-Granularity Tokenizer for Chu Bamboo Slips. In: Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE. Association for Computational Linguistics. 6201–6211. Available at: https://aclanthology.org/2025.coling-main.414/ (accessed date: 10/02/2026).

[44] Bhatt, R., Rai, A., Chanda, S., Krishnan, N.C. (2023). Pho(SC)-CTC—A hybrid approach towards zero-shot word image recognition. *International Journal on Document Analysis and Recognition*. Vol. 26. 51–63. DOI: 10.1007/s10032-022-00407-6

[45] Li, Q., Zhang, C. (2025). UCR: A unified character-radical dual-supervision framework for accurate Chinese character recognition. *Pattern Recognition*. Vol. 162. Iss. C. Art. 111373. DOI: 10.1016/j.patcog.2025.111373

[46] Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Gu, Y. (2024). A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*. Vol. 242. Art. 122807. DOI: 10.1016/j.eswa.2023.122807

[47] Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 46. Iss. 12. 9052–9071. DOI: 10.1109/TPAMI.2024.3415112

[48] Ganeriwala, P., Mitra, D. (2025). Few-shot learning for grapheme recognition in ancient scripts. *Journal on Computing and Cultural Heritage*. Art. 3773290. DOI: 10.1145/3773290

[49] Wang, M., Deng, W., Su, S. (2024). Oracle character recognition using unsupervised discriminative consistency network. *Pattern Recognition*. Vol. 148. Art. 110180. DOI: 10.1016/j.patcog.2023.110180

[50] Zhang, C. Wang, B., Chen, K., Zong, R., Mo, B., Men, Y., Almpanidis, G., Chen, S., Zhang, X. (2002). Data-driven oracle bone rejoining: A dataset and practical self-supervised learning scheme. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA. ACM. 4482–4492. DOI: 10.1145/3534678.3539050

[51] Zhou, S., Wang, X., Qiu, J., Bu, W., Wang, H. (2025). OracleNet: enhancing Oracle Bone Script recognition with Adaptive Deformation and Texture-Structure Decoupling. *Heritage science*. Vol. 13. Iss. 1. Art. 273. DOI: 10.1038/s40494-025-01839-z

[52] Wang, M., Deng, W., Liu, C.-L. (2022). Unsupervised structure-texture separation network for oracle character recognition. *IEEE Transactions on Image Processing*. Vol. 31. 3137–3150. DOI: 10.1109/TIP.2022.3165989

[53] Born, L., Monroe, M.W., Kelley, K., Sarkar, A. (2023). Learning the character inventories of undeciphered scripts using unsupervised deep clustering. In: Proceedings of the Workshop on Computation and Written Language (CAWL 2023). Toronto, Canada. Stroudsburg, PA, USA. Association for Computational Linguistics. 92–104. Available at: https://aclanthology.org/2023.cawl-1.11/ (accessed date: 10/02/2026).

[54] Bi, X., Sun, Z., Chen, Z. (2025). A novel unsupervised contrastive learning framework for ancient Yi script character dataset construction. *Heritage science*. Vol. 13. Iss. 1. Art. 39. DOI: 10.1038/s40494-025-01600-6

[55] Yue, X., Wang, Z., Ishibashi, R., Kaneko, H., Meng, L. (2024). An unsupervised automatic organization method for Professor Shirakawa's hand-notated documents of oracle bone inscriptions. *International Association for Pattern Recognition*. Vol. 27. Iss. 4. 583–601. DOI: 10.1007/s10032-024-00463-0

[56] Lion, P., Trunz, E., Klein, R. (2024). Unsupervised detection and localization of Egyptian hieroglyphs. *Eurographics Workshop on Graphics and Cultural Heritage*. The Eurographics Association. DOI: 10.2312/gch.20241259

[57] Nasser, A., Mohamed, M., Sherif, A., Mahmoud, B., Yehia, S., Saad, A., El-Rahmany, M.S., Mohamed, E.H. (2025). HieroGlyphTranslator: Automatic recognition and translation of Egyptian hieroglyphs to English. *arXiv*. DOI: 10.48550/arXiv.2512.03817.