Чесноков Сергей Александрович
магистрант
ФГБОУ ВО «Нижневартовский государственный университет»
г. Нижневартовск, ХМАО – Югра

ВОЗМОЖНОСТИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В ПЕРЕВОДЕ ИСТОРИОГРАФИЧЕСКИХ ТЕКСТОВ

Аннотация: *в статье исследуется специфика применения технологий искусственного интеллекта для перевода текстов историко-научного дискурса. На материале публикаций, взятых из крупных византиноведческих периодических изданий Byzantinische Zeitschrift и Dumbarton Oaks Papers, проводится сопоставительный анализ качества перевода системы нейросетевого машинного перевода (NMT) и большой языковой модели (LLM).*

Ключевые слова: *историко-научный дискурс, машинный перевод, византиноведение, перевод реалий, терминология.*

Введение.

Аспекты профессиональной подготовки переводчиков с использованием проектной технологии находятся сегодня в центре внимания исследователей; помимо лингвистической составляющей анализируются вопросы организации и управления конкретными переводческими проектами [4]. Подготовка современного переводчика включает в себя практическое освоение технология искусственного интеллекта как инструмента качественной автоматизированной обработки и перевода специальных текстов.

Интеграция технологий искусственного интеллекта в методологию исследований актуализирует проблему качественной автоматизированной обработки специализированных текстов. Историко-научный дискурс представляет собой специфический объект для машинного перевода, характеризующийся высокой плотностью узкопрофильной терминологии, наличием архаичных синтаксических конструкций и широким пластом безэквивалентной лексики и реалий. В

этих условиях перед исследователем (переводчиком) встает вопрос выбора оптимального технологического инструментария, способного обеспечить предельную фактологическую точность при передаче имен собственных, топонимов и кодикологических данных в процессе перевода иноязычных текстов. Целью данной статьи является сравнительный анализ эффективности двух доминирующих подходов в компьютерной лингвистике – систем нейросетевого машинного перевода (NMT) и больших языковых моделей (LLM).

Эмпирической базой для верификации данной гипотезы послужили материалы, отобранные в рамках исследовательского проекта, целью которого является сбор, перевод и первичный анализ зарубежной научной литературы на английском, французском и немецком языках, характеризующей научную деятельность известного византиниста Х. М. Лопарева и его влияние на мировое византиноведение. В качестве модельных объектов для эксперимента были использованы фрагменты англоязычных источниковедческих работ в области византинистики.

Эволюция парадигм автоматизированного перевода.

В современной компьютерной лингвистике доминирующей парадигмой машинного перевода долгое время являлся нейросетевой подход (Neural Machine Translation, NMT). NMT-системы моделируют условную вероятность последовательности слов языка перевода на основе последовательности слов исходного языка, используя архитектуру «кодировщик-декодировщик» (encoder-decoder) [9]. Однако применительно к историко-научному дискурсу NMT-системы демонстрируют ряд фундаментальных ограничений. Возникают трудности с разрешением анафорических связей и полисемии, если подсказка для правильного перевода находится за пределами текущего предложения. NMT-системы обучаются на параллельных корпусах. Поскольку параллельных корпусов по византинистике или медиевистике крайне мало, система склонна заменять редкие термины на более частотные общеупотребительные эквиваленты, что ведет к терминологической эрозии текста. Следует отметить также, что ошибки оптического распознавания, характерные для оцифрованных изданий XIX века не воспринимаются

и не анализируются NMT-системой в широком контексте, что может становится причиной некачественного перевода [5].

Появление трансформерных моделей (transformer architecture) и последующее развитие генеративных предобученных трансформеров (generative pre-trained transformer) изменили подход к задаче перевода. В отличие от NMT, где перевод является единственной целевой функцией, для LLM перевод – это лишь одна из эмерджентных способностей, проявляющаяся в результате обучения на сверхмасштабных массивах данных. LLM обучаются не только на парах «язык А – язык Б», но и на огромном массиве моноязычных текстов [10]. Это позволяет модели формировать внутреннее представление о сущностях реального мира. LLM способны выполнять задачи zero-shot (без примеров) и few-shot (с несколькими примерами), опираясь на свои «знания» о мире. В контексте перевода это означает, что модель не просто ищет лингвистическое соответствие слову, обозначающему, например, византийского императора, а обращается к информации, ассоциированной с этой исторической личностью, что позволяет корректно восстанавливать контекст [7]. При переводе текстов историко-научного дискурса требуется соблюдение академического узуса. Большие языковые модели справляются с сохранением единого стиля в рамках целого текста, что важно при переводе историографических материалов.

Для эмпирической верификации теоретических положений о преимуществе больших языковых моделей над системами нейросетевого машинного перевода был проведен детальный анализ перевода фрагментов из базы данных Dumbarton Oaks Hagiography Database и Byzantinische Zeitschrift. В качестве модели LLM использовалась система DeepSeek, в качестве системы NMT – сервис DeepL. В исходные текстах затрагиваются источниковедческие вопросы, такие как проблемы датировки, авторства и исторической интерпретации агиографических источников в контексте историографических исследований византинистики XIX-XX веков и содержится характерный набор трудностей историко-научного дискурса: исторические антропонимы, топонимы, жанровые термины и кодикологические шифры.

Был проведен сравнительный анализ двух переводов одного фрагмента текста историко-научного дискурса. Наиболее показательным является различие стратегий на ономастическом уровне. При передаче имен собственных в историческом тексте переводчик неизбежно сталкивается с дилеммой выбора между фонетической транскрипцией и использованием традиционного эквивалента. Антропонимы «Elias», «Daniel», «Sabas» переводятся системой NMT (DeepL) посредством транслитерации, как стандартные западноевропейские имена: «Элиас», «Даниэль», «Сабас». В рамках византиноведения, связанного непосредственно с восточно-христианской традицией, такой перевод приводит к культурной интерференции. В переводе LLM антропоним «Elias the Younger», связанный с исторической личностью святым Илией Эннским корректно распознается. Используются канонические формы имен, таких как «Илия», «Даниил», «Савва». Аналогичная ситуация происходит при переводе фамилии академика В. Г. Васильевского. NMT транслитерирует графемную часть как «Василевский», LLM приводит корректную форму фамилии ученого.

Существенные расхождения наблюдаются на лексико-семантическом уровне, где основной проблемой становится полисемия лексических единиц. Архитектура NMT при переводе использует наиболее частотные значения. Так, термин «vita» (лат. «жизнь», в научном дискурсе – «житие») переводится общелитературной единицей «биография», топос «in a chest» (в значении «реликварий» или «ковчег») – бытовой конструкцией «в сундуке», а сакральную реалию «Virgin's robe» («Риза Богородицы») – гиперонимом «одежда». В переводе LLM, прослеживается работа с широким контекстом. Наличие единиц «vita», «Theotokos», «homily», «church» задает тематический вектор, который определяет выбор специализированных лексических единиц «житие», «ковчег», «Богородица», «гомилия», обеспечивая тем самым терминологическую конгруэнтность перевода [3].

Рассмотрим ряд примеров в переводе с английского на русский язык.

Пример 1. *Оригинал*:

«*Elias* the Younger of Enna

His anonymous *vita* has survived only in late manuscripts, the oldest of which (Messina, *University, cod. 29*) is of 1307/8. The author was a monk of the Salinas monastery, who claims no personal acquaintance with his hero. It is usually assumed that he wrote soon after *Elias*' death, in the 930s or 940s; da Costa-Louillet posits an even earlier date of composition – 905/6. The author is generally thought to have obtained his information about *Elias* from his disciple *Daniel*, who accompanied his master on his travels. The suggestion of Loparev that the hagiographer was another disciple of *Elias*, *Sabas* by name, *cannot be verified*, nor is there any proof of da Costa-Louillet's identification of the author with *Daniel* himself» [8, с. 30].

Перевод NMT (*DeepL*):

«*Элиас* Младший из Энны

Его анонимная *биография* сохранилась только в поздних рукописях, самая старая из которых (Мессина, *Университет, cod. 29*) датируется 1307/8 годом. Автор был монахом монастыря Салинас, который утверждает, что не был лично знаком со своим героем. Обычно предполагается, что он написал ее вскоре после смерти *Элиаса*, в 930-х или 940-х годах; да Коста-Луйе предлагает еще более раннюю дату создания – 905/6 год. Обычно считается, что автор получил информацию об *Элиасе* от его ученика *Даниэля*, который сопровождал своего учителя в путешествиях. Предположение Лопарева о том, что автором жития был другой ученик *Элиаса* по имени *Сабас*, *не может быть подтверждено*, как и нет никаких доказательств идентификации автора с самим *Даниэлем*, предложенной да Коста-Луйе».

Перевод LLM (*DeepSeek*):

«*Илия* Младший из Энны

Его анонимное *житие* сохранилось лишь в поздних рукописях, древнейшая из которых (Мессина, *Университетская библиотека, кодекс 29*) датируется 1307–1308 годом. Автор жития был монахом монастыря Салинас и утверждает, что не был лично знаком со своим героем. Традиционно считается, что житие было написано вскоре после смерти *Илии* – в 930-х или 940-х годах; однако да Коста-Луйе выдвигает более раннюю датировку – 905–906 год. Предполагается,

что автор получил сведения об *Илии* от его ученика *Даниила*, сопровождавшего святого в путешествиях. Гипотеза Лопарева о том, что агиографом мог быть другой ученик Илии по имени *Савва*, *не находит подтверждения*, как и предположение да Коста-Луйе о том, что автором был сам *Даниил*».

Пример 2. *Оригинал*:

«But before we can proceed to note one or two features of this extract from the Constantinopolitan Typikon we must determine our position in regard to the account of the rediscovery of tpe *Virgin's robe in a chest* in the *church* of the *Theotokos* in *Blachernae,* of which the most convenient text is that of Kh. M. Loparev in *Yiz. Yrem*. *II* (1895) pp. 592 sqq. Loparev refers this *homily* to the attack of the Russians on Constantinople in 842 A. D.; but in the view of the present writer *Vasilevsky* has proved that this contention is untenable, and that the work refers to the *Avar surprise*» [6, c. 123].

Перевод системы NMT (*DeepL*):

«Но прежде чем мы перейдем к рассмотрению одной или двух особенностей этого отрывка из Константинопольского Типикона, мы должны определить нашу позицию в отношении рассказа о повторном обнаружении *одежды Богородицы в сундуке* в *церкви Богородицы* в *Влачернах*, наиболее удобный текст которого приводит Х. М. Лопарев в *Yiz. Yrem*. II (1895) стр. 592 и далее. Лопарев относит эту *проповедь* к нападению русских на Константинополь в 842 году н. э.; но, по мнению автора настоящей статьи, *Василевский* доказал, что это утверждение несостоятельно и что произведение относится к неожиданному *нападению аваров*».

Перевод модели LLM (*DeepSeek*):

«Однако прежде чем перейти к анализу особенностей данного отрывка из Константинопольского Типикона, необходимо определиться с интерпретацией рассказа об обретении *Ризы Богородицы в ковчеге* во *Влахернском храме*, наиболее доступный текст которого представлен в публикации Х.М. Лопарева (*Византийский временник II*, 1895, с. 592 и след.). Лопарев относит соответствующую

гомилию к нападению русов на Константинополь в 842 г., однако, по мнению автора настоящей работы, *Васильевский* убедительно доказал несостоятельность этой точки зрения, показав, что текст описывает события *аварского нашествия*».

Перевод топонима «Blachernae» может послужить примером иллюстрирующим качество передачи имен собственных. NMT в данном случае предлагается вариант «в Влачернах», который является ошибочным. Подобный результат свидетельствует о фонетической интерференции, при которой латинизированный греческий диграф интерпретируется системой неверно, вероятно, под влиянием правил чтения современных западноевропейских языков.

Особого внимания заслуживает работа систем с ошибками распознавания текста, возникающими при обработке текстовых документов. Перевод искаженной библиографической ссылки «Yiz. Yrem.» является демонстрацией возможностей NMT в исполнении данной функции. LLM, напротив, показывает способность к текстологической реконструкции. Вероятно, имеет место анализ семантического окружения, в результате чего предпочтение отдается варианту перевода «Византийский временник». Подобная экспликация происходит при переводе шифра рукописи, где LLM добавляется единица «библиотека» к слову «университет» и расшифровывает сокращение «cod.» как «кодекс».

На синтаксическом уровне различия проявляются в степени адаптации текста к нормам русского академического стиля. Системы NMT склонны к синтаксическому калькированию. Пассивная конструкция английского языка «cannot be verified» передается аналогичным русским пассивом «не может быть подтверждено», а термин «Avar surprise» – пословным переводом «неожиданное нападение аваров». Данный перевод грамматически верен, но содержит стилистические отклонения. LLM генерирует текст с использованием научных клише и устоявшихся терминологических единиц «не находит подтверждения» и «аварское нашествие».

Проведенный анализ позволяет заключить, существует значительное различие в качестве. Если системы нейросетевого машинного перевода склонны к бук-

вализму, что приводит к утрате культурного контекста, то большие языковые модели демонстрируют способность к интерпретации текста и информации в целом, выступая в качестве инструмента интеллектуальной поддержки и редактуры текста. Это делает их более предпочтительными при работе с историко-научным дискурсом. Но также необходимо отметить, что технологии искусственного интеллекта на текущем этапе развития не лишены технических ограничений, в связи с чем полная автоматизация переводческого процесса представляется преждевременной. Функциональная роль переводчика в данном случае состоит в валидации переведенного текста.

Список литературы

1. Влахов С.И. Непереводимое в переводе / С.И. Влахов, С.П. Флорин. – М.: Р. Валент, 2009. – 245 с.

2. Диденко А.Н. Перспективы использования систем искусственного интеллекта в рамках лингвистических научных исследований / А.Н. Диденко, Е.В. Рогова, С.О. Щерба // Вестник Пятигорского государственного университета. – 2023. – №4. – С. 136–140. – DOI 10.53531/25420747_2023_4_136. EDN JQDMPQ

3. Козлова О.В. Особенности передачи культурного контекста при использовании искусственного интеллекта в переводе / О.В. Козлова, М.Р. Евстифеева // Человек. Социум. Общество. – 2024. – №11. – С. 44–50. EDN EMPRAU

4. Новикова Л.В. Рынок переводческих услуг: проект как средство формирования soft skills / Л.В. Новикова // Традиции и инновации в образовательном пространстве России: материалы VII Всероссийской научно-практической конференции (Нижневартовск, 21 апреля 2018 года). – Нижневартовск: Нижневартовский государственный университет, 2018. – С. 92–94. – EDN UXDLKM.

5. Bahdanau D. Neural Machine Translation by Jointly Learning to Align and Translate / D. Bahdanau, K. Cho, Y. Bengio // Proceedings of the 3rd International Conference on Learning Representations (ICLR). – San Diego, 2015. – 15 p.

6. Baynes N. H. The Date of the Avar Surprise: A chronological study // Byzantinische Zeitschrift. 1912. Vol. 21. Pp. 110–128.

7. Brown T. B. Language Models are Few-Shot Learners / T. B. Brown [et al.] // Advances in Neural Information Processing Systems (NeurIPS). 2020. Vol. 33. Pp. 1877–1901.

8. Dumbarton Oaks Hagiography Database / directed by A. Kazhdan; Dumbarton Oaks, Trustees for Harvard University. Washington, D.C.: Dumbarton Oaks, 1998. P. 33.

9. Koehn P. Neural Machine Translation / P. Koehn. Cambridge: Cambridge University Press, 2020. 406 p.

10. Wei J. Emergent Abilities of Large Language Models / J. Wei [et al.] // Transactions on Machine Learning Research. 2022. 30 p.