DOI: 10.33184/dokbsu-2025.4.7

Лингвистические исследования в эпоху развития искусственного интеллекта: русско-китайские параллельные корпусы

А. В. Павлова, Д. В. Аскарова

Оренбургский государственный университет
Россия, 460018 г. Оренбург, пр. Победы, 18.

Email: pavlova_a@bk.ru

Статья посвящена исследованию роли искусственного интеллекта (ИИ) в современных лингвистических исследованиях. Особое внимание уделено потенциалу нейросетевых методов в анализе и выявлении скрытых языковых закономерностей в параллельных русско-китайских корпусах. Рассмотрены современные модели, технологии и ограничения их применения. Предложены перспективы развития методологии в контексте цифровизации и интерпретируемости.

Ключевые слова: лингвистика, искусственный интеллект, русский язык, китайский язык, корпус языка.

Сравнительная лингвистика как наука претерпевает значительные трансформации на фоне бурного развития технологий обработки естественного языка (NLP) и искусственного интеллекта. Там, где ранее основным методом анализа выступал ручной морфолого-синтаксический разбор и контрастивное описание, сегодня на передний план выходят автоматические инструменты, нейросетевые модели и большие корпуса, поддерживающие многоязыковую аналитику.

Русский и китайский языки – представители двух различных структурных типов: флективного и изолирующего соответственно. Их сопоставление является не только лингвистически интересным, но и методологически сложным: расхождения касаются морфологии, синтаксиса, семантики и прагматики. В эпоху ИИ появляются уникальные возможности преодолеть традиционные ограничения сравнительной лингвистики: например, выявлять закономерности, которые недоступны при интуитивном анализе, либо выходят за пределы грамматических описаний.

Современные параллельные корпуса, такие как RuZhCorp, открывают новые перспективы для масштабного изучения соответствий между языками. В сочетании с мощными языковыми моделями – BERT, LaBSE, XLM-R – становится возможным выстраивание межъязыковых карт соответствий и даже моделирование семантических перехо-

дов между конструкциями. Это особенно ценно для языковых пар с низкой степенью типологического родства и высокой культурной разницей.

Цель настоящей статьи – описать, как современные нейросетевые методы и параллельные корпуса способствуют выявлению скрытых закономерностей между русским и китайским языками. В рамках исследования рассматриваются как уже реализованные подходы (выравнивание слов, переводные эквиваленты, эмбеддинги), так и перспективные направления: интерпретируемое ИИ, трансляционные универсалии, анализ сопоставимых корпусов. Статья основана как на обзоре существующих работ, так и на предложении новых методологических решений.

В последние десятилетия параллельные корпуса стали незаменимыми ресурсами в сравнительной лингвистике, особенно в контексте работы с языковыми парами, представляющими разные типологические семьи. Русско-китайский параллельный корпус RuZhCorp (создан в 2016 г.) стал первой масштабной попыткой систематизировать переводные соответствия между двумя крайне разными языками – русским и китайским. Корпус содержит более 4,6 млн слов и представляет тексты различных жанров (художественные, публицистические, деловые и др.) с выравниванием на уровне предложений и морфологической разметкой.

Так, согласно результатам современных исследований, использование предобученных многоязычных моделей, таких как BERT и LaBSE, может значительно повысить точность автоматического выравнивания слов [1]. При дообучении LaBSE на вручную размеченном «золотом» наборе предложений в рамках RuZhCorp удалось достичь минимального значения AER (Alignment Error Rate) – 18,9%, что является значительным улучшением по сравнению со статистическими методами (например, GIZA++).

В других исследованиях показано, что LaBSE превосходит альтернативные модели (например, LASER, mUSE) по качеству извлечения межъязыковых соответствий. Кроме того, использование attention-механизмов в трансформерах позволяет моделям выявлять как прямые переводные эквиваленты, так и функционально-семантические соответствия, не всегда очевидные даже для человека [2–4].

Интерес вызывает также подход к извлечению эквивалентов из сопоставимых, а не строго параллельных текстов [5], где предлагается метод динамического выравнивания с использованием контекстуальных эмбеддингов, позволяющий извлекать смысловые аналоги даже при отсутствии формального соответствия между сегментами текста.

Дополнительно, ряд исследований посвящен выявлению так называемых трансляционных универсалий – закономерностей, характерных для текста в переводе [6–7]. В русско-китайском контексте это выражается, например, в предпочтении определен-

ных синтаксических конструкций (например, аналитических форм в китайском против синтетических в русском).

В совокупности эти работы демонстрируют, что ИИ-инструменты не только автоматизируют процесс анализа, но и выходят за пределы возможного при традиционном подходе, выявляя глубинные закономерности, ранее недоступные лингвисту.

Современные методы анализа параллельных корпусов опираются на архитектуры глубокого обучения, в частности трансформеры, а также на модели представления текста в виде контекстуальных эмбеддингов. Далее рассмотрим ключевые подходы и технологии, применимые к русско-китайскому сопоставлению.

Наиболее распространенными в межъязыковом анализе являются модели на базе трансформеров:

BERT (Bidirectional Encoder Representations from Transformers) – модель, обученная на задаче маскированного моделирования языка, широко используется для извлечения контекстно-зависимых представлений слов.

mBERT – многоязычная версия BERT, способная работать с более чем 100 языками, включая русский и китайский.

LaBSE (Language-agnostic BERT Sentence Embedding) – модель, обученная на параллельных данных для извлечения семантически сопоставимых предложений на разных языках.

XLM-R (Cross-lingual RoBERTa) – расширенная мультилингвальная модель, демонстрирующая высокую эффективность в задачах выравнивания и сопоставления.

Эти модели используются как для анализа на уровне слов, так и для более высокоуровневого представления предложений и фраз в едином эмбеддинговом пространстве.

Одной из ключевых задач при анализе параллельных корпусов является определение соответствий между лексемами. Статистические методы (например, GIZA++) постепенно уступают место нейросетевым: контекстуализированное выравнивание осуществляется через сравнение векторов слов в эмбеддинговом пространстве. Наилучшие результаты достигаются при использовании LaBSE с fine-tuning на целевой языковой паре. Supervised alignment включает ручную разметку обучающего набора, на основе которого модель обучается различать истинные и ложные пары соответствий. Attention-based alignment предполагает использование attention-весов из трансформеров, позволяющих идентифицировать пары слов с высоким взаимным вниманием в процессе перевода.

Помимо лексических соответствий, современные методы позволяют выявлять глубинные семантические связи между конструкциями. Это особенно актуально в парах с асимметрией грамматических и прагматических систем (например, при передаче вежливости, модальности, аспекта действия). Косинусная близость между эмбеддингами предложений позволяет измерить степень семантического соответствия между сегментами.

Cross-lingual retrieval – техника, при которой по фразе на одном языке ищется наиболее релевантная единица на другом языке в эмбеддинговом пространстве.

Explainable AI (XAI) – подход, нацеленный на интерпретируемость модели: выделяются признаки, наиболее сильно влияющие на сопоставление, например, с использованием методов LIME или SHAP.

Для достижения большей точности предлагаются гибридные схемы:

Комбинирование ручной аннотации и автоматических моделей;

Объединение нейросетевого выравнивания с морфосинтаксической разметкой;

Использование сопоставимых корпусов (не строго параллельных) с семантическим выравниванием.

ИИ и нейросетевые методы позволяют исследовать такие межъязыковые связи, которые ранее оставались вне поля зрения традиционной лингвистики. Особенно это актуально в случае таких языков с высокой типологической дивергенцией, как русский и китайский.

Семантическое моделирование с помощью трансформеров выявило устойчивые случаи асимметрии в лексических соответствиях. Например, в китайском языке отсутствуют точные аналоги многозначных русских слов с абстрактными значениями (например, «воля», «душа»), что отражается в высокой дисперсии эмбеддинговых соответствий. Модели LaBSE и XLM-R демонстрируют, что китайские эквиваленты таких слов тяготеют к синтаксически простым, но прагматически контекстуализированным конструкциям.

Нейросетевой анализ позволяет выявить устойчивые синтаксические преобразования между русским и китайским. Например, конструкции с деепричастиями в русском зачастую преобразуются в сложносочиненные или номинативные конструкции в китайском. С помощью attention-весов и токеновых выравниваний можно проследить, какие элементы предложения подвергаются перестройке, и как формируется прагматическая эквивалентность.

Особый интерес представляет выявление прагматических компенсаций. Например, в китайских переводах часто наблюдается добавление частиц (啊, 吧) или переформу-

лировок, компенсирующих эмоционально-экспрессивную нагрузку русских конструкций. Нейросети с интерпретируемыми слоями (например, LIME-анализ attention-весов) позволяют проследить, какие элементы исходного текста инициируют такие компенсации.

Использование моделей zero-shot или few-shot позволяет встраивать тексты в семантическое пространство и автоматически классифицировать переводческие приемы: калькирование, адаптацию, эллипсис, экспликацию. Это открывает путь к построению автоматических карт перевода с метками стратегий и их статистическим распределением.

Наконец, анализ больших параллельных корпусов с помощью ИИ позволяет подтвердить или опровергнуть гипотезы о трансляционных универсалиях. Например, частотный анализ эмбеддинговых кластеров показывает, что в китайских переводах наблюдается тенденция к снижению структурной сложности по сравнению с оригинальным русским текстом – что коррелирует с гипотезой «упрощения» (simplification).

Интеграция ИИ в сравнительную лингвистику открывает качественно новые горизонты как в теоретическом осмыслении языковых связей, так и в практическом анализе корпусов. Использование нейросетевых моделей позволяет не только автоматизировать процессы выравнивания и анализа, но и выходить за пределы традиционных методологических парадигм, выявляя скрытые уровни лингвистических соответствий.

Среди ключевых перспектив можно выделить:

- Повышение интерпретируемости моделей: дальнейшее развитие explainable AI (XAI) обеспечит лучшее понимание того, как ИИ принимает решения, что важно для лингвистической верификации результатов.

- Создание специализированных параллельных корпусов: акцент на лексико-грамматические особенности и культурные концепты позволит проводить более тонкие сопоставления.

- Учет прагматических и дискурсивных параметров: применение моделей, способных анализировать не только структуру, но и функцию высказываний, расширит границы сравнительного анализа.

- Междисциплинарная интеграция: соединение лингвистики, когнитивных наук, ИИ и переводоведения приведет к созданию комплексных моделей языковой интерпретации.

Таким образом, будущее сравнительной лингвистики в эпоху ИИ связано не только с автоматизацией, но и с переосмыслением самой природы лингвистического знания. Русско-китайская языковая пара служит ярким примером того, как технологии могут пролить свет на ранее неочевидные взаимосвязи между далекими по структуре и культуре языками.

Развитие этих направлений требует тесного сотрудничества между лингвистами, программистами, переводчиками и специалистами по данным. Только в этом случае потенциал ИИ будет реализован в полной мере, обеспечивая как научную, так и прикладную ценность для многоязычного мира.

Литература

1. Politova E., Bonetskaya N., Tihonova M., Podolskiy A., Panchenko A. Cross-Domain Analysis of Multilingual Pre-Trained Language Models // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021. Pp. 1121–1129.
2. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT Sentence Embedding // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022). 2022. Vol. 1: Long Papers. Pp. 828–842.
3. Wang Z., Mayhew S., Roth D. How do Languages Influence Each Other? Studying Cross-Lingual Data Sharing During LM Fine-tuning // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2023. Pp. 13994–14012.
4. Feng G., Zhang R., Zhang M., Liu Q. A General and Robust Framework for Importance Weighting Correction of Multilingual Contrastive Learning // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). 2024. Vol. 1: Long Papers. Pp. 12583–12600.
5. Danihelka J., Rychlý P. A Contextual Embeddings-Based Approach for Advanced Word Alignment in Parallel Corpora // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024. Pp. 11886–11896.
6. Baker Mona. Corpus Linguistics and Translation Studies: Implications and Applications // Text and Technology: In Honour of John Sinclair / Edited by Mona Baker, Gill Francis, Elena Tognini-Bonelli. – Amsterdam / Philadelphia: John Benjamins Publishing Company, 1993. Pp. 233–250.
7. Chesterman Andrew. Beyond the Particular // Translation Universals: Do They Exist? / Edited by Anna Mauranen, Pekka Kujamäki. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2004. Pp. 33–49.

Статья рекомендована к печати кафедрой английского и китайского ОГУ
(докт. филол. наук, доцент, И. А. Солодиловой).

Linguistic Research in the Age of Artificial Intelligence: Russian-Chinese Parallel Corpora

A. V. Pavlova*, D. V. Askarova

Orenburg State University
18 Pobeda av., 460018 Orenburg, Russia.

**Email: pavlova_a@bk.ru.*

The article is devoted to the study of the role of artificial intelligence (AI) in modern linguistic research. Special attention is paid to the potential of neural network methods in the analysis and identification of hidden linguistic patterns in parallel Russian-Chinese corpora. Modern models, technologies, and limitations of their application are considered. Prospects for the development of methodology in the context of digitalization and interpretability are proposed.

Keywords: linguistics, artificial intelligence, Russian, Chinese, language corpus.