ISSN 3060-4567

ЭТАПЫ И ПРИНЦИПЫ СОЗДАНИЯ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВЫХ КОРПУСОВ

Файзиева Сафия Алишеровна

Преподаватель факультета языков БГПИ

fayziyevasafiya@buxdpi.uz

Абдуллаева Исмигул Баходир кизи

Студентка 3-ого курса факультета языков БГПИ

ismigulabdullayeva@icloud.com

Аннотация. *В статье рассматриваются этапы и принципы построения параллельных текстовых корпусов, используемых в лингвистических исследованиях и образовательных целях. Выделены основные этапы создания корпусов: от отбора репрезентативных текстов исходного языка до их согласования с переводами на другие языки и внедрения многоуровневой разметки. Описаны критерии, обеспечивающие репрезентативность и целевую направленность корпусов, включая жанровое разнообразие, тип языковых данных и доступность для компьютерной обработки. Особое внимание уделено характеристикам, определяющим функциональность параллельных корпусов, таким как контекстуальная согласованность, синхронность текстов и адаптация под профессионально-ориентированные задачи. Проанализированы дидактические возможности параллельных корпусов, обеспечивающие точное изучение языковых закономерностей и лексической вариативности.*

Ключевые слова: *параллельные текстовые корпусы, принципы создания корпусов, этапы построения корпусов, репрезентативность, контекстуальная согласованность, многоуровневая разметка, профессиональная лексика, лингвистический анализ.*

Введение. Параллельные текстовые корпусы представляют собой инструмент, позволяющий проводить систематический анализ языковых

явлений на основе текстов оригиналов и их переводов. Они обладают высокой ценностью для лингвистических исследований и образовательных целей, обеспечивая точность интерпретации языковых единиц и анализ их употребления в сопоставимых контекстах.

Процесс создания таких корпусов требует строгой методологической последовательности. Это включает отбор текстов, репрезентативных по жанровому, стилистическому и временному признакам, синхронизацию исходных данных с переводами и внедрение разметки для последующего анализа. Принципы формирования корпусов определяют их практическую ценность: репрезентативность, многоязычие, разметка на морфологическом и синтаксическом уровнях, а также возможность использования в профессионально-ориентированных задачах.

Изучение этапов и принципов построения параллельных текстовых корпусов необходимо для совершенствования методов анализа языка, разработки современных учебных материалов и построения автоматизированных систем обработки текста. Настоящее исследование направлено на систематизацию подходов к созданию корпусов и их адаптацию под требования научных и образовательных задач.

Материалы и методы. Для исследования этапов и принципов создания параллельных текстовых корпусов были отобраны тексты на двух языках, соответствующие жанровым и тематическим требованиям. Исходными материалами служили научные статьи, художественные произведения и техническая документация. Принципы отбора текстов базировались на репрезентативности жанрового состава, синхронности текстов по времени публикации и их переводной идентичности.

Обработка текстов включала предварительную подготовку: разделение на сегменты (предложения или абзацы), токенизацию и выравнивание оригиналов и переводов. Использовались специализированные инструменты для выравнивания, такие как HunAlign и Bleualign, что обеспечило точное соответствие текстовых сегментов.

Для разметки корпуса применялись автоматизированные системы аннотирования, включая Stanford NLP и SpaCy, с целью определения морфологических и синтаксических характеристик текстов. Процесс разметки включал идентификацию частей речи, лемматизацию и построение синтаксических деревьев. Дополнительно были внедрены семантические маркеры для анализа лексических единиц в контексте.

Результатом процесса стало формирование базы данных, объединяющей тексты и их переводы в единой структуре. Для анализа корпуса использовались программы AntConc и Sketch Engine, позволяющие осуществлять поиск языковых закономерностей, анализировать частотность употребления лексических единиц и их контексты.

Созданный корпус прошел оценку качества, включая проверку соответствия оригинала и перевода, корректность сегментации и полноту разметки. Эти этапы обеспечили возможность использования корпуса для дальнейших исследований в области лингвистики и прикладных задач.

Литературный обзор. Процесс создания параллельных текстовых корпусов включает несколько ключевых этапов. В первую очередь осуществляется сбор текстов на исходном языке и их переводов. Для достижения репрезентативности отбираются материалы, охватывающие разные жанры и стили, что обеспечивает возможность изучения широкого спектра языковых особенностей [1]. На следующем этапе проводится выравнивание текстов, при котором каждая единица исходного языка соотносится с соответствующей единицей в переводе. Этот процесс базируется на алгоритмах автоматизированного выравнивания, учитывающих лексические и синтаксические различия между языками [2].

После выравнивания текстов осуществляется лингвистическая разметка, включающая определение морфологических и синтаксических характеристик языковых единиц. Данная разметка необходима для анализа лексико-грамматических особенностей и выявления закономерностей перевода. Современные программные инструменты позволяют проводить

автоматическую разметку с высокой степенью точности [3]. На заключительном этапе формируется база данных корпуса, которая предоставляет доступ к параллельным текстам и инструментам для их анализа. Это включает возможность выполнения поисковых запросов, изучения контекстов употребления и анализа частотных характеристик [4].

Принципы, лежащие в основе создания параллельных текстовых корпусов, направлены на обеспечение репрезентативности текстов, точности выравнивания и универсальности разметки. Следование этим принципам позволяет использовать корпуса в лингвистических исследованиях, разработке языковых технологий и образовательной деятельности [5].

Результаты и обсуждение. В рамках исследования был сформирован экспериментальный параллельный текстовый корпус на основе научных статей, посвященных инженерным и гуманитарным наукам, и их переводов. Объем корпуса составил 250 000 словоформ, включающих 125 000 словоформ на русском языке и 125 000 словоформ на английском языке. Основное внимание уделялось анализу структурных и лексических особенностей текстов.

Для выравнивания текстов использовалась система HunAlign. Выравнивание предложений показало точность 96,2%, что позволило сформировать сопоставимые пары на уровне предложений. После этого проведена лексико-грамматическая разметка текстов с использованием инструмента SpaCy, включающая автоматическое определение частей речи, синтаксической структуры и лемматизацию.

Сравнительный анализ синтаксических конструкций выявил различия между текстами оригинала и перевода. На русском языке преобладают сложноподчиненные конструкции (в 34,7% предложений), в то время как на английском чаще используются простые предложения (в 41,2% случаев). Средняя длина предложений в русском корпусе составила 14,9 словоформ, в английском — 12,4 словоформ.

Наиболее частотными лексическими единицами в обоих текстах стали специализированные термины, такие как *analysis*, *methodology*, *data* на английском языке и *анализ*, *методология*, *данные* на русском. Общая доля терминологической лексики составила 15,4% от общего объема корпуса. Корреляция частот терминов между русским и английским текстами достигла 88,1%, что указывает на высокую степень лексического соответствия.

Таблица 1.

Сравнительный анализ синтаксических конструкций

Тип синтаксической конструкции	**Русский текст (%)**	**Английский текст (%)**
Простые предложения	25.8	41.2
Сложные предложения	39.5	32.8
Сложноподчиненные конструкции	34.7	26.0

Полученные данные демонстрируют характерные различия в синтаксических и лексических структурах текстов на русском и английском языках. Преобладание сложноподчиненных конструкций в русском языке отражает его синтаксическую специфику, что подтверждает необходимость адаптации автоматизированных инструментов разметки для работы с корпусами на разных языках.

Точность автоматического выравнивания предложений (96,2%) свидетельствует о надежности используемых алгоритмов, однако оставшиеся 3,8% ошибок требуют последующей ручной проверки. Высокий уровень терминологической корреляции между текстами (88,1%) подчеркивает качество перевода и сопоставимость материалов для использования в лингвистических исследованиях.

Дальнейшее развитие проекта предусматривает увеличение объема корпуса до 500 000 словоформ, а также включение текстов из других научных областей для повышения репрезентативности. Это позволит проводить более детальный анализ языковых закономерностей и повысить универсальность корпуса для решения научных задач.

Заключение. Результаты исследования подтвердили значимость этапов и принципов, лежащих в основе создания параллельных текстовых корпусов. Проведенная работа по формированию экспериментального корпуса объемом 250 000 словоформ продемонстрировала высокую точность автоматизированного выравнивания текстов (96,2%) и позволила выявить различия в синтаксических структурах русского и английского языков. Преобладание сложноподчиненных конструкций в русском языке и более частое использование простых предложений в английском подчеркивают необходимость учета синтаксической специфики при разработке параллельных корпусов.

Лексический анализ показал высокий уровень терминологического соответствия между текстами оригинала и перевода, что подтверждает возможность использования параллельных текстовых корпусов для изучения лексических закономерностей, разработки учебных материалов и повышения качества перевода.

Дальнейшая работа будет направлена на увеличение объема корпуса, включение текстов из различных научных областей и внедрение дополнительных уровней разметки. Это расширит возможности использования корпусов в лингвистических исследованиях, автоматизированной обработке текста и образовательной практике.

СПИСОК ЛИТЕРАТУРЫ

1. Mikhailov M., Cooper R. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research // Target. – 2016. – Vol. 28. – No. 2. – P. 224–239.

2. Čermák F. Corpus Linguistics: Theoretical Foundations and Practical Applications // Corpus Linguistics and Linguistic Theory. – 2017. – Vol. 13. – No. 1. – P. 31–58.

3. Bowker L., Pearson J. Working with Specialized Corpora: Tools and Techniques for Translators. – Routledge, 2020. – 240 p.

4. Елисеева Т.В., Назарова А.А. Корпусная лингвистика: инструменты и методы анализа параллельных текстов // Известия РАН. Серия литературы и языка. – 2019. – Т. 78. – № 3. – С. 5–21.

5. Тютюнов А.Л., Горбаневский М.А. Технологии параллельных корпусов в переводоведении // Вестник СПбГУ. Серия 9. – 2018. – № 2. – С. 145–159.

6. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge University Press, 2012. – 320 p.

7. Anthony L. AntConc: Design and Use of a Free Corpus Analysis Toolkit for TESL/TEFL // TESL-EJ. – 2004. – Vol. 8. – No. 1. – P. 1–16.

8. Baker M. Corpus-based Translation Studies: The Challenges That Lie Ahead // Benjamins Translation Library. – 1995. – Vol. 18. – P. 175–186.

9. Сичинава Д.В. Параллельные тексты в составе Национального корпуса русского языка: новые направления развития и результаты // Труды Института русского языка им. В.В. Виноградова. – 2015. – № 21. – С. 195–204.

10. Garside R., Leech G., McEnery T. Corpus Annotation: Linguistic Information from Computer Text Corpora. – Routledge, 1997. – 320 p.

11. Fayziyeva S. A., Dilnoza J. THE ARTISTIC ORIGINALITY OF L. PETRUSHEVSKAYA'S CREATIVITY //ОБРАЗОВАНИЕ НАУКА И ИННОВАЦИОННЫЕ ИДЕИ В МИРЕ. – 2024. – Т. 47. – №. 6. – С. 162-165.

12. Файзиева С. А. Жанровое своеобразие литературной сказки //Development of pedagogical technologies in modern sciences. – 2024. – Т. 3. – №. 1. – С. 52-56.

ISSN 3060-4567

13. Файзиева С. А. GENRE ORIGINALITY OF A LITERARY FAIRY TALE //Web of Teachers: Inderscience Research. – 2023. – Т. 1. – №. 7. – С. 58-62.

14. Zanettin F. Corpus Methods for Translation Studies: Bridging the Gap Between Theory and Practice. – Routledge, 2012. – 256 p.

15. Серегин А.С., Мельникова Н.А. Параллельные корпусы и их применение в переводоведении и изучении языка // Вестник ВГУ. Серия Лингвистика и межкультурная коммуникация. – 2017. – № 2. – С. 112–119.