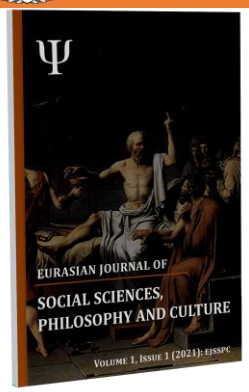
**Письмо-согласие**

Я, Юсупова Альфия Шавкетовна, написала в соавторстве с Авезовым Сухробом Собировичем научную статью под названием «Процесс обработки узбекского параллельного корпуса в условиях недостаточности данных», опубликованную в Международном журнале «Eurasian Journal of Social Sciences, Philosophy and Culture». Я хотела бы подтвердить, что я согласна с тем, чтобы докторант использовал эти статьи в своих исследованиях, и я не возражаю против этого. 60% этих материалов принадлежат исследователю.


Доктор филологических наук,
профессор кафедры общего языкознания и тюркологии
КФУ, Институт филологии и межкультурной коммуникации
Юсупова Альфия Шавкетовна

EURASIAN JOURNAL OF SOCIAL SCIENCES, PHILOSOPHY AND CULTURE
Innovative Academy Research Support Center
UIF = 8.2 | SJIF = 6.051 www.in-academy.uz




# ПРОЦЕСС ОБРАБОТКИ УЗБЕКСКОГО ПАРАЛЛЕЛЬНОГО КОРПУСА В УСЛОВИЯХ НЕДОСТАТОЧНОСТИ ДАННЫХ

**[1]Авезов Сухроб Собирович**
Бухарский государственный университет
1990senigama@gmail.com,
**[2]Юсупова Альфия Шавкетовна**
КФУ / Институт филологии и межкультурной коммуникации
Alfiia.Iousoupova@kpfu.ru.
https://www.doi.org/10.5281/zenodo.7714460

**ARTICLE INFO**
Received: 28th February 2023
Accepted: 09th March 2023
Online: 10th March 2023
**KEY WORDS**
*Параллельные корпуса, системы машинного перевода, морфология языка, методы выравнивания, методы фильтрации, неконтролируемый метод выравнивания, NLP (обработка естественного языка).*

**ABSTRACT**
*Параллельные корпуса являются необходимым компонентом для разработки качественных систем машинного перевода, однако сбор соответствующих данных представляет собой сложную задачу. Когда богатая морфология языка увеличивает разреженность данных, необходимо иметь точные методы выравнивания и фильтрации, которые позволят эффективно использовать имеющуюся информацию, максимально увеличивая количество корректно переведенных сегментов в корпусе и минимизируя наличие шума, путем удаления неправильных переводов и сегментов, содержащих посторонние данные. В данной статье описывается план исследования по улучшению методов выравнивания и фильтрации параллельных текстов в условиях ограниченных ресурсов. Предлагается эффективный неконтролируемый метод выравнивания, способный решить проблему выравнивания, а также стратегия дополнения современных моделей автоматически извлекаемой информацией, с использованием основных инструментов NLP для эффективной обработки богатой морфологии языков.*

## 1. Введение

Наличие нейронных систем машинного перевода (NMT) значительно улучшило качество машинного перевода (MT). Тем не менее, в условиях ограниченных ресурсов и несоответствия доменов, преимущество в качестве по сравнению с системами статистического машинного перевода (SMT) уменьшается. В последнее время неконтролируемый NMT, обученный только на одноязычных корпусах, привлекает значительное внимание и используется в сценариях, где отсутствуют двуязычные данные. Однако эти методы хорошо работают только для родственных языковых пар.

Для языков, которые значительно отличаются друг от друга, неконтролируемые методы становятся менее эффективными. Кроме того, неконтролируемый NMT чувствителен к несоответствию доменов, что создает проблему для языковых пар с низким уровнем ресурсов. Поэтому для достижения высокого качества машинного перевода необходимо иметь тексты, выровненные по предложениям на двух или более языках. Ряд исследований также показывает, что контролируемые и полуконтролируемые методы с небольшим параллельным корпусом превосходят лучшие неконтролируемые системы для языков, как родственных, так и неродственных.

В научных исследованиях установлено, что системы машинного перевода (NMT) обладают чувствительностью к шуму в обучающих данных, при этом шумом считаются сегменты, негативно сказывающиеся на качестве перевода, полученном при обучении на таких данных. В связи с этим, важно обладать точностью в выравнивании многоязычных текстов и аккуратно фильтровать несоответствия и некорректные переводы, чтобы не допустить ухудшения производительности системы. Один из предыдущих экспериментов, нацеленных на изучение влияния различных типов шумов на качество машинного перевода, выявил, что наиболее критическими являются непереведенные и смещенные сегменты. Однако смещения могут быть разнообразными, такие сегменты могут содержать одно лишнее слово, в два раза больше информации, чем аналог в другом языке, либо иметь средний уровень смещения. В связи с этим, особенно полезным является понимание тонкостей воздействия разных видов и уровней шума, а также оценка того, насколько важно их избегать и какие их типы являются более приемлемыми, чем другие. Описанная проблематика приводит к первому исследовательскому вопросу: «Как различные виды несоответствий в параллельном корпусе влияют на качество перевода системы машинного перевода (SMT или NMT), обученной на этом корпусе?»

Если будет возможность оценить влияние различных типов несоответствий в параллельных корпусах, это позволит разработать более эффективные методы фильтрации таких корпусов для обучения систем машинного перевода.

При первоначальном осознании полезности параллельных корпусов для машинного перевода, стало очевидно, что согласование таких текстов является серьезной проблемой. Кроме того, сбор многоязычных текстов является затратным и трудоёмким процессом, а для некоторых языков получение даже небольшого количества текстов может быть осложнено. Таким образом, необходимо максимально эффективно использовать имеющиеся ресурсы [1].

В данной статье описывается метод выравнивания параллельных корпусов с использованием только одноязычных текстов для обучения, с помощью неконтролируемой системы SMT Bleualign и Monoses. Предлагаемый метод не зависит от конкретной пары языков и основан на использовании только невыровненных битекстов и одноязычных корпусов для каждого из языков. Этот метод представляет собой первый шаг в направлении ответа на второй вопрос исследования: как эффективно создавать полезные параллельные корпуса из двуязычных текстов без использования дополнительных ресурсов, кроме одноязычных корпусов?

В рамках морфологической типологии языков можно обнаружить различие между аналитическими и синтетическими языками. Аналитические языки основываются на использовании порядка слов и служебных слов для передачи значения, в то время как синтетические языки опираются на использование морфологических правил [2]. Особенно «морфологически богатые» языки, которые относятся к синтетическим, характеризуются большим числом грамматических форм для каждого лексического элемента, что может привести к проблеме разреженных данных и привести к появлению слов вне словарного запаса (OOV), которые необходимо учитывать при использовании алгоритмов машинного обучения.

В нашей работе мы сосредоточимся на построении параллельного корпуса для узбекско-русской языковой пары и решении проблем, возникающих при работе с менее ресурсным и морфологически богатым языком.

При выравнивании предложений и фильтрации шума из параллельных корпусов проблема разреженности, вызванная богатой морфологией, приводит к более низким показателям достоверности для пар сегментов, что приводит к более низкой точности классификации и, следовательно, к меньшим или менее точным данных параллельных корпусов. Когда UzbRusCorp был скомпилирован, процесс фильтрации привел к сокращению размера корпуса примерно на 20%. Из того, что осталось, около 5% были неисправны. Мы будем работать с теми же данными с целью минимизировать эти числа. Это подводит нас к третьему и последнему исследовательскому вопросу, вокруг которого сосредоточено это исследовательское предложение: «Как мы можем фильтровать параллельные корпуса, чтобы свести к минимуму шум и при этом потерять мало или совсем не потерять полезные данные из исходных текстов?»

Для ответа на исследовательские вопросы, связанные с выравниванием и фильтрацией текстов, мы предлагаем экспериментальный подход, основанный на использовании широко распространенных методов из соответствующей литературы. Для этого мы разработаем набор инструментов, способный применять различные известные методы и сравнивать их эффективность [3]. В рамках исследования мы будем изучать как использование лемматизации, тегирования частей речи (PoS) и синтаксического анализа может помочь преодолеть проблему разреженности данных, а также определить, какие известные методы могут дать наилучшие результаты. Для оценки эффективности разработанных методов будут созданы наборы данных, и результаты оценки будут измеряться по ряду критериев. Наконец, мы планируем обучить и оценить нашу систему на другой паре языков с аналогичными проблемами.

**2. Связанные работы**

Фильтрация параллельных данных является процессом удаления ошибочных переводов, шума и других неточностей из двух или более выровненных текстов. В свою очередь, выравнивание многоязычных текстов заключается в поиске соответствующих сегментов целевого текста для каждого исходного сегмента. Несмотря на то, что эти две задачи кажутся различными, некоторые методы могут быть применены к обеим проблемам. Фильтрация часто осуществляется путем оценки предложений и удаления менее оцененных, в то время как в процессе выравнивания предложения с наивысшей оценкой могут служить опорными элементами, которые надежно выравниваются и

могут быть использованы для дальнейшей обработки. В следующих разделах мы опишем методы выравнивания и фильтрации, примененные в предыдущих исследованиях.

**2.1. Выравнивание и фильтрация**

Изначально автоматическое выравнивание предложений осуществлялось на основе длины. В работе Гейла и Черча было обнаружено, что корреляция между длиной абзаца в символах и длиной его перевода является значительной. Основываясь на этом, они описывают метод выравнивания предложений, основанный на статистической модели длины символов. Браун и соавторы также используют длину, однако в качестве единицы измерения используются токены вместо символов. Кроме того, они используют сигналы в разметке для определения опорных точек, которые помогают разбить корпус на более мелкие фрагменты.

Кей и Рёшайзен использовали двуязычные лексиконы, созданные в результате выравнивания корпуса, для автоматического выравнивания параллельных текстов. Харуно и Ямазаки продемонстрировали, что использование индуцированного словаря совместно с внешним словарем приводит к улучшению качества автоматического выравнивания. Папагеоргиу и соавторы предложили использовать части речи, которые сохраняются при переводе, для определения оптимального выравнивания на основе PoS-тегов. Чорн и Люделинг усовершенствовали меру расстояния на основе словаря, используя морфологический анализатор, а Ма повысил эффективность выравнивания на основе лексической информации, присваивая больший вес низкочастотным переводимым словам.

Сеннрих и Волк применяют метод машинного перевода в сочетании с метрикой BLEU для выявления надежных сопоставлений, которые могут быть использованы в качестве опорных точек. Оставшиеся промежутки между опорными точками заполняются эвристиками, основанными на метрике BLEU и длине фрагментов.

Томпсон и Коэн описывают метод, основанный на встраивании двуязычных предложений, используя сходство между вложениями в качестве функции оценки для выравнивания.

В современной лингвистике находят все большее применение методы машинного обучения, в том числе и нейронные сети, для выявления опорных точек и несоосностей в параллельных текстах [4]. Многие из этих методов основаны на обучении классификаторов, которые могут определять, являются ли исходное и целевое предложения параллельными, что позволяет извлекать параллельные предложения из сопоставимых корпусов.

Ранее проведенные исследования в области выравнивания параллельных корпусов включают использование моделей IBM для сопоставления слов. Хадиви и Ней применяют модели IBM 1 и 4, а также модели, основанные на длине, для фильтрации шумного контента в корпусе и оценивают выравнивание на основе их линейной комбинации. Тагипур и коллеги обнаруживают выбросы и показывают, что использование отфильтрованного корпуса приводит к улучшению качества перевода, даже если некоторые предложения были удалены [5]. Сарыкая и коллеги применяют экстраполяцию контекста для расширения охвата пар предложений, оценивая, имеют

ли предложения наивысший показатель сходства в пределах определенного окна, несмотря на то, что они находятся на расстоянии от точки привязки. В настоящее время в данной области все чаще используются нейронные сети для поиска опорных точек и выявления несоосностей путем обучения классификаторов, определяющих, являются ли исходные и целевые предложения параллельными.

В последние годы для решения задачи поиска эквивалентностей между словами разных языков использовались межъязыковые вложения слов. Некоторые исследователи рассматривают задачу фильтрации как задачу supervised регрессии и демонстрируют, что расстояние Левенштейна между исходным и целевым переведенными на машинный перевод текстами, а также косинусное расстояние между вложениями предложений исходного и целевого текстов являются важными характеристиками. Несмотря на то, что некоторые исследователи используют метод InferSent, недавние исследования показали, что использование модели BERT для расчета межъязыкового семантического текстового сходства дает хорошие результаты в задаче обнаружения несоответствий.

Zipporah применяет модель логистической регрессии для классификации пар предложений, обучая ее на синтезированных данных с зашумлением в качестве отрицательных примеров. BiCleaner использует набор жестких правил, созданных вручную, для выявления ошибочных предложений, а затем использует классификатор случайного леса, основанный на лексических переводах и неглубоких функциях, таких как длина, совпадающие числа и знаки препинания. Наконец, BiCleaner оценивает беглость речи предложений с помощью 5-граммовых языковых моделей.

В рамках четвертой конференции по машинному переводу (WMT), которая прошла в 2019 году, была поставлена задача по параллельной фильтрации корпусов в условиях ограниченных ресурсов. Одним из основных методов для эффективного представления предложений было использование межъязыковых вложений, обученных из параллельных пар предложений. Artetxe и Schwenk представили сходный подход, который также решает проблемы несоответствий косинусного подобия путем анализа окрестностей данной пары предложений, превосходя системы, использующие только косинусное сходство.

**3 Экспериментальная структура (Framework) и базы данных**

Разнообразие морфологически богатых языков можно описать в терминах континуума, где агглютинативные языки на одном конце используют дискретные морфемы для флексии, а флективные языки на другом конце используют одну флективную морфему для выражения нескольких признаков. Вопрос о том, какие методы могут использоваться для обработки различных языковых категорий, требует дальнейших исследований, поскольку агглютинативные и флективные языки могут потребовать разных подходов для решения задачи обработки текста. Для решения проблемы неизвестных слов (OOV) может быть полезно использовать разложение для агглютинативных языков, а для флективных языков требуются дополнительные методы для обработки внутренних изменений. В данном исследовании мы сфокусировались на агглютинативных языках и выбрали языковую пару узбекский-русский в качестве тестового примера.

Корпус UzbRusCorp, содержащий параллельные тексты на узбекском и русском языках, был собран на основе 100 000 сегментов перевода. После выравнивания с помощью LF Aligner, корпус был отфильтрован с помощью алгоритма подсчета предложений, использующего двуязычный словарный мешок слов и метод сравнения исходного сегмента и машинного перевода. В результате фильтрации было получено 70 000 сегментов. Примерно 2000 пар образцов из корпуса были оценены вручную, при этом было выявлено около 5% ошибочных переводов. Более 50% удаленных сегментов были оценены как ошибочные с использованием автоматических методов.

По результатам анализа было выявлено, что необработанный корпус UzbRusCorp, состоящий из 100 тысяч сегментов, содержит около 30 тысяч ошибочных сегментов, вызванных в основном смещением. В связи с этим, для создания наилучшего корпуса необходимы усовершенствованные методы выравнивания, позволяющие уменьшить количество ошибочных выравниваний, а также классификаторы, способные определить качество сегментов с высокой точностью и отзывом. Это позволит собрать как можно больше сегментов с минимальным количеством ошибок. В настоящее время мы сосредоточены на работе с необработанными данными UzbRusCorp, исходя из описанных принципов.

**3.1 Оценка результатов. Инструменты и модели**

В рамках исследования мы подготовили три оценочных набора, извлеченных из корпуса UzbRusCorp, предназначенных для оценки качества выравнивания, фильтрации и машинного перевода. Оценочный набор машинного перевода включает в себя 3000 сегментов, выровненных вручную и не содержащих ошибок. Набор оценки выравнивания состоит из 2000 предложений, также выровненных вручную. Набор фильтрации содержит 2000 автоматически выровненных сегментов, каждый из которых был отнесен к одному из четырех классов: «правильный», «частично смещенный», «частично неверный перевод» и «неправильный».

С целью оценки эффективности наших методов в контексте машинного перевода, мы планируем использовать наши выровненные и отфильтрованные корпуса для обучения систем машинного перевода SMT и NMT. Далее, мы проведем сравнительный анализ полученных результатов с базовым уровнем, где для обучения систем машинного перевода используется необработанный корпус UzbRusCorp.

В данной работе рассматриваются различные методы обработки текстов, среди которых применение доступных инструментов и моделей, а также разработка собственных. Для PoS-маркировки русских текстов будет использоваться ABLTagger, использующий biLSTM и внешний морфологический словарь. Лемматизация осуществляется с помощью Nefnir. Для обработки узбекского языка в данной работе будут использоваться инструменты, доступные в наборах инструментов NLTK или SpaCy.

В данной работе мы сосредоточимся на изучении наиболее распространенных моделей встраивания слов, таких как word2vec, GloVe, FastText и ELMo. Кроме того, мы также рассмотрим возможность использования контекстуализированных моделей встраивания слов на основе двуязычных вложений предложений с помощью BERT. Однако, следует отметить, что использование таких моделей требует значительных

вычислительных ресурсов для обучения, что может ограничить наши возможности и стать значимым препятствием в нашей работе.

Мы планируем провести эксперименты с использованием различных инструментов и методов для выравнивания и фильтрации параллельных текстов. Среди таких инструментов будут Bleualign, Hunalign и vecalign для выравнивания предложений, Giza++ для выравнивания слов, а также Zipporah, BiCleaner и LASER для фильтрации и возможной поддержки в привязке параллельных текстов для более эффективного выравнивания. Наша цель – найти наилучшие методы для выравнивания и фильтрации, которые могут быть применены в последующих исследованиях в области машинного перевода и связанных задач.

Для реализации машинного перевода методом SMT на основе фраз мы воспользуемся системой Мозес. Для реализации метода NMT мы будем использовать эталонную реализацию архитектуры на основе трансформатора Vaswani, входящую в состав пакета Tensor2Tensor.

**4 План исследования**

Одной из наших первоочередных задач является настройка неконтролируемого конвейера для автоматического выравнивания параллельных текстов. Несмотря на то, что это первый шаг в решении третьей задачи, которая была поставлена ранее, также необходимо разработать методологию для ответа на первую задачу. В данном контексте мы обрисуем общие принципы, которые мы собираемся использовать, чтобы ответить на эти вопросы, а также на третий вопрос. Наша вторая цель – изучение методов улучшения неконтролируемого конвейера путем использования базовых инструментов обработки естественного языка (NLP) для решения проблемы разреженности данных, которая характерна для многих морфологически богатых языков. В последующих подразделах мы описываем подробнее, как мы планируем исследовать эти вопросы.

**4.1 Unsupervised выравнивание**

Наша исходная методология для выравнивания параллельных текстов заключается в обучении конвейера лишь на одноязычных корпусах. Этот подход является отправной точкой для тех языковых пар, для которых ранее не было доступных параллельных корпусов или глоссариев, используемых для выравнивания. Тем не менее, он также служит отправной точкой для дальнейшего сравнения с улучшенными методами обработки, включая лемматизацию и другие инструменты NLP.

| | **LF Aligner** | | | **Bleualign + Monoses** | | |
|---|---|---|---|---|---|---|
| | **Нормативные тексты** | **Литературные тексты** | **Всего** | **Нормативные тексты** | **Литературные тексты** | **Всего** |
| **Выровненные пары** | 184 | 69 | 253 | 166 | 61 | 227 |
| **правильные** | 143 | 57 | 79,1% | 154 | 54 | 91,6% |
| **ошибочные** | 41 | 12 | 20,9% | 12 | 7 | 8,4% |
| **Выровненн** | 2470/248 | 1652/1652 | 4122/41 | 2427/2485 | 1539/1652 | 3966/41 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **ые слова** | 5 | | 37 | | | 37 |
| **правильные** | 1980 | 1337 | 80,5% | 2110 | 1539 | 92,0% |

Таблице 1. Здесь приведены результаты сравнения обеих систем, а также количество слов исходного языка в каждом сравнении. В случае, если выравнивание не было обнаружено, соответствующие сегменты были исключены из анализа.

Как было указано ранее, в данной работе для неконтролируемого выравнивания используется Bleualign. В отличие от методов, основанных на длине, таких как Sennrich и Volk, мы используем Monoses для создания машинных переводов выравниваемых предложений, которые затем передаются в Bleualign. Для обучения Monoses мы создаем межъязыковые вложения слов на основе одноязычных корпусов, используя word2vec и Vecmap, и строим таблицу фраз. Затем система SMT обучается на этих данных и используется для перевода одноязычных корпусов на один из двух языков. Преобразованные данные затем используются для обучения стандартной системы SMT в обратном направлении. Процесс повторяется три раза для окончательной модели, при этом каждый раз строится новая таблица фраз.

Для оценки эффективности предлагаемого метода выравнивания мы провели сопоставление двух случайно выбранных параллельных текстов из набора данных UzbRusCorp с помощью Bleualign и Monoses, а также сравнили полученные результаты с методом LF Aligner, использующим Hunalign. Для более точной оценки качества выравнивания были составлены оценочные наборы. Предварительные результаты, полученные путем ручной оценки выравнивания и представленные в таблице 1, показали, что метод Bleualign + Monoses обеспечивает более высокую точность выравнивания пар текстов. Конкретно, 92% полученных пар правильно выровнены с помощью данного метода по сравнению с только 80,5% совпадений при использовании метода LF Aligner. Несмотря на то, что предлагаемый метод приводит к 10% меньшему числу выровненных пар, он обеспечивает более правильные выравнивания как в абсолютных значениях, так и в процентном соотношении, независимо от того, рассматриваем ли мы выровненные пары или выровненные слова. Эти результаты говорят о том, что предлагаемый метод выравнивания является перспективным в решении задачи создания качественных параллельных корпусов.

Существует ряд методов, позволяющих усовершенствовать unsupervised метод выравнивания. Путем обучения более масштабных моделей встраивания слов можно расширить словарный запас, а изучение часто встречающихся n-граммов может способствовать более точному выделению фраз, состоящих из нескольких слов. Дальнейшее улучшение качества возможно при расширении процесса итерации на битексты, выборе пар наивысшего качества после обучения и выравнивания и добавлении их в обучающий набор системы SMT. Это позволит получить более точные данные для обучения и вероятно приведет к лучшим результатам перевода после каждой итерации, повышая уверенность в выборе оптимального выравнивания.

**Заключение**

После выполнения настройки конвейеров выравнивания и создания оценочных наборов, мы приступим к процессу фильтрации, применяя методы и стратегии, которые демонстрировали ранее высокую эффективность для других языковых пар.

Одним из важных аспектов процесса фильтрации в параллельных корпусах является определение наиболее значимых типов шума, которые необходимо исключить. Несмотря на то, что некоторые ученые обратили внимание на важность фильтрации определенных типов шума, для достижения более точных результатов требуется дополнительное исследование. В данном исследовании мы специально изучим различные классы несоосностей, что поможет определить, нужно ли рассматривать все несоосности одинаково или некоторые из них следует рассматривать как более нежелательные, чем другие.

Для получения как можно более чистого корпуса параллельных текстов мы применим доступные инструменты для агрессивной фильтрации возможных ошибочных выравниваний. Затем мы систематически изменяем выравнивания, чтобы ввести в корпус различные типы смещения. Воздействие этих вариаций будет исследовано с помощью обучения систем машинного перевода SMT и NMT и сравнения изменений в результирующих переводах. Этот метод предназначен для решения первого вопроса, поставленного выше. Мы будем использовать полученные результаты для принятия решения о настройке системы фильтрации.

Далее в процессе мы повторно запускаем фильтрацию с использованием информации о том, какие типы ошибок, скорее всего, окажут наихудшее влияние на системы машинного перевода, обученные на данных. Чтобы ответить на третий вопрос, мы исследуем практическую применимость различных механизмов для оценки предложений. Мы рассмотрим функции, такие как длина предложения, сходство слов на основе поиска в словаре, как с использованием внешнего словаря, так и с использованием словаря, созданного на основе необработанных параллельных данных, сходство слов из встраивания слов, дистанцию между машинным переведенным исходным предложением и целевым предложением, а также оценки сходства предложений на основе двуязычных вложений предложений. Мы намерены проанализировать эти функции с целью выявления наиболее эффективных механизмов оценки и фильтрации предложений в параллельных корпусах.

После того, как мы провели исследование о влиянии несоответствий на системы машинного перевода и найдем подходящий баланс между различными механизмами для оценки выровненных сегментов, мы продолжим наше исследование, чтобы оценить, как этот баланс зависит от языковой пары.

После того, как мы провели исследование о влиянии несоответствий на системы машинного перевода и найдем подходящий баланс между различными механизмами для оценки выровненных сегментов, мы продолжим наше исследование, чтобы оценить, как этот баланс зависит от языковой пары. Мы запустим тот же процесс для других пар языков.

Для языковой пары узбекский-русский не существует машиночитаемого словаря, и в случае использования методов полууправляемого машинного перевода необходимо создать лексикон из параллельных или одноязычных данных, либо из обоих типов данных. Другие методы создания глоссария могут включать использование внешних источников данных, таких как Викисловарь или Википедия, а также доступных словарей на разных языках с последующей их адаптацией.

Одним из результатов данного исследования будет создание набора инструментов для формирования параллельных корпусов из многоязычных текстов. Предоставляемое программное обеспечение должно обладать следующими функциями: автоматическое выравнивание двуязычных параллельных текстов; фильтрацию двуязычных параллельных корпусов; модульность; независимость от конкретной языковой пары, хотя возможно использование языко-специфических дополнительных функций; использование внешних инструментов для лингвистической аннотации, таких как PoStagging, синтаксический анализ, лемматизация, машинный перевод или другие методы, которые могут оказаться полезными; предоставление различных стратегий выравнивания и фильтрации в зависимости от доступных ресурсов; и достижение точности при оптимизации скорости работы.

## References:

1. Sobirovich A. S. Development of a Parallel Corpus of the Uzbek and Russian Languages //Vital Annex: International Journal of Novel Research in Advanced Sciences. – 2022. – Т. 1. – №. 5. – С. 152-155.
2. Авезов С. О корпусной лингвистике, трудностях перевода и принципах организации параллельных корпусов текстов //«УЗБЕКСКИЕ НАЦИОНАЛЬНЫЕ ОБРАЗОВАТЕЛЬНЫЕ ЗДАНИЯ ТЕОРЕТИЧЕСКОЕ И ПРАКТИЧЕСКОЕ СОЗДАНИЕ ВОПРОСЫ" Международная научно-практическая конференция. – 2022. – Т. 1. – №. 1.
3. Khamidovna N. L. Expression of the Harmony of Language and Culture in World and Uzbek Lexicography //resmilitaris. – 2023. – Т. 13. – №. 1. – С. 233-244.
4. Нигматова Л. Х. Некоторые проблемы узбекской лексикографии и разработки словаря //Ташкент Журнал Вопросы филологии. – 2020.
5. Sharipov S. ТАРЖИМАВИЙ ЛЕКСИКОГРАФИЯНИНГ ТАРИХИЙ ВА ХРОНОЛОГИК ХУСУСИЯТЛАРИ //ЦЕНТР НАУЧНЫХ ПУБЛИКАЦИЙ (buxdu. uz). – 2022. – Т. 15. – №. 15.