ЛИНГВОМЕТОДИЧЕСКИЙ ПОТЕНЦИАЛ ПАРАЛЛЕЛЬНОГО КОРПУСА

Хидирова Гулнора
кандидат филологических наук, доцент
Университет Маъмуна.

Аннотация: В статье рассматриваются структура и назначение параллельного корпуса. Под корпусом понимается совокупность устных и письменных текстов, хранящихся в базе данных компьютера, также в корпусе представлено подробное описание точного времени написания собранных материалов, стиль текстов. Корпус может определять диапазон слов, денотативные и коннотативные значения искомого слова, описывает частоту или статистику использования того и иного слова в языке писателя. Параллельный корпус описывается как новый вид лингвистических ресурсов, которые все больше используются при обучении языкам. На основе данных корпусов выявляются несовершенства современных учебников. В параллельном корпусе собраны тексты оригинала и его переводов. Параллельные корпуса служат обучающими данными для статистических систем машинного перевода. С точки зрения переводоведения важной задачей параллельного корпуса является обнаружение переводческих изменений. Основная цель создания параллельного корпуса – разработка эффективных и адекватных методов перевода. Посредством параллельного корпуса достигается улучшение качества перевода.

Ключевые слова: корпусная лингвистика, текст, стиль, параллельный, структура, сопоставительный, преподавание языка, переводоведение, машинный перевод, словари, процесс автоматической нормализации текстов, семантика, статистические технологии.

Annotation: The article discusses the purpose and functions of the parallel corpus. A corpus is understood as a collection of oral and written texts stored in a computer database; the corpus also provides a detailed description of the exact time of writing of the collected materials and the style of the texts. Modern methodological literature actively uses examples and descriptions from parallel corpora when teaching a language; increasingly, the corpus is used to analyze and identify the shortcomings of existing textbooks in teaching foreign languages. The corpus can determine the range of words, the denotative and connotative meanings of the searched word, and describes the frequency or statistics of the use of a particular word in the writer's language. The parallel corpus is described as a new type of linguistic resource. Parallel corpora serve as training data for statistical machine translation systems. The parallel corpus contains the texts of the original and its translations. From the point of view of translation studies, an important task of a parallel corpus is the detection of translation changes. The main goal of creating a parallel corpus is to develop effective and adequate translation methods. By using a parallel corpus, translation quality is improved.

Keywords: corpus linguistics, text, style, parallel, structure, comparative, language teaching, translation studies, machine translation, dictionaries, process of automatic text normalization, semantics, statistical technologies.

Введение. Одной из глобальных проблем XXI века является сохранение национальной идентичности естественных языков. Актуальной задачей мировой науки

стало последовательное проведение исследований нейролингвистического программирования и технологий в создании и развитии электронных языковых корпусов. Научные и практические исследования в области иностранной корпусной лингвистики доказали, что корпус является желательным и необходимым основанием не только для ученых лингвистов, занимающихся работой со словом, но и для развития и сохранения нации.

И в Узбекистане одной из важнейших задач, стоящих сегодня перед прикладной лингвистикой, стало создание Национального корпуса узбекского языка. Помимо научно-исследовательских задач, корпус даст импульс в достижении множества экстралингвистических целей, в частности:

- повышение престижа узбекского языка в обществе;

- создание Национального корпуса узбекского языка, содержащего всю научную, теоретическую и практическую информацию на узбекском языке;

- популяризация узбекского языка в мировой информационной сети Интернет и обеспечение его достойного места на международном уровне;

- создание узбекских приложений программных продуктов, компьютерного обучения узбекскому языку, работы по созданию компьютерных программ для редактирования текстов на узбекском языке [Указ Президента - №5850; 2019].

На настоящий момент это направление в тюркологии также активно разрабатывается, отметим труды лишь некоторых ученых, например, вопросы корпусного представления турецкого языка исследовали Дениз, Аксан, Зейрек, Кемал Офлазер, Умут Озге Булар; уйгурского языка - Юсуп Аибаидулла, Ким Тенг Луа; башкирского языка - З.А. Сиразитдинов, Л.А. Бускунбаева; хакасского языка - А.В. Шеймович, А. В. Дыбо, В. С. Мальцева, Э. В. Султрекова; татарского языка - Ж. Сулейманов, О. Невзорова, Б. Хакимов; тувинского языка - А.Я. Салчак, Мира В. Бавуу-Сюрюн, Чодураа М. Монгуш [Абдурахмонова Н.; 2020].

Концептология Национального корпуса узбекского языка разрабатывается коллективами ученых Государственного университета узбекского языка и литературы, Узбекского национального университета, Ташкентского университета информационных технологий, Самаркандского, Андижанского, Термезского госуниверситетов.

Корпус — это электронная коллекция текстов, которая формируется на основе различных аспектов. Различают самые разнообразные классификации корпусов: по видам базы данных (устная, письменная), языку текстов (русский, немецкий, турецкий, параллелизму переводов текстов (двуязычный, трехъязычный), стилю (художественный, научный…), доступности базы (открытая, закрытая), географическому фактору (относящийся к одному государству/…) и др.

Таким образом, под корпусом понимается совокупность устных и письменных текстов, хранящихся в базе данных компьютера. Также в корпусе представлено подробное описание точного времени написания собранных материалов, стиль текстов. Пользователь может ссылаться на тексты, представленные в художественном, научном, официальном или публицистическом стилях, в зависимости от целей и задач. Это особенно полезно при обучении языку. Ввиду того, что корпус представляет собой систематизированную библиотеку с очень широким охватом и высокой степенью достоверности, учитель в школе может получить действенную поддержку в оперативном составлении заданий учащимся на

этапе закрепления знаний. Он прост в использовании, значительно экономит время [Абдурахмонова Н., Урдишев К., 2019].

Современная западная методическая литература уже активно использует примеры и описания из параллельных корпусов при обучении языку. Примеры из корпуса дают студентам практический материал, с которым они столкнутся при использовании языка в реальных ситуациях межкультурной коммуникации. Нужно отметить, что основным недостатком многих учебников являются изобретенные примеры, а их описания строятся на интуиции или субъективных выводах [Кокорева А. А., 2013]. В работе с корпусом этот момент исключается совершенно, т.к. он выдает примеры из реальных текстов. В связи с этим, все чаще корпус используется для анализа и выявления недостатков существующих материалов в преподавании иностранных языков.

Поиск по корпусу позволяет пользователю найти все формы указанного слова в различных контекстах, что выгодно отличает электронную поисковую систему от других. Наглядно показывает варианты слова, какое место они занимают в словаре. Он может определять диапазон слов, денотативные и коннотативные значения искомого слова, которые могут быть спряжены. Также описывает частоту или статистику использования того и иного слова в языке писателя.

Основная часть. Направления корпусной лингвистики, в том числе проекты электронных корпусов текстов, активно развиваются и имеют значительный прикладной потенциал в методике обучения иностранным языкам и переводу, а также в компьютерной лингвистике. Вопросам обучения языку и переводу на базе параллельных корпусов уделено большое внимание в работах таких ученых как М.Барлоу, Маккенери, Бэйкера, Уилсона, Дэнилсон и Райдингс, Заннети, Аренберга, Бланка, Брауна, Черча и Гейла, Девиса, Фостера и многих других.

Использование электронного корпуса текстов на иностранном языке при переводе художественных текстов позволяет найти способ устранения ошибок, заключающихся в употреблении таких конструкций и принятии таких решений, которые не являются очевидным выбором носителя языка. К ним относятся проблемы лексико-грамматической сочетаемости, орфографические, пунктуационные трудности и трудности, связанные с выбором оптимальной грамматической конструкции [Груздев Д. Ю., 2013].

К началу 21 века в мире было создано множество двуязычных/многоязычных корпусов, к примеру: EUROPARL; CHEMNITZ GERMAN-ENGLISH TRANSLATION CORPUS; KACENKA; Lancaster's ITU, Johansson and Hofland; HKUST; Erjavec et al.; "Agenda 21"; Kraaij; Resnik et al., Tadic и др.

В настоящее время действуют или находятся на стадии разработки корпуса многих языков мира: Национальный корпус русского языка [НКРЯ, 2003] - представительная коллекция текстов на русском языке общим объемом более 2 млрд слов, оснащенная лингвистической разметкой и инструментами поиска [https://ruscorpora.ru]. В состав НКРЯ вошли 38 корпусов. Некоторые из корпусов включают подкорпусы, в частности, Параллельный корпус содержит 22 параллельных подкорпуса общим объемом 165 млн слов.

TUD - Национальный корпус турецкого языка [https://www.tnc.org.tr/tr/] версии 3.0 представляет собой исчерпывающий, сбалансированный и представительный корпус-

справочник общего назначения объемом 50 миллионов слов, охватывающий период в 24 года (1990–2013 гг.)

С 2012 года разрабатывается Алматинский корпус казахского языка [АККЯ] – один из возможных вариантов Национального корпуса казахского языка [http://web-corpora.net/KazakhCorpus]. С 2019 года разрабатывается Национальный корпус таджикского языка (НКТЯ). Корпус создан в результате совместной работы таджикских и российских специалистов. На настоящем ресурсе [https://tajik-corpus.org/] находится письменный корпус литературного таджикского языка объемом 40,8 млн словоупотреблений. Каждая разобранная словоформа включает грамматическую информацию и перевод на русский и английский язык.

Проводятся теоретические и практические исследования по формированию веб-корпуса узбекского языка: Ташкентский государственный университет узбекского языка и литературы, Национальный университет Узбекистана, Ташкентский университет информационных технологий, филиал Самаркандского государственного университета ТАТУ. В 2020-2021 годах в Ташкентском государственном университете узбекского языка и литературы реализован прикладной проект «Создание учебного корпуса узбекского языка».

В 2021 году под руководством профессора Национального университета - Абдурахмоновой Н.З. запущен проект по созданию Электронного корпуса узбекского языка [https://uzbekcorpus.uz]. Современное состояние «Узбекского электронного корпуса» представляет собой комплекс лингвистических словарей, веб-страниц, морфологической базы данных узбекского языка, учебной и научной литературы, официальных и художественных текстов различных жанров. Корпус содержит набор устных и письменных текстов, программное обеспечение с возможностью управления и использования данных [https://uzbekcorpus.uz].

Под руководством профессора Самаркандского государственного университета Суюн Каримова проводится прикладное исследование «Проектирование национального корпуса узбекского языка и разработка программного комплекса». Основным результатом исследования станет программный комплекс, разработанный для создания национального корпуса узбекского языка. В 2022 году программное обеспечение было протестировано и готово к использованию в рамках созданного корпуса текстов эпоса «Алпомиш».

Рассмотрим параллельный корпус как новый вид лингвистических ресурсов. Автономная часть электронного корпуса способна вместить в себя огромное количество необходимой информации. В направлении машинного перевода существуют многоязычные корпуса, специально отформатированные для параллельного сравнения, называемые систематизированными параллельными корпусами.

Спектр использования параллельного корпуса достаточно широк: в обычном и машинном переводе, в исследовании проблем теории и практики перевода, лингвистики, вычислительной лингвистики. С начала 1980-х годов корпуса перевода использовались в вычислительной лингвистике для машинного перевода, а также для извлечения терминов, определения значения слов и др. [Абдурахмонова Н., 2016]. Вопросы построения, состава и возможностей параллельного корпуса исследуются в работах Д.О.Добровольского, Ю.Тао, В.Захарова, А.А.Кокоревой, Е.П.Сосниной.

Таким образом, параллельный корпус представляет собой коллекцию текстов-оригиналов вместе с их переводами на другой язык или языки. Параллельный корпус рассматривается как база для проведения эмпирических исследований в лексикографии и теории перевода, и как один из элементов в методике обучения иностранному языку. В переводоведении и контрастивной лингвистике параллельные корпуса занимают центральное место. Многие параллельные корпуса доступны через простые в использовании конкордансы, что значительно облегчает изучение межъязыковых явлений.

Далее опишем сам процесс создания электронного параллельного корпуса, состоящий из переводов на русский язык художественных произведений узбекских писателей. Разработка параллельного корпуса на основе узбеко-русского языковой пары, наряду с возможностью проведения на его основе исследований самых разных направлений, представляет собой новый научный метод исследования, осуществляемый непосредственно на представительном материале художественных переводов с узбекского языка на другие языки мира (метод может быть применен как для прямого, так и для обратного направлений языковых пар). Такой подход к решению переводческих проблем называется корпусным.

Параллельный корпус разрабатывается в целях создания электронной платформы по исследованию перевода и обучению переводу. В связи с этим общий алгоритм создания параллельного корпуса будет следующим: определение содержания коллекции → сбор оригинальных текстов и их переводов → установление программного инструмента → статистика корпусных данных → работа с данными корпуса. Создание корпуса тесно связано с его назначением, поэтому на основе корпуса будет возможным проведение лингвистического анализа художественного перевода произведений. Таким образом решается задача – как нужно структурировать параллельный корпус, предназначенный для анализа переводов, для создания переводных словарей и используемый в создании новых переводов. В связи с этим приоритетной задачей является определение необходимых параметров и характеристик, какими следует наделить корпус электронных текстов, позволяющие улучшить качество перевода художественного дискурса на русский язык.

Для этой цели мы предприняли попытку проанализировать опыт составителей национальных и других представительных корпусов текстов различных языков с тем, чтобы выработать методические рекомендации по составлению и использованию параллельных корпусов. Особое внимание к процедуре поиска информации в корпусе объясняется тем, что для ее извлечения из такого корпуса необходимо специальное программное обеспечение, т.н. корпус-менеджеры. На основе нескольких составленных корпусов будут выработаны рекомендации по поиску информации в них с помощью таких программ [Груздев Д. Ю., 2016]. Для разработки параллельного корпуса в качестве образца выбраны 3 действующих электронных корпуса: Национальный корпус русского языка, Национальный корпус японского языка и Национальный корпус турецкого языка.

Параллельный корпус предназначен для анализа, исследования переводов, поэтому, как упоминалось выше, создаётся целый комплекс сложных задач.

В процессе создания параллельного корпуса будет определена общая структура параллельного корпуса, дана характеристика отобранных текстов, описаны метаданные и поиск словоизменительных парадигм, таким образом, решается задача создания электронного корпуса текстов.

Во-вторых, создаётся программа - конкорданс (программное обеспечение) параллельных текстов. Для согласования и создания конкордансов параллельных текстов стоит задача выбора, установки и адаптации одного из вариантов программы-конкордансер ParaConc для исследуемой пары языков (узбекский-русский). Данная задача нуждается в дополнительном анализе для принятия решения о целесообразности использования веб-приложений вместо программ-конкордансеров, а также использования Интернета вместо корпуса текста, что в определенной степени позволит освободить переводчиков от необходимости составлять тематические корпусы [Груздев Д. Ю., 2016].

Разрабатывается сайт корпуса, через веб-интерфейс которого реализуется поиск морфологических форм, поиск с добавлением элементов метаданных и поиск стилистических единиц (н-р, реалии, экзотизмы, языковая картина писателя).

Далее также намечается разработка программы предварительной обработки текста, автоматическая лемматизация слов в текстах в обоих (русскоязычной и узбекоязычной) частях корпуса. Поиск в корпусе осуществляется посредством запроса по буквам или в виде набора символов (выбора параметров из меню) разной степени сложности. Электронный формат корпуса обусловливает обязательное наличие специальной разметки - лингвистической или метаязыковой информация о выбранных единицах корпуса: текст, предложение, текстоформа, звук.

Создание корпуса начинается с процедуры обработки текстов. Проведение предварительной обработки состоит из следующих этапов: 1) заполнение текстовых таблиц. Импорт данных таблицы в параллельный корпус; 2) выравнивание. На первом этапе выравнивание корпуса выполняется автоматически с использованием Paraconc. На втором этапе ошибки выравнивания исправляются в ручном режиме. Затем производится загрузка, которая осуществляется с помощью следующих действий: file – export – export corpus files. Выравненные тексты загружаются автоматически на платформе Paraconc. Для того чтобы файл открывался, загруженный файл должен быть сохранён вместе с микрософтом Paraconc.

Особое внимание к процедуре поиска информации в корпусе объясняется тем, что для ее извлечения из такого корпуса необходимо специальное программное обеспечение, т.н. корпус-менеджеры. На основе составленного корпуса будут выработаны рекомендации по поиску информации в них с помощью таких программ.

На основе анализа возможностей других электронных ресурсов, в основе которых лежат корпусы текстов, например, программы накопительного перевода, был выявлен ряд присущих им потенциальных недостатков, а именно – 1) использование переводных текстов и 2) навязывание переводчику единственно возможного варианта перевода. Ввиду отсутствия таких ограничений у тематических корпусов текстов мы предполагаем, что они дают возможность придать процессу перевода творческий характер. Они также являются более надежным источником фоновой информации. Все это в комплексе должно способствовать повышению качества переводимых на русский язык художественных текстов в целом. Для проверки данного предположения требуется проведение лингвистического эксперимента.

Поиск в корпусе осуществляется на основе языка регулярных выражений (regular expressions). Морфологическая нормализация для узбекского языка не производится. Так как целью создания корпуса является исследование переводов, то основной режим поиска

- поиск на платформе Paraconc. Кроме того, создаётся платформа поиска в Интернете. Там будет реализован поиск по лексическим единицам/словоформам с добавлением элементов метаданных (retrieval intersection). Например, мы ищем контексты для слова «мусофирлар» в текстах, по автору и конкретному переводчику.

Генерация частотного словника. Материал корпуса – художественный текст и, естественно, в оригиналах и переводах много слов и выражений, отражающих идиостиль писателя. Их перевод является приоритетным, поэтому создание словника или базы данных языка писателя - одна из функций параллельного корпуса. На первом этапе мы определяем частотность слов из текстов на узбекском языке и их переводов в текстах на русском языке и выравниваем их. Выровненные слова и обороты составляют новый текст (в формате txt). На втором этапе мы импортируем выровненные тексты по строчке с помощью инструментов (eclipse и java) в базу данных (database) и выполняем генерацию частотного словника по данным корпуса [Захаров В., Yuan Tao., 2015]. Словник составит основу Толково-переводного словаря произведений писателя.

Далее разрабатывается сайт корпуса, через веб-интерфейс которого реализуются поиск морфологических форм, поиск с добавлением элементов метаданных и поиск частотных слов [Захаров В., Yuan Tao., 2015]. Метаданные корпуса включают следующую информацию: язык, тип, автор, переводчик, время издания и название. В метаданных мы указываем два языка – узбекский и русский, три типа: лингвистика, литературоведение и переводоведение, два вида времени издания - год публикации оригинала и перевода, название на двух языках – узбекском и русском [Захаров В., Yuan Tao., 2015]. Таким образом, электронный корпус даст возможность изучать стандартные лексические и синтаксические соответствия между двумя языками, анализировать путем сопоставления различных переводов принципы, лежащие в основе той или иной переводческой стратегии [Маник С. А., 2019].

Параллельный корпус (ПК) будет состоять из двух частей:

1. Переводной корпус текстов на узбекском языке (например, все произведения А.Кадыри) и их переводов на русский язык и

2. Сопоставимый корпус (СК) из тематических текстов на русском языке. Тематические тексты представят художественные произведения русскоязычных писателей Узбекистана, например, таких как - С. Бородин, М. Шевердин, В. Ян, Я. Ильясов, А. Удалов, Б. Пармузин.

Совпадение стиля текстов параллельного и сравнительного корпусов обеспечивает соспоставимость особенностей переводного и оригинального русского языка.

Переводной русский язык сопоставляется не только с оригинальным узбекским языком в переводном корпусе, но и соригинальным русским языком в сравнительном корпусе. Оригинал и перевод каждого произведения переводного корпуса создают свой подкорпус, и оригинал каждого произведения сравнительного корпуса создает соответствующий подкорпус. Таким образом, сопоставляются и количествено анализируются не только сами корпуса, но и пары подкорпусов. Отметим, что учет баланса количества слов между подкорпусами повышает надежность исследования.

Структуру корпуса узбекского и русского языков можно представить в виде следующей схемы:

Как отмечалось выше, в рамках параллельного корпуса возможно проведение корпусных исследований, то есть исследование языка с помощью корпусных методов. Данный подход основывается на использовании электронных корпусов текстов - огромных массивов естественных текстов на иностранном языке, собранных на магнитном носителе и должным образом упорядоченных и размеченных для более быстрого поиска интересующей лингвистической информации. Корпусный метод основан на статистической обработке языка, т.е. в корпусных исследованиях используются количественные методы. Параллельные корпуса выявляют закономерности в тексте оригинала и текста перевода, так как одинаковая коммуникативная направленность в двух разноязычных текстах является объектом сопоставительных корпусов.

Следование корпусному подходу при решении переводческих проблем представляет возможности повышения качества перевода. В настоящего времени исследователи корпусной лингвистики решают задачу составления корпусов текстов и выработки методологии работы с ними для проведения дальнейших лингвистических исследований и использования данного ресурса в качестве одного из способов решения проблемы улучшения перевода. Исследования в русле корпусного подхода к решению переводческих проблем стали привлекать внимание широкого круга ученых лишь в недалеком прошлом, к примеру международная конференция по данному вопросу состоялась в России в 2004 году. Хотя методология проведения микроисследования с помощью корпуса текстов в процессе перевода не нова: неэлектронные корпуса текстов применялись переводчиками в работе и раньше, но очень редко, так как их громоздкость и малая репрезентативность вызывали критику и не позволяли оперативно использовать их в реальной переводческой деятельности [Владимов Н. В., 2005].

Также вторым назначением создаваемой платформы является использование корпуса для обучения переводу. В корпусной методике обучения языку и переводу задействован грамматико-переводной метод. Обучение языку осуществляется в параллельных электронных корпусах текстов с использованием программ-конкордансов параллельных текстов.

Известно, что на практике перевод ориентируется на возможность постредактирования, также исследователь сравнивает и оценивает различные стратегии и интерпретации текста. Много времени тратится на обращение к словарям. Такой расход времени при использовании лингвистических компьютерных технологий и электронных параллельных корпусов ощутимо сокращается, также при изучении приемов и способов перевода корпус предоставляет образцы профессионального перевода.

Параллельный корпус художественных переводов с узбекского языка на русский позволит решать различные лингвистические, переводческие и образовательные задачи. Например,

1. Исследование принципов перевода на русский язык на основе анализа перевода частей речи, грамматических категорий (залог, переходность глагола), основе анализа перевода предложений;

2. Определение основной переводной единицы на основе корпуса;

3. Практическое исследование перевода на основе корпуса (перевод художественного дискурса, правила трансформации простых и сложных предложений).

На платформе корпуса возможно осуществление оценки переводных текстов, исследование природы и универсальности переводного языка, изучение идиостиля писателя. Вследствие чего в рамках проекта будет возможной разработка толково-переводной словарь художественного языка писателя.

На самом деле создание корпусов имеет большое значение, прежде всего, для лексикографии. Известно, что время находит свое отражение в словаре языка, которое постоянно «растет». Зеркалом, которое отчетливо показывает богатство языка, является его словарь. Объем словарного содержания считается фактором, повышающим его ценность.

В преподавании языка такие корпуса также являются богатым источником материалов. Кроме того, параллельные корпуса служат обучающими данными для статистических систем машинного перевода.

В параллельном корпусе собраны тексты оригинала и его переводов. В переводоведении основное внимание уделяется выявлению особенностей, отличающих переводы от оригинальных текстов. Эти изменения могут быть индивидуальными для конкретной переводческой задачи или языковой пары, но также в соответствии с широкими лингвистическими особенностями перевода они могут указывать на общие черты, которые отличают текст перевода. Очевидно, что эти исследования являют собой наглядный эмпирический метод определения особенностей корпусных переводов, и с 1990-х годов применены многими учеными, к примеру, Beyker (1993; 1996), Johansson и Ebeling (1996), Hansen (2003); Teich (2003); Mauranen и Kujamäki (2004) и Hansen-Schirra, Neumann и Steiner (2012). Кроме того, параллельные корпуса используются в качестве справочного материала при обучении переводу и в сфере профессионального перевода, поскольку они обеспечивают быстрый и интерактивный доступ к переводческим решениям (например, к памяти переводов).

С точки зрения переводоведения важной задачей параллельного корпуса является обнаружение переводческих изменений. Как в переводоведении, так и в контрастивной лингвистике в последнее время многоязычные корпуса стали использоваться для изучения переводческих явлений, т.е. переводческих изменений или особенностей перевода, а также контрастивных различий между ними.

Мировая лингвистическая ситуация явилась побудительной причиной в формировании приоритетной задачи узбекского языкознания – это создание параллельных корпусов. Процесс автоматической нормализации текстов связан с морфологическим анализом, при котором используются методы токенизации, лемматизации и стемминга.

Токенизация – разделение речевых единиц естественного языка в соответствии с грамматическим значением, определение формы слов. Лемматизация – приведение словоформ к словарной (начальной) форме, т.е. в соответствии с их представлением в словаре, определяется основа и окончание, таким образом приводится к форме морфем. Стемминг – определение корня слова.

Токенизация	Лемматизация	Стемминг
лошадку	лошадк - а	лошадь
женского	женск - ий	жен

При использовании параллельных текстов сначала необходимо идентифицировать текстовые сегменты (фразы или предложения), а также адаптировать и скоординировать тип и жанр параллельных текстов. Алгоритм данной процедуры таков:

1 шаг. В целях лингвистического обеспечения параллельных корпусов сбор и сортировка коллекции англо-узбекских, узбеко-английских параллельных текстов.

2 шаг. Анализ репрезентативности текстов.

3 шаг. Сегментация слов, сочетаний и предложений в соответствии с уровнем сопоставления.

4 шаг. Создание переводных словарей, соответствующих текстам.

5 шаг. По завершению процесса сегментации сохранить все параллельные тексты в памяти перевода.

При создании автоматического анализатора текста сначала создается электронный морфологический словарь. По итогам морфологического анализа выдаются статистические данные о частеречной принадлежности слов. Так создается большая логико-семантическая база лексем. Морфологический анализ также важен для синтаксического анализа. Параллельный корпус служит связующим звеном для выявления глубинной семантики лексических единиц языков [Абдурахманова Н., Тулиев У., 2018].

Сравнительный корпус содержит важные для параллельного корпуса ключевые слова. Известно, что ключевые слова используются часто и передают основную смысловую нагрузку. Для выделенной части текста в отдельную группу собираются единицы с самостоятельным значением (т.е. знаменательные части речи отделяются от служебных), затем создается таблица единиц, классифицированных как ключевые слова. Выделение ключевых слов в тексте параллельные корпуса намного ускоряет работу.

Заключение. Параллельный корпус считается важной реальностью для широко распространённого межкультурного общения современной эпохи. Посредством параллельных корпусов возможно выявление универсалий для сопоставляемых языковых сред и культур, а также специфических ментальных характеристик языков, реалий и лакунарных единиц. Корпус параллельных текстов служит также для развития автоматического перевода, обеспечивает формирование компьютерной лексикографии. С

помощью корпуса параллельных текстов разрабатываются программы согласования и возможно создание различных специализированных словарей.

В настоящее время корпус представляет собой действенный инструмент, экономящий время и усилия. Проводятся теоретические и практические исследования по формированию … «корпусного обучения языку, parsing на основе теории зависимостей, технологии FST в морфологическом анализе корпуса, авторского корпуса, программного обеспечения и лингвистического обучения в создании национального корпуса, морфологического и семантического анализаторов корпуса, по созданию параллельных корпусов машинного перевода на основе нейротехнологии, учебного корпуса узбекского языка» [Абдурахманова Н., Тулиев У., 2018]. Применение параллельного корпуса эффективно не только в языкознании, но также и в переводоведении, двуязычной лексикографии и в областях, где необходимо сравнивать языки. Одним из преимуществ электронного корпуса является возможность посредством их наблюдать изменения слов, появление историзмов, неологизмов, расширение и сужение значения, появление в языке новых фразеологизмов. В языковых корпусах, представляющих собой электронную базу, постоянно вводится и размещается новая информация, в них находит свое отражение процесс периодического обогащения языка. Это радикально сокращает объем человеческого труда, так как требуется много усилий и средств для включения в словари печатной формы все те изменения, происходящие со временем в языке. А корпус дает возможность решить эту проблему в короткие сроки и без материальных затрат.

Ожидаемым результатом применения параллельных корпусов является полная автоматизация текста, который может быть использован не только для ручного (человеческого) перевода, но и для машинного. В настоящее время мы находимся в процессе создания технологии машинного перевода, использующей статистический способ. Создание памяти переводов, являющейся лингвистической основой для статистических технологий перевода (таких как Google), способствует развитию нейронной технологии машинного перевода. Еще одним важным аспектом параллельных корпусов является то, что они препятствуют появлению переводческих ошибок. До сих пор наблюдается в Интернете проблема неверного перевода, когда появляются слова, не имеющие никакого отношения к переводимому текста. Основная цель создания параллельного корпуса – разработка эффективных и адекватных методов перевода. Посредством параллельного корпуса достигается улучшение качества перевода.

Параллельный корпус представляет собой набор электронных текстов на языке оригинала (узбекский язык) и его перевод (русский язык). Параллельный корпус сохраняет оригинальную последовательность текста и служит:

- во-первых, для сопоставления различных языковых структур, фраз и слов в определенном контексте, т.е. для сравнительного анализа структур двух языков;
- во-вторых, в области переводов: для поиска эквивалентов оригинала в других языках;
- в-третьих, при обучении движков (систем) машинного перевода при изучении языка и при составлении словарей.

В статье рассмотрен процесс разработки параллельного корпуса узбекского и русского языков и возможных на его основе направлений исследования. Этот параллельный и компаративный корпус узбекского и русского языков предназначен для анализа переводов. Основными задачами проекта являются разработка дизайна корпуса,

формирование коллекции текстов, проведение предварительной обработки, создание поисковой системы и метаданных. В результате исследования создается параллельный корпусный веб-сервис. Разрабатывается программа-конкордансер (программное обеспечение) параллельных текстов. В рамках апробации и функционирования параллельного корпуса возможно издание толково-переводной словарь художественного языка писателя, а также новых переводов на русский язык художественных произведений.

Продуктом, результатом разработки является веб-сервис «Параллельный корпус» - идентифицируемая уникальным URL-адресом программная система со стандартизированными интерфейсами. Отметим, что данный веб-сервис является продуктом совместной научно-исследовательской деятельности лингвистов и программистов.

Особо актуальным будет использование веб-сервиса в качестве корпусного метода в научных исследованиях по художественному переводу. Также будет значимым применение продукта в качестве нового, эффективного инструмента обучения переводу студентов и молодых специалистов. Третьей важной областью приложения веб-сервиса будем сам образовательный процесс в вузах, посредством которого можно будет обучать иностранному языку. Таким образом, обозначаются два основных направления в наличии спроса на результаты разработки: во-первых, на основе параллельного корпуса, применив корпусный метод, станет возможным проведение научных исследований в сферах художественного перевода, сопоставительного языкознания, корпусной лингвистики, лексикографии. Во-вторых, параллельный корпус станет образовательным инструментом или новой методикой в процессе обучения иностранному языку и в преподавании теории и практики перевода.

На основе технологии создания веб-сервиса возможна разработка других оригинальных программных систем, имеющих приложение не только в лингвистике и переводоведении, но и в других направлениях общественно-гуманитарной сферы. Станет возможным открытие современных информационных центров узбекского языка в филиалах зарубежных вузов. Также отметим, что параллельный корпус художественных текстов войдет составной частью Национального корпуса узбекского языка.

Список использованной литературы:

1. Абдурахманова Н. Komputer lingvistikasi [учебник / Globe edit, 2020, 395 b.
2. Абдурахманова Н. Основы автоматического морфологического анализа для машинного перевода. - Известия Киргизского государственного технического университета. 2016; 2 [38]:12-7.
3. Abdurakhmonova, N., & Tuliyev, U. (2018). Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language. *Literature, Education*, *3*, 68.
4. Abdurakhmonova, N., & Urdishev, K. (2019). Corpus based teaching Uzbek as a foreign language. *Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL)*, *6*(1-2019), 131-7.
5. Владимов Н. В. Корпусный подход к решению переводческих проблем: на материале письменных переводов с русского языка на английский: диссертация ... к.ф.н.: - Москва, 2005. - 198 с.

6. Груздев Д. Ю. Электронный корпус текстов как эффективный инструмент переводчика: диссертация ... к.ф.н.: Москва. 2013.
7. Захаров В., Yuan Tao. Создание и использование параллельного корпуса русского и китайского языков. - 2015.
8. Кокорева А. А. Корпус параллельных текстов в обучении иностранному языку // Вестник ТГУ. 2013. №2 (118).
9. Маник С. А. Параллельный корпус в практике перевода общественно-политических текстов (с английского на русский) // СИСП. 2019. №4.
10. Многоязычный корпус: лингвистические требования и технические перспективы. Семинар. Великобритания. Ланкастер - 2003.
11. Соснина Е.П. Параллельные корпусы в обучении языку и переводу. http://spr.fld.mrsu.ru/2015/09/e-p-sosnina-parallelnye-korpusy-v-obuchenii-yazyku-i-perevodu/
12. Указ Президента Республики Узбекистан «О мерах по кардинальному повышению роли и авторитета узбекского языка в качестве государственного языка», mУП - №5850. 2019.
13. https://uzbekcorpus.uz

MATNNING DISKURSIV TAHLILI

Mirvaliyeva Dildora Anvarovna
Mirzo Ulug‘bek nomidagi O‘zbekiston
Milliy universiteti stajyor o‘qituvchisi.
E-mail: dildoramirvalieva07@gmail.com

Annotatsiya: Mazkur maqola diskurs tushunchasi, uning kelib chiqishi va o‘rganishga bo‘lgan turlicha yondashuvlar hamda matnni diskursiv tahlil qilishning ahamiyati bayonotini o‘z ichiga oladi.

Kalit so’zlar. Diskurs, intuitiv, tahlil, yondashuv, matn, muloqot, fraza, adresat.

Abstract: This article includes a statement of the concept of discourse, its origins and different approaches to its study, and the importance of discursive text analysis.

Keywords: Discourse, intuition, analysis, approach, text, communication, phrase, addressee.

Аннотация: В данной статье изложено понятие дискурса, его истоки и различные подходы к его изучению, а также важность дискурсивного анализа текста.

Ключевые слова: Дискурс, интуиция, анализ, подход, текст, общение, фраза, адресат.

Zamonaviy tilshunoslikda hozirgi kunga qadar diskurs masalasining turli ko‘rinishlari (siyosiy, ijtimoiy, pedagogik, diniy, didaktik, xususiy, ommaviy) o‘rganilmoqda va ko‘pchilik tilshunoslar diskursni nutqiy faoliyat jarayoni hamda so‘zlashish uslubi kabi mazmunlarda talqin etishadi.

Diskurs- hissiy, bevosita, intuitiv, ya’ni muhokama orqali vosita bilan hosil qilinadigan mantiqiy dalildir. V.G. Borbotko: “Diskurs ham matndir, lekin u tilning kommunikativ birliklaridan - jumlalardan va ularning uzluksiz ichki semantik aloqada bo‘lgan kattaroq birliklarga birlashmasidan iborat bo‘lib, uni integral shakllanish sifatida qabul qilish imkonini beradi” degan ta’rifni beradi. “Haqiqiy vaziyatlarda so‘zlovchilarning tildan haqiqiy foydalanishini o‘rganish” , deb yozgan T. A. van Deyk “Diskurs tahlili qo‘llanmasi” da. Matnning tuzilish xususiyatlarini qamrab olishi va konteks tahlili jarayoni yanada ochib berishini alohida ta’kidlab o‘tadi.

*Diskurs [*Temirova F., 2022: 272*] – matndan ko’ra kengroq tushuncha bo‘lib, bir vaqtning o‘zida ham tilga xos jarayondan, hamda shu jarayonning natijasi bo‘lgan matndan iborat bo‘ladi. Diskurs so‘zi zamirida nutqning barcha kognitiv-kommunikativ funksiyalarini tushunish yotadi. Izlanishlardan ko‘rinib turibdiki, diskurs ham jarayon, ham matndir. Matnning turg‘un, tayyor mahsulot, diskursning esa kechayotgan nutqiy muloqot jarayoni sifatida talqin qilinishi ularning keskin farqlanishiga sabab bo‘ladi. Lingvistik adabiyotlarda “diskurs” atamasining qat’iy bir ma’nosi yo‘q, u ifodalaydigan hodisalar diapazoni juda keng, ya’ni “matnning qismi” dan yaxlit “nutq”gacha bo‘lgan hodisalarni ifodalash uchun ishlatiladi*[Abdurakhmonova, N. 2019,2021,2022].

Tilni pragmatik tadqiq etishga bag‘ishlangan ishlarda [Миловидов В., 1998: 39] diskurs atamasi sakkiz xil: so‘z muqobili, frazalardan o‘lchami bo‘yicha ortadigan birlik, fikrning

adresatga ta’siri, suhbat, so‘zlovchi pozitsiyasi, lisoniy birliklardan foydalanish, fikrning cheklangan turi va matn hosil bo‘lish shartlarini tadqiq etishga mo‘ljallangan nazariy qurilmalar ma’nolarida qo‘llangan.

Diskurs tahlili - bu kontekstda tildan ma’no olish uchun ishlatiladigan sifatli tahlil usuli hisoblanadi. Diskursiv tahlil odatda o‘zaro bog‘liq bo‘lgan ikkita usulda aniqlanadi [Austin A., 2023: 134]. Birinchidan, u real muloqotning lingvistik hodisalarini tadqiq qiladi. Ikkinchidan, u tilning shaklini emas, balki uning asosiy funksiyalarini hisobga oladi. Bu ikki jihat ikki xil kitobda: Maykl Stubbs o‘zining “Diskurs tahlili” asarida tahlilni lingvistik pragmatikaga havola qilsa, Jon Braun “chiziqlar orasidagi” tilni o‘rganishga harakat qiladi.

Shuning bilan birga, o‘zbek tilshunosligida ham diskurs borasida lingvistik tadqiqotlar tilshunos Sh.Safarov tomonidan amalga oshirilgan. U matn va diskurs muammolari to‘g‘risida quyidagilarni bayon etadi: “Agarda matn va diskursning har ikkalasi ham inson lisoniy faoliyatining natijasi bo‘lsa, ularni faqatgina zohiriy – formal ko‘rsatkichga asosan “og‘zaki” va “yozma” sifatlari bilan farqlash imkoniga gumonim bor”.

Diskursiv [Taxirov Z., 2021:214] (mantiqda mulohazaga asoslangan, mantiqiy qismlar ketma-ketl qatori) tahlil ma’noni aniqlash uchun korpus yoki ma’lumotlar to‘plamida taqdim etilgan tildan foydalanadi. Ushbu ma’lumotlar to‘plami intervyular yoki guruh muhokamasi stenogrammalarini o‘z ichiga olishib, nutqni tahlil qilishning ba’zi shakllari tilning o‘ziga xos xususiyatlariga (masalan, tovushlar yoki grammatikaga), boshqa shakllar bu tilning maqsadlariga erishish uchun qanday ishlatilishiga e’tibor beradi. Wodak va Krzyzanowski ta’kidlaganidek: “Diskuriv tahlili muammoga yo‘naltirilgan ijtimoiy tadqiqotlar uchun umumiy asos yaratadi”.

Diskurs tahlilining maqsadi tilning funksiyalarini va turli kontekstlarda ma’no qanday tuzilganligini, nutqning ijtimoiy, madaniy, siyosiy va tarixiy sharoitlarini o‘z ichiga olganligini o‘rganishdir [Kamolova A., 2024:191].

Matnni diskurs tahlil qilishda quyidagi bosqichlar muhim ahamiyatga ega:

-bir matnning diskurs tahlili uchun birinchi bosqich mazmun va kontekstni tushunish bo‘lib, bunda matnning qanday maqsadda yozilganligini, qaysi muhitda yoki texnologiyada yaratilganligini aniqlash, mazmunning asosiy tushunchalar va matndagi asosiy mavzuni anglash diskurs tahlili uchun boshlang‘ich nuqta vazifasini bajaradi.

-diskursiv tahlil jarayonida matndagi muhim so‘zlar, atamalarning tavsifi va ularga oid ta’riflar ko‘zdan kechirilib, ba’zi so‘zlar yoki atamalar belgilangan ma’noni o‘zgartirish uchun ishlatilishi mumkinligini o‘rganish tahlilning muhim qismi sanaladi.

-matnning muhim qismlarini aniqlash, ularning ta’kidlanishi yoki tavsifi diskurs tahlida muallifning maqsadi, fikrlaridagi tashqi va ichki bog‘liqlar tahlil qilinishi hamda bu tahlil matndagi fikr, xulosa, mazmunning ustuvliklarini aniqlashni o‘z ichiga oladi.

-matndagi fikr va nazariyalarni o‘rganish va baholash muallifning matnda qanday muammo yoki tanqidiy qarashlarga e’tibor qilganligini, fikrlarini isbotlash usullarini hamda matndagi muhim mantiqiy o‘zgarishlarni tushunishni jamlaydi.

-matndagi natijalarni ko‘rsatish va baholashda diskurs tahlilining natijalari, muallifning maqsadi, mazmun-mohiyati hamda matnni o‘rganuvchilar uchun qanday ma’noni qamrab oladiganligi baholanadi.

Diskurs tahlili ko‘p qadamli jarayon bo‘lib, matnni tushunish, o‘rganish va tahlil qilishning ko‘p tomonlama o‘zgarishlarini o‘z ichiga oladi[Abdurakhmonova, N. 2019,2021,2022].

Faoliyat doirasini ko‘rib chiqadigan bo‘lsak, qonunchilik, ish yuritishga (ishga aloqador yozishmalar, huquqiy munosabatlar sohasi) oid hujjatlarning diskurs tadqiqoti ma’muriy faoliyat sohasida tahlil va tushunchalar bilan bog‘liq bo‘lib, ish yurituvchi organlar tomonidan ishlatiladigan hujjatlarni o‘rganish, kommunikatsiya jarayonlarini tahlil qilish va ish yuritish protsessidagi muammolarni aniqlash bilan bog‘liq[Abdurakhmonova, N. 2019,2021,2022].

Shuningdek, yozma nutqning rasmiy uslubida, asosan, quyidagi munosabatlar doirasidagi hujjatlar tuziladi:

1. Huquqiy munosabatlarga oid: qonun, fuqarolik va jinoyat aktlari, nizom, shartnoma va boshqalar.

2. Idoraviy-ma’muriy shaklga oid: dalolatnoma, buyruq va farmoyishlar, turli ish qog‘ozlari (ariza, tavsiyanoma, tilxat, ma’lumotnoma kabi).

3. Diplomatik munosabatlarga doir: bayonot, nota, bitim, memorandum va boshqalar.

Ish yuritishga oid hujjatlarning diskurs tadqiqida mamuriy hujjatlarning matnlari tahlil qilinadi, ularning ma’noni o‘rganish, muammo yuzaga kelishi, so‘zlar va ifodalar orqali munozaralarda o‘tishi kabi asosiy choralarga e’tibor qaratiladi. Tadqiqot maqsadi hujjatlarning tuzilishini, ularning kommunikativ vazifalarini va muhitlarini, hujjatlarning o‘zgarishlarga olib kelishi, protokollar, qonunchilik va boshqaruvni ta’minlash jarayonlarini tushunishdir.

Tahlil natijalari ish yuritish tizimi va uning faoliyati bilan bog‘liq qarorlarni olish, boshqaruv organlari va ishlab chiqarish tashkilotlari uchun qarorlarni o‘rganishda yordam berishi mumkin. Bundan tashqari, bu tadqiqot o‘rganilgan ma’lumotlar asosida boshqaruv tizimini yanada mustahkamlash va uning faoliyatini mukammallashtirish uchun strategiyalarni ishlab chiqishga ham yordam berishi mumkin.

Foydalanilgan adabiyotlar ro‘yxati:

1. A.Austin. Discursive analysis: concept and role in modern linguistics.2024
2. В. Миловидов. Текст, контекст, интертекст: Введение в проблему сравнительного литературоведения. – Тверь: ТвГУ, 1998, с. 39.
3. Z. Taxirov. O‘zbek tilining amaliy stilistikasi [Matn]: darslik. - Toshkent: Tafakkur avlodi, 2021.
4. Abdurakhmonova, N., Tuliyev, U., Ismailov, A., & Abduvahobo, G. (2022). Uzbek electronic corpus as a tool for linguistic analysis. In *Компьютерная обработка тюркских языков. TURKLANG 2022* (pp. 231-240).
5. Abduraxmonova, N. Z. Q., & Urazaliyeva, M. Y. (2022). O ‘zbek tili elektron korpusida (http://uzbekcorpus. uz/) og ‘zaki matnlar korpusini yaratishning nazariy va amaliy masalalari. *Academic research in educational sciences*, *3*(3), 644-650.
6. Mengliev, D., Barakhnin, V., & Abdurakhmonova, N. (2021). Development of intellectual web system for morph analyzing of uzbek words. *Applied Sciences*, *11*(19), 9117.
7. F. Temirova. Badiiy diskurs va diskursiv tahlil. FarDU ilmiy xabarlari**,** 2022**.**
8. M.Saidova, M.Axrorova. Zamonaviy tilshunoslikda diskurs tushunchasi va uning tahlili. Miasto Przyszłości Kielce, 2023.
9. Sh.Bobojonova. Diskursiv tahlil va uning akademik matnlarda ifodalanishi. O‘zbekiston Milliy universiteti xabarlari, 2022. [1/12].
10. O.Ablakulova. Diskurs tushunchasi va uning tadqiqi. Academic Research in Educational Sciences, 2021.

11. Abdurakhmonova, N. (2019). Dependency parsing based on Uzbek Corpus. In *of the International Conference on Language Technologies for All (LT4All)*.
12. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021, January). Uzwordnet: A lexical-semantic database for the uzbek language. In *Proceedings of the 11th Global Wordnet conference* (pp. 8-19).
13. https://muhaz.org/6-mavzu-ish-yuritish-tili-va-uslubi-hujjat-turlari-va-xususiya.html