IJTIMOIY-GUMANITAR SOHADA ILMIY-INNOVATSION TADQIQOTLAR

ILMIY METODIK JURNALI

DOI: 10.67227

ISSN 3060-5059

VOL.3 № 6 2026

ALEKSANDR SERGEYEVICH PUSHKIN ASARLARINING MUALLIFLIK PARALLEL KORPUSINI YARATISH TAMOYILLARI

Jumayeva Zarnigor Zokirovna
Buxoro davlat universiteti, izlanuvchi

Annotatsiya

Maqola Aleksandr Sergeyevich Pushkin asarlarining asl matnlari va ularning o'zbek tiliga tarjimalarini birlashtiruvchi mualliflik parallel korpusini yaratish tamoyillariga bag'ishlangan. Rus tilining milliy korpusidagi 31 ikki tilli juftlik orasida rus-o'zbek juftligi yo'qligi, O'zbekistonda esa bu juftlik uchun maxsus yozuvchi korpusi hali yaratilmagani ko'rsatilgan. XML va TEI P5 sxemasi asosidagi arxitektura, HunAlign va LF Aligner vositasida tenglashtirish, MyStem va UzMorphAnalyzer orqali morfologik belgilash tavsiflangan. Flektiv rus va agglyutinativ o'zbek tuzilishi farqini hisobga oluvchi tarjima o'zgarishlari tasnifi taklif etilgan.

Kalit so'zlar: korpus lingvistikasi, parallel korpus, mualliflik korpusi, tenglashtirish, belgilash, A.S.Pushkin, rus-o'zbek korpusi, tarjimashunoslik, metama'lumotlar, idiostil.

ПРИНЦИПЫ СОЗДАНИЯ АВТОРСКОГО ПАРАЛЛЕЛЬНОГО КОРПУСА ПРОИЗВЕДЕНИЙ

Жумаева Зарнигор Зокировна
Бухарский государственный университет, соискатель

Аннотация

Статья посвящена принципам построения авторского параллельного корпуса, который объединяет подлинные тексты Александра Сергеевича Пушкина и их переводы на узбекский язык. Показано, что среди 31 двуязычной пары Национального корпуса русского языка русско-узбекская пара отсутствует, а специализированного писательского корпуса для этой пары в Узбекистане пока нет. Описаны архитектура ресурса на основе XML и схемы TEI P5, выравнивание средствами HunAlign и LF Aligner, морфологическая разметка через MyStem и UzMorphAnalyzer. Предложена классификация переводческих преобразований, учитывающая различие флективного русского и агглютинативного узбекского строя.

Ключевые слова: корпусная лингвистика, параллельный корпус, авторский корпус, выравнивание, разметка, А.С.Пушкин, русско-узбекский корпус, переводоведение, метаданные, идиостиль.

PRINCIPLES OF CREATING AN AUTHOR PARALLEL CORPUS OF ALEXANDER SERGEYEVICH PUSHKIN WORKS

Jumayeva Zarnigor Zokirovna
Bukhara State University, Researcher

Abstract

The article presents the principles of building an author parallel corpus that combines the original texts of Alexander Sergeyevich Pushkin with their translations into Uzbek. It shows that the Russian-Uzbek pair is absent among the 31 bilingual pairs of the Russian National Corpus and that no dedicated writer corpus for this pair has yet been built in Uzbekistan. The proposed design relies on XML and the TEI P5 scheme, sentence alignment by HunAlign and LF Aligner, and morphological annotation through MyStem and UzMorphAnalyzer. A classification of translation shifts is offered that accounts for the difference between inflectional Russian and agglutinative Uzbek.

Keywords: corpus linguistics, parallel corpus, author corpus, sentence alignment, annotation, A.S.Pushkin, Russian-Uzbek corpus, translation studies, metadata, idiostyle.

В начале 1960-х годов У. Н. Фрэнсис и Г. Кучера собрали в Брауновском университете первый машиночитаемый корпус английского языка, включивший 1 014 312 словоупотреблений из 500 текстовых выборок изданий 1961 года. Эта дата считается точкой отсчёта корпусной лингвистики. Опираясь на электронные собрания текстов, языкознание получило возможность проверять гипотезы на больших данных, а не на единичных наблюдениях. Среди множества типов корпусов особое место заняли параллельные, в которых оригинал и перевод соотнесены по

предложениям. Такой ресурс служит сопоставительной грамматике, теории перевода и двуязычной лексикографии.

Цель статьи состоит в обосновании принципов построения авторского параллельного корпуса, который объединит подлинные тексты А. С. Пушкина и их переводы на узбекский язык. Под авторским понимается собрание, ограниченное произведениями одного писателя. Параллельным называют собрание, где каждому предложению оригинала поставлено в соответствие предложение перевода. Узбекские переводы А. С. Пушкина известны с конца XIX века, а основной массив создан после правительственного постановления 1937 года к столетию гибели поэта, когда «Евгения Онегина» перевёл Айбек, «Капитанскую дочку» — Абдулла Каххар, сказки — Эльбек. Накопленный переводческий материал достаточен для отбора и выравнивания. Препятствием остаётся отсутствие согласованной методики разметки для столь различных по строю языков — флективного русского и агглютинативного узбекского.

Методы и обзор литературы. Работа опирается на несколько методов. Описательный метод применяется при характеристике существующих корпусов и их параметров. Сопоставительный метод служит для сравнения русских оригиналов с узбекскими переводами на уровне предложения и слова. Методом корпусного моделирования задаётся архитектура будущего ресурса, его разметка и состав метаданных. Приёмы автоматического выравнивания текстов реализуются программой HunAlign в оболочке LF Aligner с ручной посткоррекцией. Морфологический анализ русской части ведётся анализатором MyStem, тогда как узбекская часть обрабатывается средствами UzMorphAnalyzer. Количественные данные о составе корпусов получены из открытых источников Национального корпуса русского языка.

Теоретические основания корпусной лингвистики разработаны в трудах В. П. Захарова [1; 2], А. Н. Баранова [3] и В. А. Плунгяна [4; 5]. Методику параллельных корпусов Национального корпуса русского языка наиболее полно описал Д. В. Сичинава [6; 7], а вопросы фразеологии на материале параллельных собраний рассмотрели Д. О. Добровольский и Анна А. Зализняк [8]. Общую характеристику Национального корпуса дали В. А. Плунгян, Т. И. Резникова и Д. В. Сичинава [9]. Узбекское направление компьютерной обработки текста представлено работами Н. З. Абдурахмоновой. Вопросы машинного перевода для тюркских языков и построения русско-тюркских параллельных собраний разработаны в международном проекте TurkLang.

Результаты. Построение любого корпуса начинается с уяснения того, что считать корпусом. В учебно-методическом пособии В. П. Захарова закреплено определение, прочно вошедшее в российскую традицию: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [1, с. 3]. Шесть признаков этого определения, от электронной формы до филологической выверки, задают требования к авторскому собранию пушкинских текстов. Электронная форма обеспечивается оцифровкой академического Полного собрания сочинений А. С. Пушкина в 17 томах. Структурность достигается единой схемой кодирования, а размеченность обеспечивается приписыванием грамматических и метатекстовых помет каждому слову и каждому тексту.

От одноязычного собрания текстов проектируемый авторский ресурс принципиально отличается наличием переводного компонента. Если одноязычный корпус ориентирован преимущественно на изучение лексико-грамматических, стилистических и тематических особенностей текстов в пределах одного языка, то параллельный корпус позволяет сопоставлять оригинал и перевод, выявлять переводческие соответствия, способы передачи художественных образов, синтаксических конструкций, культурно маркированной лексики и авторской стилистики. Именно поэтому для исследования произведений А. С. Пушкина и их узбекских переводов наиболее продуктивной является модель параллельного корпуса, в котором оригинальный русский текст представлен в тесной связи с его переводным вариантом.

Определение параллельного корпуса в современной корпусной лингвистике формулируется следующим образом: «Особым типом корпуса является так называемый параллельный корпус, в котором тексту сопоставлен перевод этого текста на другой язык. Между единицами оригинального и переводного текста (обычно — между предложениями) с помощью специальной процедуры устанавливается соответствие; эта процедура называется выравниванием» [8]. Данное определение подчёркивает два ключевых признака параллельного корпуса: наличие как минимум двух языковых версий одного текста и установление формализованных связей между

соответствующими единицами оригинала и перевода. В условиях исследования художественного перевода такая структура особенно важна, поскольку позволяет проследить, как трансформируются семантика, образность, ритмико-интонационная организация и стилистическая выразительность пушкинского текста при передаче на узбекский язык.

Архитектура проектируемого ресурса опирается на три взаимосвязанных уровня. Первый уровень предназначен для хранения подлинных текстов А. С. Пушкина и их узбекских переводов в формате XML по схеме TEI P5. Использование TEI P5 обеспечивает структурированное представление текста: выделение заголовков, строф, абзацев, реплик, примечаний, имен собственных, дат, жанровых и композиционных элементов. Благодаря этому корпус становится не простым электронным архивом, а научно организованным ресурсом, пригодным для филологического, лингвистического и переводоведческого анализа. На данном уровне сохраняется целостность текста, фиксируются сведения об источнике, издании, переводчике, времени публикации и редакторских особенностях.

Второй уровень связан с фиксацией соответствий между оригиналом и переводом. Для этого используются пары предложений или более крупных смысловых фрагментов в формате TMX. Данный формат позволяет представить переводческую память, в которой каждой единице исходного русского текста соответствует определённая единица узбекского перевода. Выравнивание может осуществляться на уровне предложения, фразы, стихотворной строки или абзаца в зависимости от жанровой природы текста. Например, при работе с прозаическими произведениями целесообразно использовать преимущественно предложенческий уровень, тогда как при анализе поэтических текстов важным становится сохранение строки, строфы и ритмико-композиционной структуры. Такой подход позволяет исследователю быстро находить соответствующие фрагменты, сравнивать переводческие решения и выявлять закономерности межъязыковой трансформации.

Третий уровень включает морфологическую разметку и систему метаданных. Морфологическая разметка обеспечивает возможность анализа грамматических форм, частей речи, словоизменительных моделей, синтаксических связей и частотности языковых единиц в оригинале и переводе. Это особенно важно при изучении способов передачи пушкинской лексики, эпитетов, глагольных форм, обращений, архаизмов, реалий и эмоционально-экспрессивных средств. Метаданные, в свою очередь, фиксируют информацию о произведении, жанре, дате создания, переводчике, языке перевода, источнике публикации, типе текста и степени выравнивания. Наличие таких данных делает ресурс удобным не только для чтения, но и для комплексного научного поиска.

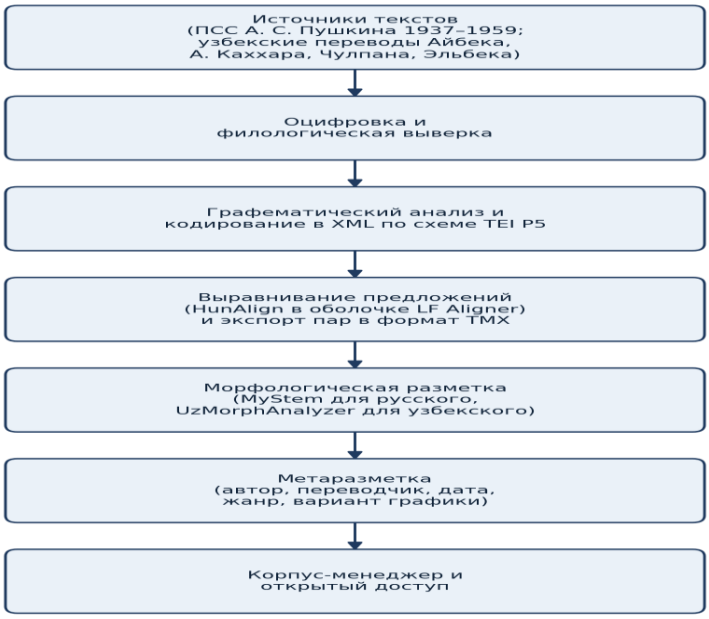

Рис. 1. Этапы создания и многоуровневая архитектура авторского параллельного русско-узбекского корпуса произведений А.С.Пушкина

Сердцевину параллельного корпуса образует выравнивание. По описанию разработчиков Национального корпуса русского языка, выравнивание понимается следующим образом: «Важный

элемент разметки параллельных корпусов — выравнивание: каждому предложению (как минимум, абзацу) на языке X соответствует предложение на языке Y. Благодаря выравниванию параллельный корпус становится полезным инструментом для нескольких категорий пользователей». Автоматическое сопоставление пар выполняет программа HunAlign в составе оболочки LF Aligner, после чего пары проходят ручную проверку. Сложность для русско-узбекской пары создаёт несовпадение порядка слов. Русское предложение «Татьяна любит Онегина» строится по схеме: подлежащее, сказуемое, дополнение, тогда как узбекский перевод «Татьяна Онегинни севади» ставит сказуемое в конец. При выравнивании на уровне слов соединительные линии пересекаются, потому что глагол севади перемещается вправо, а дополнение Онегинни занимает место перед ним. На уровне предложения соответствие остаётся однозначным, и для исследователя перевода важна именно эта пара.

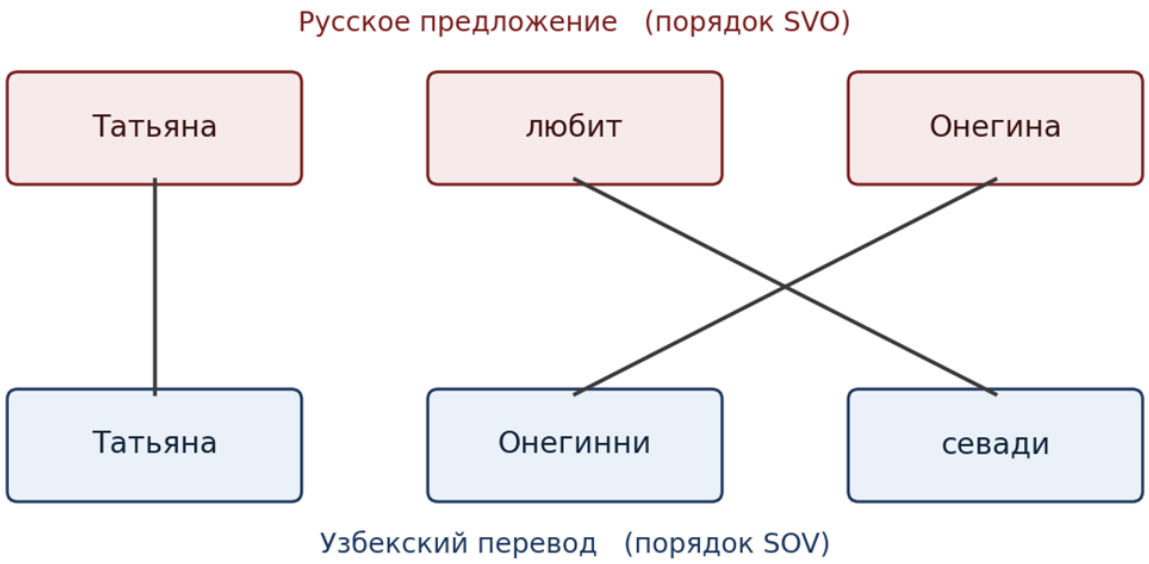

Рис. 2. Выравнивание на уровне слов при несовпадении порядка слов в русском и узбекском предложениях

Качество собрания определяется его разметкой. В очерке В. А. Плунгяна это положение высказано так: «Корпус некоторого языка — это, в первом приближении, собрание текстов на данном языке, представленное в электронной форме и снабжённое научным аппаратом. Аппарат, „встроенный“ в корпус, обычно называется „разметкой“, или „аннотацией“, корпуса; корпус тем лучше, чем полнее и совершеннее его аннотация» [4].

Сравнение проектируемого собрания с двумя действующими ресурсами сведено в таблицу 1. Англо-русская пара Национального корпуса насчитывает 1556 текстов и 52 036 252 словоупотребления. Узбекский национальный корпус развивает параллельный компонент, однако открытых количественных данных по нему пока нет. Авторский русско-узбекский корпус А. С. Пушкина отличается от обоих по охвату материала и по назначению.

Таблица 1. Сравнение параметров авторского корпуса А.С.Пушкина с действующими ресурсами

Параметр	**НКРЯ, англо-русская пара**	**Узбекский национальный корпус**	**Авторский корпус А.С.Пушкина (проект)**
Тип	двуязычный, поливариантный, общий	многокомпонентный, национальный	двуязычный, поливариантный, авторский
Объём	1556 текстов, 52.036.252 слова	разрабатывается, данные не опубликованы	около 30 произведений, по 1-3 перевода
Языки	русский и английский (31 пара в НКРЯ)	узбекский, русский, английский	русский и узбекский
Выравнивание	HunAlign и оболочка	HunAlign и собственные	HunAlign в LF

	Euclid	модули	Aligner, ручная коррекция
Морфология	MyStem	UzMorphAnalyzer	MyStem и UzMorphAnalyzer
Формат	XML, внутренний формат НКРЯ	XML	XML по схеме TEI P5 и TMX
Метаданные	автор, переводчик, дата, жанр	автор, дата, источник	автор, переводчик, дата, жанр, графика
Координация	Д.В.Сичинава	Н.З.Абдурахмонова	определяется автором проекта

Идея авторского подсобрания заложена в самой модели большого корпуса. По наблюдению В. А. Плунгяна, выбор пользователем нужного подмножества предусмотрен изначально: «Поиск возможен не только по всему корпусу, но и по определённому подмножеству текстов, выбранному пользователем: например, тексты определённого автора, определённого периода, определённого жанра и т. п. (в любых комбинациях...)» [4]. В Национальном корпусе таким способом выделены подсобрания 12 классиков, и А. С. Пушкин стоит среди них первым. Различие строя русского и узбекского языков рождает устойчивые типы соответствий.

Таблица 2. Типология переводческих преобразований при выравнивании русско-узбекских пар

Тип преобразования	**Причина**	**Пример (русский-узбекский)**	**Влияние на выравнивание**
Перестановка сказуемого	переход от SVO к SOV	Татьяна любит Онегина - Татьяна Онегинни севади	пара предложений сохраняется
Утрата рода	в узбекском нет категории рода	Она пришла, Он пришёл - У келди	пара при потере родовых сведений
Развёртывание безличной конструкции	в узбекском нет безличных форм	Мне грустно - Мен ғамгин	возможна пара 1 к 1 или 1 к 2
Свёртка предлога и падежа	флексия с предлогом против аффикса	в Петербург - Петербургга	узбекское предложение короче
Развёртывание словоформы	синтез морфем в узбекском	уйимиздагилар - те, кто в нашем доме	русский эквивалент длиннее
Метрическая перестройка	ямб против силлабики бармоқ	онегинская строфа - строфа из 14 строк	выравнивание по строфам и строкам

За методикой стоит самостоятельная область знания. По определению В. П. Захарова, её предмет очерчен строго: «Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с использованием компьютерных технологий» [1, с. 3].

Материал для пушкинского собрания обширен. Первый узбекский перевод пушкинской сказки о рыбаке и рыбке появился в газете «Туркистон вилоятининг газети» 14 марта 1899 года. Основной массив переводов создан после постановления правительства Узбекской ССР от 2 февраля 1937 года, когда «Евгения Онегина» перевёл Айбек, «Капитанскую дочку» — Абдулла Каххар, «Бориса Годунова» и «Дубровского» — Чулпан, сказки — Эльбек. В середине 1950-х годов вышло четырёхтомное собрание сочинений А. С. Пушкина на узбекском языке, разошедшееся почти мгновенно. У одного только «Евгения Онегина» исследователи насчитывают около 12 узбекских переводчиков, среди которых Миртемир, Зульфия, Аскад Мухтар, Эркин Вахидов, Абдулла Арипов, Джамал Камал.

Обсуждение. Готовый ресурс открывает несколько прикладных возможностей. Сопоставление флективного и агглютинативного строя на материале одного автора даёт типологам выверенную пару текстов с проверенным переводом. Наличие нескольких узбекских переводов одного произведения позволяет сравнивать переводческие решения разных поэтов, например передачу онегинской строфы у Айбека и у позднейших переводчиков. Двуязычные словари пушкинской поры получают эмпирическую основу, поскольку каждое словоупотребление

снабжено контекстом и переводным соответствием. Системы машинного перевода между русским и узбекским языками получают художественный параллельный материал для дообучения нейронных моделей, которого сейчас недостаёт.

Препятствия проекта связаны прежде всего с историей узбекской письменности. Узбекские переводы А. С. Пушкина выходили арабским письмом в 1920-е годы, латиницей — с 1929 по 1940 год, кириллицей — с 1940 по 1992 год и современной латиницей — после 1993 года, поэтому в метаданных каждого текста фиксируется вариант графики. Автоматическое выравнивание стихотворных произведений осложняется тем, что силлабический узбекский стих **barmoq** не совпадает с четырёхстопным ямбом подлинника, и ручная проверка строф становится обязательной.

Заключение. Создание авторского параллельного русско-узбекского корпуса произведений Александра Сергеевича Пушкина опирается на проверенную методику корпусной лингвистики и на доступный переводческий материал, накопленный за более чем сто лет. Отсутствие русско-узбекской пары среди 31 пары Национального корпуса русского языка и отсутствие писательского корпуса такого рода в Узбекистане делают проект новым по постановке задачи. Предложенная трёхуровневая архитектура, инструменты HunAlign, MyStem и UzMorphAnalyzer, а также классификация переводческих соответствий образуют рабочую основу, которую предстоит проверить на пробном корпусе объёмом не менее 200 000 словоупотреблений.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Захаров В.П. Корпусная лингвистика: учебно-методическое пособие. – СПб.: Изд-во СПбГУ, 2005. – 48 с.
2. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. – СПб.: Изд-во СПбГУ, 2020. – 234 с.
3. Баранов А.Н. Введение в прикладную лингвистику: учебное пособие. 4-е изд. – М.: ЛЕНАНД, 2017. – 368 с.
4. Плунгян В.А. Зачем мы делаем Национальный корпус русского языка? // Отечественные записки. – 2005. – № 2 (23).
5. Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. – 2008. – № 2 (16). – С. 7–20.
6. 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. – Вып. 6, № 3. – С. 194–235.
7. Сичинава Д.В. Параллельные корпуса в составе НКРЯ: новые языки и новые задачи // Труды Института русского языка РАН. – 2019. – № 21. – С. 41–60.
8. Добровольский Д.О., Зализняк А.А. Корпусный подход к исследованию фразеологии: новые результаты по данным параллельных корпусов // Вестник Санкт-Петербургского университета. Язык и литература. – 2020. – Т. 17, № 3. – С. 398–411.
9. Плунгян В.А., Резникова Т.И., Сичинава Д.В. Национальный корпус русского языка: общая характеристика // Научно-техническая информация. Сер. 2. – 2005. – № 3. – С. 9–13.