Научная статья
УДК 81’42:811.112.2

Языковые средства описания внешности литературного героя: сопоставительный анализ на базе корпуса параллельных текстов

Д. В. Степанова[1], А. С. Яновская[2]

[1,2]*Минский государственный лингвистический университет, Минск, Республика Беларусь*
[1]*daryastepanova79@gmail.com*
[2]*yanovska.nastya33@gmail.com*

Аннотация. Цель настоящего исследования – выявить специфику построения моделей внешности литературного персонажа на основе сопоставительного анализа корпусов параллельных текстов. Означенная цель осуществима при помощи современных методов обработки естественного языка. В качестве инструмента использовался самостоятельно разработанный авторами программный код на языке программирования Python, а также некоторые готовые программные решения, в частности, российский программный комплекс «Генератор сбалансированного лингвистического корпуса и корпусный менеджер». В результате разработана система признаков словесного портрета и выведена универсальная модель описания внешности литературного персонажа.

Ключевые слова: корпусная лингвистика, корпус параллельных текстов, словесный портрет, детективный жанр, внешность литературного персонажа, сопоставительный анализ, русский язык, английский язык, испанский язык, белорусский язык

Для цитирования: Степанова Д. В., Яновская А. С. Языковые средства описания внешности литературного героя: сопоставительный анализ на базе корпуса параллельных текстов // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. Вып. 13 (894). С. 112–120.

Original article

Linguistic Means of a Fictional Hero Appearance Description: a Comparative Analysis on the Basis of a Parallel Text Corpus

Darya V. Stepanova[1], Nastassia S. Yanovskaya[2]

[1,2]*Minsk State Linguistic University, Minsk, Belarus*
[1]*daryastepanova79@gmail.com*
[2]*yanovska.nastya33@gmail.com*

Abstract. The present study aims to identify the specifics of constructing models of a fictional hero appearance based on a comparative analysis of parallel text corpora, which is achieved using modern methods of natural language processing. The tool used was a programming code developed by the authors in the Python programming language, as well as some ready-made software solutions, in particular, the Russian software package “Balanced Linguistic Corpus Generator and Corpus Manager”. As a result, a system of verbal portrait features was developed and a universal model for describing the appearance of a fictional hero was derived.

Keywords: corpus linguistics, corpus of parallel texts, verbal portrait, detective genre, appearance of a fictional hero, comparative analysis, Russian, English, Spanish, Belorussian

For citation: Stepanova, D. V., Yanovskaya, N. S. (2024). Linguistic means of describing the appearance of a literary hero: a comparative analysis on the basis of a parallel text corpus. Vestnik of Moscow State Linguistic University. Humanities, 13(894), 112–120.

ВВЕДЕНИЕ

Интерпретация художественного текста с использованием лингвистических корпусов актуальна в наше время. Автоматизация процесса интерпретации художественного произведения способствует созданию более точных и объективных методов анализа литературных произведений, в частности, в отношении оценки качества его перевода и описания идиостиля автора [Баранов, Добровольский, 2024; Баранов, Добровольский, Фатеева, 2021; Горожанов, Гусейнова, Степанова, 2022а; Горожанов, Гусейнова, Степанова, 2022б].

В исследовании решаются следующие задачи:

1) отобрать программные решения для проведения автоматизированного сопоставительного анализа художественных текстов и определить основные этапы создания корпуса параллельных текстов;
2) выявить основные особенности создания словесного портрета героя в художественном произведении в рамках различных подходов;
3) разработать систему признаков описания внешности литературного персонажа на основе выявленных особенностей;
4) представить в формализованном виде исследовательский материал, осуществить выравнивание параллельных текстов с учетом структурных особенностей английского, испанского, русского и белорусского языков;
5) отобрать языковые единицы лексико-семантического поля «внешность человека» и распределить данные единицы по элементам системы признаков;
6) выявить сходства и различия в характеристиках моделей внешности литературного персонажа в тексте оригинала на английском языке и в текстах перевода на белорусский, русский и испанский языки.

Лингвистическим материалом исследования послужили художественные тексты: повесть Агаты Кристи «Dead Man's Mirror» (английский язык), ее переводы на русский («Последний из Чевеникс-Горов»), белорусский («Нябожчыкава люстэрка»), испанский («El espejo del muerte») языки. Тексты представлены в полном объеме. Создан лингвистический корпус объемом 86202 словоупотребления: 23646 – текст оригинала на английском языке; 19786 – текст перевода на русский язык; 20317 – текст перевода на белорусский язык; 22453 – текст перевода на испанский язык.

Теоретическая значимость исследования состоит в расширении корпусной методологии, а именно в применении инструментов корпусной лингвистики для сравнительного анализа описания внешности литературного героя.

КОРПУС ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ КАК ИНСТРУМЕНТ АНАЛИЗА ТЕКСТА

Создание параллельного корпуса текстов представляет собой многоэтапный процесс, требующий тщательной и систематической обработки данных. Процесс создания параллельного корпуса начинается со сбора текстов на целевых языках. Тексты должны быть репрезентативными и разнообразными, чтобы обеспечить адекватное покрытие лексического и синтаксического разнообразия.

Повесть «Dead Man's Mirror» впервые была опубликована в сборнике «Murder in the Mews and Other Stories» в 1937 году как расширенная версия рассказа «The Second Gong», опубликованного в 1932 году в номере 499 журнала «Strand Magazine». Произведение входит в цикл «Эркюль Пуаро», относится к жанру классического английского детектива и переведено более чем на 30 языков. В созданный лингвистический корпус вошли следующие тексты переводов:

- русский язык: «Последний из Чевеникс-Горов» (1990 год, перевод В. Постникова);
- белорусский язык: «Нябожчыкава люстэрка» (1988 год, перевод П. Мартинович);
- испанский язык: «El espejo del muerto» (1962 год, перевод К. Перер дель Молино).

За сбором лингвистического материала следует этап выбора типа и объема исследуемого корпуса. На основе данного произведения создан *письменный литературный художественный статический многоязычный синхронический сбалансированный исследовательский лингвистический* корпус текстов. Для анализа художественной литературы часто используются сбалансированные корпусы, которые включают тексты всех или определенных произведений конкретного автора, либо даже работы писателей из одного литературного направления. Они позволяют проводить детальный анализ и интерпретацию художественных текстов, выявлять особенности стиля и тематики авторов, исследовать эволюцию их литературного творчества, а также анализировать характерные черты литературных направлений [Горожанов, Степанова, 2022].

Общий объем корпуса составил более 86 тыс. словоупотреблений, что, несомненно, затрудняет ручной анализ произведения и указывает на необходимость использования компьютерных технологий.

В нашем исследовании отдельным этапом выделен процесс фрагментации текста, т. е. разбиение

текста на отдельные блоки, в данном случае – предложения. Для иллюстрации возможностей существующих инструментов было принято решение использовать несколько приложений, предназначенных для работы с лингвистическими корпусами.

Одним из готовых продуктов такого типа является корпусный менеджер, основанный на модели реляционной базы данных для оперирования сбалансированным лингвистическим корпусом художественного произведения. Данный программный инструмент разработан в лаборатории фундаментальных и прикладных проблем виртуального образования Московского государственного лингвистического университета [Gorozhanov, Guseynova, Stepanova, 2024]. Корпусный менеджер выделил 2 тыс. 439 предложений в тексте оригинала и 2 тыс. 384 предложений в переводе на русский язык. Различие этих показателей несущественно.

Таким образом, был сделан вывод об относительной точности фрагментации текста корпусным менеджером. Наиболее существенной проблемой для данного программного инструмента стало выделение отдельных реплик диалога (в русскоязычном тексте) и прямой речи (в тексте оригинала) за счет пунктуационных различий в оформлении прямой речи в русском и английском языках (см. табл. 1):

Таблица 1

РАЗЛИЧИЯ ФРАГМЕНТАЦИИ ПРЯМОЙ РЕЧИ

id>0: [300] : “Gentlemen,” he said, “this door must be broken open immediately!”	id>0: [296] : – Джентльмены!
	id>0: [297] : – сказал он.
	id>0: [298] : – Нужно немедленно взломать дверь!

Указанная выше модель использует в качестве внутреннего инструмента библиотеку обработки естественного языка SpaCy. Не менее продуктивной платформой, используемой для разработки программ задач NLP на Python, является Natural Language Toolkit (NLTK). Она предоставляет простые в использовании интерфейсы к более чем 50 корпусам и лексическим ресурсам, таким как WordNet, а также набор библиотек обработки текста для классификации, токенизации, стемминга, тегирования, синтаксического анализа и семантических исследований, оболочки для профессиональных библиотек NLP и активный дискуссионный форум.

В качестве альтернативы готовому программному решению для решения задачи фрагментации текста можно использовать непосредственно библиотеку, например NLTK. Нами был написан программный код, соответствующий поставленной задаче, в ходе работы которого было выделено 2 тыс. 449, 2 тыс. 411, 2 тыс. 396, 2 тыс. 259 предложений в текстах оригинала, русского, белорусского и испанского языков соответственно. Такой подход позволяет легко обрабатывать текстовые файлы и разбивать их на предложения для дальнейшего анализа.

Выравнивание параллельных текстов

Важнейшим этапом создания корпуса является выравнивание текстов, где для каждой пары языков производится сопоставление текстовых фрагментов на уровне предложений. Этот этап является одним из наиболее трудоемких, поскольку требует учета различий в грамматике и структуре предложений между разными языками. Проведем обзор нескольких решений задачи выравнивания параллельных текстов.

В первую очередь, с целью сопоставительного анализа доступного программного обеспечения, предназначенного для работы с параллельными корпусами текста, использовано проприетарное (несвободное) программное обеспечение memoQ. Оно служит для автоматизированного перевода и предоставляет возможности использования терминологического словаря. С помощью означенной электронной программы можно также создавать корпус текстов в формате «исходник-перевод», работать со статистическими модулями сетевой базы данных, а также осуществлять перевод. (Он базируется на системе Translation Memory). Исходный документ и его перевод сегментируются на единицы перевода, а затем соответствующие единицы перевода сопоставляются с использованием статистических и лингвистических алгоритмов. Одной из функций программы является составление корпуса текстов, в том числе, параллельных, а также их автоматическое выравнивание (Text Aligning). Недостаток означенной системы проявляется, однако, в ограничении языков – таблица соответствий составляется на основе только двух языков. В данном случае, таковыми являются английский и русский по причине их распространенности.

С помощью программы memoQ было выделено 2 тыс. 156 строк, результат ее работы оказался более несовершенным, нежели вывод корпусного менеджера. В приведенном ниже примере можно увидеть, насколько сильны несовершенства в соответствии предложений (рис. 2). MemoQ не справилась с выравниванием реплик диалога. Фрагментация проведена частично, границы расставлены неверно, что нарушает смысловую целостность предложений.

Аналогичным программным обеспечением для выравнивания параллельных текстов на уровне

предложений является специализированный продукт LF Aligner. Он используется в переводческих и лингвистических исследованиях для создания выравненных корпусов текстов. Основные преимущества LF Aligner для нашего исследования – одновременная поддержка более чем двух языков и возможность вывода результата не только в формате TMX, но и стандартном файле XLS.

Исходя из вышеописанных подэтапов в выравнивании текстов, выделим некоторые сложности при работе с многоязычными корпусами текстов:

1. Тексты могут быть представлены в различных форматах как на этапе ввода, так и на этапе вывода результата выравнивания (.txt, .docx, .pdf и т. д.), что требует дополнительных усилий для их конвертации и нормализации.
2. Различные языки могут иметь различную структуру предложений, что затрудняет точное выравнивание.
3. Существующие алгоритмы выравнивания не всегда работают идеально, особенно для языков с сильно различающимися грамматическими структурами.
4. Для некоторых языков нет достаточного количества инструментов автоматической обработки текста (например, токенизаторов, лемматизаторов, теггеров POS) или качество существующих инструментов недостаточно высокое.

АНАЛИЗ ВЕРБАЛИЗАЦИИ ЛЕКСИКО-СЕМАНТИЧЕСКОГО ПОЛЯ ВНЕШНОСТИ ЛИТЕРАТУРНЫХ ГЕРОЕВ

Одним из существенных этапов предобработки текста с целью последующей интерпретации является выделение непосредственных объектов анализа. В рамках данной работы было принято решение в пользу использования метода выделения именованных сущностей, Named-entity recognition (NER), – процесса идентификации в тексте слов или их сочетаний, которые обозначают объекты или явления некоторой категории, такие как названия организаций, имена людей и т. д. [Petasis et al., 2001; Красикова, 2024; Чжу, Митрофанова, 2024]. В данном случае под выделением именных сущностей будем подразумевать выделение персоналий.

В пределах поставленной задачи по автоматизированному выделению основных персонажей литературного произведения были написаны несколько программных кодов на Python с использованием библиотек SpaCy, NLTK, Stanza и Polyglot.

В первую очередь, следует проанализировать текст оригинала, представленный на английском языке. С помощью библиотеки spaCy были получены списки ключевых персонажей повести:

1. Gervase Chevenix-Gore: 170
2. Hercule Poirot: 165
3. Ruth Chevenix-Gore (Lake): 69
4. Vanda Elizabeth Chevenix-Gore: 46
5. Riddle: 41
6. Miss Lingard: 36
7. Miss Cardwell: 34
8. Ogilvie Forbes: 31
9. Colonel Ned Bury: 14
10. Godfrey Burrows: 8
11. Captain John Lake: 9
12. Hugo Trent: 9
13. Mr Satterthwaite: 4

Исходя из списка частотности персонажей произведения, можно сделать следующие выводы:

1. Герои с наибольшей частотностью упоминаний, такие как Gervase Chevenix-Gore и Hercule Poirot, очевидно, являются главными персонажами и играют ключевую роль в сюжете, так как Gervase Chevenix-Gore – жертва убийства, Hercule Poirot – детектив.
2. Ruth Chevenix-Gore и Vanda Elizabeth Chevenix-Gore связаны сюжетом друг с другом, так как они являются приемными дочерью и матерью соответственно.
3. Персонажи, у которых относительно низкая частотность упоминаний, такие как Colonel Ned Bury и Godfrey Burrows, могут играть менее значительные роли в сюжете или вступают в него на определенных этапах.
4. Некоторые персонажи, например, Riddle и Miss Lingard, могут быть важными для развития сюжета, учитывая относительно высокую частотность их упоминаний.
5. Встречающиеся реже персонажи, такие как Captain John Lake и Mr Satterthwaite, могут быть связаны с определенными ключевыми моментами сюжета или представлять определенные аспекты истории, но, вероятнее всего, не влияют на финальную разгадку.

Для работы с белорусским, русским и испанским языками были использованы библиотеки и Polyglot, NLTK и Stanza соответственно. Так, обнаружилось, что особенностью библиотеки Stanza является то, что при выводе результатов кроме непосредственных имен персонажей выдается также их титулы, должностные характеристики и вежливые обращения: *inspector*, *señorita*, *señor*, *señora, lord*, *mademoiselle*, *coronel*. Также было отмечено, что Stanza выделяет повелительное наклонение в качестве именованной сущности, если оно стоит

в начале предложения и, соответственно, пишется с заглавной буквы: *Oiga*, *Vamos*, *Llévense*, *Créame*.

Особенностью библиотеки Polyglot, являющейся решающей при отборе программных инструментов для белорусского языка, является актуализация множества редких языков, которые не всегда поддерживаются в других библиотеках для NLP. NLTK может работать медленнее, нежели другие современные библиотеки, такие как spaCy. Для работы многих функций NLTK необходимо загружать дополнительные ресурсы и корпусы, что может быть неудобно. Именно поэтому данная библиотека используется для обработки текста на русском языке, где не требуется подключение дополнительных модулей.

Получив списки на выводе из программ, разработанных для русского и белорусского языков, сопоставив их с имеющимися списками имен литературных персонажей, получим языковые соответствия именованных сущностей – основных персоналий произведения (см. табл. 2).

В ходе анализа литературных произведений и изучения образов героев одной из важных задач является выделение лексических единиц, описывающих их внешность. Отбор таких лексических единиц играет ключевую роль в создании живого и наглядного образа литературного героя.

В первую очередь, вручную был проведен отбор словоупотреблений и коллокаций, служащих для описания каждого литературного героя по отдельности. Фрагменты распределения лексических единиц по каждому персонажу представлены ниже (см. табл. 3).

Таблица 2

ЯЗЫКОВЫЕ СООТВЕТСТВИЯ ИМЕН ПЕРСОНАЖЕЙ ПОВЕСТИ

Английский язык	Испанский язык	Русский язык	Белорусский язык
Gervase Chevenix-Gore	Gervasio Chevenix-Gore	Джервас Чевеникс-Гор	Гервазы Шэвені-Гарэ
Hercule Poirot	Hércules Poirot	Эркюль Пуаро	Эркюль Пуаро
Ruth Chevenix-Gore	Ruth Chevenix-Gore	Рут Чевеникс-Гор	Рут Шэвені-Гарэ
Vanda Elizabeth Chevenix-Gore	Vanda Chevenix-Gore	Ванда Джервас Чевеникс-Гор	Ванда Шэвені-Гарэ
Major Riddle	Mayor Riddle	Майор Ридл	Маёр Рыдл
Miss Lingard	señorita Lingard	мисс Лингар	міс Лінгард
Susan Cardwell	Susana Cardwell	Сюзн Кардуэлл	Сьюзен Кардвэл
Ogilvie Forbes	Oswaldo Forbes	Освальд Форбс	Освальд Фобз
Colonel Bury	coronel Bury	полковник Бэри	палкоўнік Бэры
Godfrey Burrows	Godfrey Burrows	Годфри Бэрроус	Годфры Бараўз
John Lake	John Lake	Джон Лейк	Джон Лэйк
Hugo Trent	Hugo Trent	Хьюго Трент	Х'юга Трэнт
Mr Satterthwaite	señor Satterthwaite	мистер Сэттертуэйт	містэр Сэтазвэйт
Snell	Snell	Снелл	Снэл

Таблица 3

РАСПРЕДЕЛЕНИЕ ЛЕКСИЧЕСКИХ ЕДИНИЦ ПО ОТДЕЛЬНЫМ ПЕРСОНАЖАМ

Персонаж	Английский язык	Испанский язык	Русский язык	Белорусский язык
Vanda Chevenix-Gore	quite a handsome woman	una mujer atractiva	довольно красивая женщина	прыгожая жанчына
	frightfully vague	terriblemente incierta	ужасно рассетшая	
	wears amulets and scarabs	lleva amuletos y escarabajos	носит амулеты скарабеев	носіць амулеты і скарабеяў
	tall woman	una mujer alta	высокая женщина	высокая жанчына
	dark hair was threaded with gray	cabellos oscuros con hebras de plata	темные волосы были подернуты сединой	сіваватая
	flattish, indeterminate features	las facciones imprecisas	плоские мягкие черты	дробныя, невыразныя рысы
	She was wearing an oriental-looking garment of purple and orange silk wound tightly round her body.	Vestía una túnica de aspecto oriental de seda morada y naranja, ceñida alrededor de su cuerpo.	Ее тело плотно облегало восточное с виду одеяние из пурпурно-оранжевого шелка.	Яна была апранута ў нейкі ўсходні ўбор пурпурова-аранжавага шоўку, які туга аблягаў яе цела.
	Her face was serene and her manner collected and calm.	Su rostro estaba sereno y sus ademanes eran quietos y pausados.	На лице – безмятежное выражение, манеры – спокойные, собранные.	Твар яе быў спакойны, і ўся яна была спакойная і засяроджаная.

Исходя из полученных характеристик, можно сделать следующие выводы:

1) внешность персонажей женского пола проработана гораздо четче, что указывает на создание образов сильных женских персонажей;
2) некоторые описательные элементы, не являющиеся типичными и присутствующие в тексте оригинала, упускаются переводчиками: *intelligent eyes*, *egg-shaped head*;
3) описание внешности жертвы значительно выделяется на фоне «живых» персонажей. Джерваса Чевеникс-Гора Агата Кристи описывает подобно объекту, а не литературному персонажу;
4) описание внешности убийцы отличается недостаточным количеством оценочных элементов вопреки выдвинутой гипотезе. Однако, автор, как и в случае с жертвой, пренебрежительно объективизирует личность персонажа посредством описания внешности: *funny old thing*;
5) главные участники фабулы «расследование» (*Hercule Poirot*, *Major Riddle*) не имеют развернутого словесного портрета, в отличие от персонажей фабулы «преступление».

ЛЕКСИЧЕСКИЕ ЕДИНИЦЫ, ОПИСЫВАЮЩИЕ КАТЕГОРИЮ ВНЕШНОСТИ

Значимый аспект анализа словесного портрета заключается в исследовании лингвистических средств, содействующих успешному формированию целостного образа персонажа. Так, И. Я. Чернухина выделяет категорию персонажа в качестве «универсальной смысловой категории художественного текста» [Чернухина, 1987], а его языковую реализацию – одной из задач лингвистики текста. Литературного героя целесообразно поделить на следующие категории:

– элементы неотъемлемой идентичности персонажа – черты лица, части тела, позы, жесты, мимика, движения и возраст;
– одежда и ее описания [Галанов, 1967].

Совокупность лексических категорий создания словесного портрета характеризуется как лексико-семантическое поле «внешность человека». Оно включает в себе отдельные семантические микрополя: черты лица, части тела, одежда и т. д. Из числа наиболее распространенных частей речи, участвующих в формировании поля внешности выделяются существительные (*face*, *hair*, *mouth*, *lips*, *cheeks*, *jaw*, *body*, *posture*, *built*, *etc*), прилагательные (*tall*, *short*, *slim*, *muscular*, *pale*, *green*, *rough*, *pink*, *charming*), глаголы (*to be*, *to be painted*, *to wear*, *to gaze*, *to radiate*, *to decorate*) и наречия (*darkly*, *heavily*, *constantly*) [Малетина, 2007].

На основе полученных данныхобнаруживаются следующие признаки, являющиеся микрополями и подкатегориями в общем словесном портрете:

1. Элементы неотъемлемой идентичности персонажа: черты лица, части тела, волосы, возраст, рост, телосложение, отличительные черты внешности, общая оценка образа.
2. Одежда и ее атрибуты.

Отметим, что данная структура лексико-семантического поля внешности разработана на фактическом материале, что ни в коей мере не противоречит общепринятым теоретическим аспектам, а только дополняет существующие положения в контексте избранного литературного произведения.

На основе полученных результатов, были сделаны следующие выводы:

1. В английском и русском языках описания могут быть более детализированными и содержать дополнительные художественные элементы, тогда как в испанском и белорусском такие описания могут быть более сжатыми.
2. Описания в русском языке могут содержать неформальные элементы, например, *старушенция*, что придает тексту разговорный оттенок.
3. Многие термины имеют прямые переводы на все четыре языка, что делает описание внешности интуитивно понятным независимо от языка, например: *tall – alta – высокая – высокая*.
4. Описания черт лица имеют различный стилистический окрас. В английском языке – *flattish*, *indeterminate*, *serene*, *well-chiseled nose*, *aquiline* – используются лексемы, нетипичные для данного контекста, в русском и белорусском языках используются более нейтральные и типичные характеристики – *плоские*, *мягкие*, *безмятежное*, *точеный*, *орлиный нос – дробныя*, *невыразныя*, *спакойны*, *вытачаны*, *арліны нос*.

Сделанные выводы подчеркивают важность сопоставительного анализа описаний внешности литературных героев на разных языках и раскрывают разнообразие языковых и культурных особенностей в описании персонажей.

Более того, детальные описания внешности создают определенное впечатление о персонаже и могут влиять на восприятие читателем его характера. Например, описание *пожилой человек маленького*

роста создает образ некого скромного и неприметного, в то время как *очаровательная девушка в современном стиле* может сформировать позитивный отклик и повлиять на целостное восприятие персонажа.

Учитывая все вышеизложенное, созданы многоязычные словесные портреты основных персонажей. На основе их обобщения и созданной системы признаков разработана универсальная модель описания внешности литературного героя, представленная в виде формулы.

- [Имя героя] – [возрастная категория], [общая оценочная характеристика].
- [Рост], [телосложение], [цвет волос], [прическа], [особенности лица]: [форма носа], [линия подбородка], [глаза (цвет, форма, выражение)].
- [Общий стиль одежды и аксессуары]: [тип одежды], [цвет одежды], [аксессуары].
- [Дополнительные особенности]: [уникальные черты], [особенности поведения], [общий стиль].

Формула описания внешности литературного героя может быть полезна для систематизации описания внешности героя, обеспечивая комплексный и детализированный портрет, который легко адаптировать к различным литературным персонажам. Ее использование позволяет поддерживать единый стиль и уровень детализации при описании различных персонажей, что способствует целостности и согласованности повествования. Более того, описания подобного рода могут быть использованы для создания визуальных образов персонажей, соответствующих авторскому замыслу.

ЗАКЛЮЧЕНИЕ

Проведенное исследование продемонстрировало, что методы корпусной лингвистики и автоматизированного анализа текста обладают значительным потенциалом для решения задач лингвистического и литературного анализа. Они позволяют не только глубоко и всесторонне изучить тексты, но и автоматизировать процессы, которые ранее требовали значительных временных и человеческих ресурсов.

Существующие программные средства для создания параллельного корпуса текстов все еще несовершенны. К основным проблемам эксплуатации таких инструментов можно отнести качество автоматического выравнивания по причине различий в структуре, длине предложений, а также отсутствие данных и ресурсов для работы с непопулярными языками.

Описание внешности персонажей отражает социокультурный контекст произведения и эпохи, в котором они действуют. Все персонажи характеризуются в соответствии с их внешним видом, что позволяет читателям сформировать визуальное представление о них. В описаниях указывается возраст, особенности внешности (рост, форма головы, цвет волос и т. д.), а также некоторые дополнительные детали, такие как ношение украшений или отличительные черты характера.

Продолжение исследований в этой области может привести к еще более широкому внедрению технологий автоматизированного анализа текста в различных сферах гуманитарных наук, а также к оптимизации существующих инструментов обработки естественного языка.

СПИСОК ИСТОЧНИКОВ

1. Баранов А. Н., Добровольский Д. О. Авторские особенности идиоматики в художественных текстах XIX века // Труды института русского языка им. В. В. Виноградова. 2024. № 1. С. 11–23. DOI: 10.31912/pvrli-2024.1.1. EDN JBETAS.
2. Баранов А. Н., Добровольский Д. О., Фатеева Н. А. Идиостиль Ф.М. Достоевского: направления изучения // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2021. Т. 12. № 2. С. 374–389. DOI: 10.22363/2313-2299-2021-12-2-374-389. EDN QGGOJP.
3. Горожанов А. И., Гусейнова И. А., Степанова Д. В. Инструментарий автоматизированного анализа перевода художественного произведения // Вопросы прикладной лингвистики. 2022а. № 45. С. 62–89. DOI: 10.25076/vpl.45.03. EDN IWBHQI.
4. Горожанов А. И., Гусейнова И. А., Степанова Д. В. Стандартизированная процедура получения статистических параметров текста (на материале цикла рассказов Дж. Лондона «Смок Белью. Смок и Малыш») // Вестник Минского государственного лингвистического университета. Серия 1: Филология. 2022б. № 4(119). С. 7–13. EDN PXAVUX.
5. Горожанов А. И., Степанова Д. В. Интерпретация художественного произведения: корпусный подход // Филологические науки. Вопросы теории и практики. 2022. Т. 15. №. 1. С. 203–208. EDN TCZLAF.
6. Gorozhanov A. I., Guseynova I. A., Stepanova D. V. Natural Language Processing and Fiction Text: Basis for Corpus Research // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2024. Т. 15. № 1. С. 195–210. DOI: 10.22363/2313-2299-2024-15-1-195-210. EDN FKVAOI.

7. Petasis G. et al. Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems // ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. 2001. P. 426–433.
8. Красикова Е. А. Роль корпусного менеджера в анализе употребления имен собственных в текстах электронных СМИ (на примере англоязычного корпуса CNN) // Филологические науки в XXI веке: актуальность, многополярность, перспективы развития: сборник научных трудов. Краснодар: Кубанский государственный университет, 2024. С. 45–49. EDN JPRHAE.
9. Чжу Х., Митрофанова О. А. Автоматическое выделение именованных сущностей в китайско-русском корпусе параллельных и сопоставимых текстов политической тематики // Филологические науки. Вопросы теории и практики. 2024. Т. 17. № 9. С. 3030–3042. DOI 10.30853/phil20240430. EDN GPIUBV.
10. Чернухина И. Я. Общие особенности поэтического текста. Воронеж: Издательский дом ВГУ, 1987.
11. Галанов Б. Е. Искусство портрета. М.: Просвещение, 1967.
12. Малетина О. А. Функционально-коммуникативный подход к изучению словесного портрета в художественном произведении // Вестник ВолГУ. Серия 2: Языкознание. 2007. Вып. 6. С. 154–157. EDN KDNMKB.

REFERENCES

1. Baranov, A. N., Dobrovol'skij, D. O. (2024). Specific features of author's idioms in Russian literary texts of the 19th century. Proceedings of the V. V. Vinogradov Russian Language Institute, 1, 11–23. 10.31912/pvrli-2024.1.1. EDN JBETAS. (In Russ.)
2. Baranov, A. N., Dobrovol'skij, D. O., Fateeva, N. A. (2021). Individual style of Dostoevsky: dimensions of investigation. RUDN journal of language studies, semiotics and semantics, 12(2), 374–389. 10.22363/2313-2299-2021-12-2-374-389. EDN QGGOJP. (In Russ.)
3. Gorozhanov, A. I., Guseynova, I. A., Stepanova, D. V. (2022a). Tools for automated analysis of fiction work translation. Issues of applied linguistics, 45, 62–89. DOI: 10.25076/vpl.45.03. EDN IWBHQI. (In Russ.)
4. Gorozhanov, A. I., Guseynova, I. A., Stepanova, D. V. (2022б). Standardized procedure for obtaining statistical parameters of a text (on the material of the stories by J. London "Smoke Bellew. Smoke and Shorty"). Minsk State Linguistic University Bulletin. Series 1. Philology, 4(119), 7–13. EDN PXAVUX. (In Russ.)
5. Gorozhanov, A. I., Stepanova, D. V. (2022). Work of fiction interpretation: corpus approach. Philology. Theory & Practice, 1(15), 203–208. EDN TCZLAF. (In Russ.)
6. Gorozhanov, A. I., Guseynova, I. A., Stepanova, D. V. (2024). Natural Language Processing and Fiction Text: Basis for Corpus Research. RUDN Journal of Language Studies, Semiotics and Semantics, 15(1), 195–210. 10.22363/2313-2299-2024-15-1-195-210. EDN FKVAOI.
7. Petasis, G. et al. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. ACL '01 (pp. 426–433): Proceedings of the 39th Annual Meeting on Association for Computational Linguistics.
8. Krasikova, E. A. The role of the corpus manager in analyzing the use of proper names in electronic media texts (on the material of the English-speaking CNN corpus). Filologicheskie nauki v xxi veke: aktual"nost", mnogopolyarnost", perspektivy razvitiya = Philological sciences in the 21st century: relevance, multipolarity, development prospects (pp. 45–49). Conference papers. Krasnodar: Kuban State University. 2024. EDN JPRHAE.
9. Zhu, H., Mitrofanova, O. A. (2024). Automatic extraction of named entities in the Chinese-Russian corpus of parallel and comparable texts on political topics. Philology. Theory & Practice, 17(9), 3030–3042. 10.30853/phil20240430. EDN GPIUBV. (In Russ.)
10. Chernukhina, I. Y. (1987). Obshchie osobennosti poeticheskogo teksta = General features of a poetic text. Voronezh: Izdatel'skii dom VGU. (In Russ.)
11. Galanov, B. E. (1967). Iskusstvo portreta = The art of the portrait. Moscow: Prosveshchenie. (In Russ.)
12. Maletina, O. A. (2007). Functional-communicative approach to the study of verbal portrait in a work of fiction. Volgograd State University Bulletin. Series 2. Yazykoznanie, 6, 154–157. EDN KDNMKB. (In Russ.)

ИНФОРМАЦИЯ ОБ АВТОРАХ

Степанова Дарья Валерьевна
кандидат филологических наук, доцент
начальник учебно-методического управления Минского государственного лингвистического университета

Яновская Анастасия Сергеевна
преподаватель
кафедры теоретической и прикладной лингвистики
Минского государственного лингвистического университета

INFORMATION ABOUT THE AUTHORS

Stepanova Darya Valeryevna
PhD (Philology), Associate Professor
Head of the Educational and Methodological Department
Minsk State Linguistic University

Yanovskaya Anastassia Sergeevna
Lecturer
at the Department of Theoretical and Applied Linguistics
Minsk State Linguistic University

Статья поступила в редакцию	29.09.2024	The article was submitted
одобрена после рецензирования	17.10.2024	approved after reviewing
принята к публикации	21.10.2024	accepted for publication